Deliverable 6.1

European Health Data Space Data Quality Framework

18 May 2022

This project has been co-funded by the European Union's 3rd Health Programme (2014-2020) under Grant Agreement no 101035467.

0 Document info

0.1 Authors

Author	Partner
Enrique Bernal-Delgado	Aragon Institute of Health Sciences (IACS)
Sarah Craig	Health Research Board (HRB)
Thomas Engsig-Karup	Central Denmark Region (RM)
Francisco Estupiñán-Romero	Aragon Institute of Health Sciences (IACS)
Nina Sahlertz Kristiansen	Central Denmark Region (RM)
Jesper Bredmose Simonsen	Central Denmark Region (RM)

0.2 Keywords

Keywords	TEHDAS, Joint Action, Health Data, Health Data Space, Data Space, HP-JA-2020-1, Data Quality

Accepted in Project Steering Group on 26 April 2022. The European Commission gives final approval to all joint action's deliverables.

Contents

1 Executive summary 3
2 Introduction 3
2.1 TEHDAS Work Package 6: Excellence in Data Quality 5
2.2 TEHDAS Task 6.1: Develop the EHDS Data Quality Assurance Framework (DQAF) 5
3 Methodology 6
3.1 Overall Methodology 6
3.2 Thematic Work Groups and Workshops 7
3.3 Review of existing initiatives 8
3.3.1 Identification of Initiatives 8
3.3.2 Inclusion Process 8
3.3.3 Data Collection and Analysis 9
3.4 Literature Scoping Review 13
3.4.1 Data Sources and Search Strategy 13
3.4.2 Inclusion Process and Data Extraction 14
4 Results 14
4.1 Thematic Work Groups: Findings and recommendations 14
4.1.1 Definition of data quality 14
4.1.2 Key Dimensions of Data Quality 15
4.1.3 Dimensions at Data Source Level 16
4.1.4 Data Quality Dimension at the Institution level 16
4.1.5 Data quality assessment and benchmarking 17
4.1.6 Level of assessment 18
4.1.7 Minimum criteria for a data quality framework (DQF) 18
4.1.8 Benchmarking model/system 19
4.2 Review of Existing Initiatives 19
4.2.1 Characteristics of initiatives 19
4.2.2 Semantic interoperability 20
4.2.3 Data Quality Definition and Dimensions 20
4.2.4 Quality Assessment in the Data Lifecycle 21
4.2.5 Legal aspects and governance of data quality management 21
4.2.6 Technical services or tools 22
4.3 Literature Scoping Review 22
4.3.1 Knowledge Summary 29
5 Recommendations 29
6 References and documentation 31
6.1 Background Materials and Documentation 31
6.2 Literature Scoping Review References 32
Annex 34

1 Executive summary

This report explores and synthesizes the existing knowledge and experiences on data quality frameworks (DQFs) in the context of cross-border sharing of federated secondary use health data with the aim to identify good practice within this area and make recommendations. The report builds on the work regarding data quality already undertaken the TEHDAS Joint Action and will be further updated with chapters on interoperability standards. This first part of the final report contains recommendations on the European Health Data Space (EHDS) data quality framework.

The recommendations are synthesized from the results of three parallel lines of work, based on three different methodologies. These three approaches are 1) thematic workshops and partner meetings, 2) analysis of existing data sharing initiatives and 3) a literature scoping review.

The main recommendations are:

- The adoption of a working definition of data quality that focuses on data "fitness for purpose" and how well data reflects the reality it represents.
- Reliability, relevance, timeliness, coherence, coverage and completeness should be adopted as measurable dimensions of data quality and incorporated in a DQF.
- Promote a focus on transparency at the level of institutions across Member States in relation to regular audits, a well-developed DQF and clear procedures in relation to processing the data.
- National competent institutions should audit data holder institutions on procedures of quality assurance and their data sets in accordance with the EHDS DQF.
- Data holders should be obligated to publish their data preparation procedures, as well as metadata about their collections including information on data provenance, relevance and coverage of the data collection.
- Initiatives should focus on continuous improvement, encouraging good practice, design, development and implementation of toolkits for quality assessment and allocate resources to support data quality-focused work.
- In the medium to longer-term promote the development of a benchmarking process which will assist data managers and institutions with alignment against a Europe-wide approach to measuring data quality.

2 Introduction

This TEHDAS Deliverable 6.1 contains recommendations on the EHDS Data Quality Framework (DQF) and is the first part of the report, which will be updated with chapters on the more technical and architectural aspects of data quality, including interoperability standards, in month 24 of the TEHDAS project. This final update will also draw on the outputs of WP5 and WP7.

The purpose of TEHDAS Task 6.1, as summarized in this report, was to explore and synthesize the existing knowledge and experiences on data quality frameworks (DQFs) in the context of cross-border sharing of federated secondary use health data, and thereby identify good practice within this area and make recommendations.

As explained below, this deliverable is the main output of subtask 6.1.1 “Assessment of DQAF good practices”, whereas the milestone document M6.1 deals with subtask 6.1.2 “Assessment of existing European legislation” and of course the features of an EHDS DQF that could be legally bound.

It should be noted that this report refers to a Data Quality Framework (DQF) instead of Data Quality *Assurance* Framework (DQAF), which is the term used in the TEHDAS Project Plan. The change is made to shift focus to continuous improvement and promotion of data quality.

As such, this Deliverable is a representation of the overall work done in TEHDAS WP6 during the first 14 months of the project. The content of the current deliverable builds on the framing and narrowing of the scope of TEHDAS Task 6.1 done through partner meetings and thematic working groups from May 2021 to the present. The first stage of this work has been summed up in the TEHDAS Milestone 6.1 document “Identifying those data quality features that could be legally bound and providing advice to the European Commission” from October 2021. The main conclusions of the Milestone 6.1 document are seen as key parts of the work in Task 6.1 and are presented here along with other key findings and conclusions of this Deliverable.

The thematic working groups and the partner meetings early in the process of Task 6.1 have been crucial for developing a common understanding of the task, assessing the current status of data sharing and scoping the work and deliverables. To briefly sum up, there was agreement on four main points, which all play a part in scoping the framework for Task 6.1 overall. These four points, taken from the Milestone 6.1 document, are:

- Collection, use and storage of healthcare data is organised differently across Member States. This makes it difficult to compare data between data sharing initiatives and between the Member States.
- Data quality is multidimensional. Quality is relative to the need of the user and hence a particular data set may meet the quality requirements of one user, but not of another. Generic metrics for quality measurements cannot be directly applied to all data sets.
- The task given to Work Package 6, when looking at data quality in an EHDS setup, has been to recommend a DQF which can accommodate all relevant institutions in all Member States. This premise of inclusiveness means that every Member State should be able to take part in the EHDS and that the levels of data quality and auditing should balance this premise.
- All the references and recommendations on governance and legal matters in the Milestone 6.1 document must be seen as relating to data quality and the introduction of a DQF, which should be differentiated from the work on legislation and governance done in TEHDAS Work Package 5 “Sharing data for health”.

Based on these points, as confirmed by the Task 6.1 partners, the scope of the task has been focused on the DQF at any stage of the data life cycle, and in particular, from data collection to the point where research finalizes. Data collection, in this context, refers to the processes implemented by data processors for its use in secondary purposes – policy making, regulation and research. A DQF aimed at the recording of patient data at the point of care is deemed out of scope for this work.

The Deliverable 6.1 builds on a three-strand approach to the task. The first are the workshops of the Thematic Working Groups, the second is the Analysis of existing initiatives, and the third is the Literature Scoping Review. This approach ensures input and consensus from Work Package 6 and other TEHDAS partners in the process towards the recommendations, while at the same time the scientific approach of the literature scoping review ensures that all identified relevant initiatives and publications are considered in the recommendations. Each of the three strands makes up a chapter in the deliverable, whereas the recommendations are made based on synthesizing the knowledge from the three strands, as well as the M6.1 report. The approach is explained in detail in the Methodology section.

2.1 TEHDAS Work Package 6: Excellence in Data Quality

The TEHDAS Project Plan defines the overall scope and objectives of Work Package 6 as follows:

Work Package 6 of the TEHDAS Joint Action will be providing solutions for the trustworthy secondary use of health and health care data with a view to fostering the digital transformation of the European health systems.

This overarching objective will be developed throughout two operational objectives:

- Developing the EHDS data quality assurance framework for a secondary use of real-world health data
- Developing the EHDS Semantic Interoperability framework

2.2 TEHDAS Task 6.1: Develop the EHDS Data Quality Assurance Framework (DQAF)

Task 6.1 will deliver on the first operational objective: “Developing the EHDS data quality assurance framework for a secondary use of real-world health data”, based on the work done in the two subtasks:

Subtask 6.1.1: Assessment of DQAF good practice

Subtask 6.1.2 Assessment of existing European legislation

The Milestone 6.1 dealt with the assessment of existing European legislation, and specifically the identification of those features that could be legally bound. Thus, this deliverable, D6.1, will be focusing on the Subtask 6.1.1: “Assessment of DQAF good practice” with the aim of offering recommendations on the quality definitions and dimensions important for the inclusion in the EHDS and data quality governance. It is important to notice that the actual legal and governance aspects of the EHDS in a TEHDAS context is covered in WP5. Issues that require liaison with WP5 are for example, what institutions should supervise the data

quality framework implementation at national and EU level, if a labelling system were to be implemented, what institution should be in charge of the system, etc..

3 Methodology

As presented in the Introduction, Deliverable 6.1 incorporates the work in Milestone 6.1, as the milestone broadly covers subtask 6.1.2 “Assessment of existing European legislation” in dealing with the legal aspects of data quality within the EHDS and the identification of those features relating to data quality that could be legally bound.

This deliverable covers TEHDAS Task 6.1 in its entirety, dealing with both subtask 6.1.1 “Assessment of DQAF good practice” and subtask 6.1.2 “Assessment of existing European legislation”.

In order to be able to conclude on Task 6.1 as a whole, including both subtasks, this chapter will briefly sum up on the overall methodologies used during the entirety of the Task 6.1 work. The four points outlined in the introduction show the scoping procedure of Task 6.1, which has a direct impact on the outline of this Deliverable, this subchapter explains the framework and scope in the light of which Deliverable 6.1 should be understood.

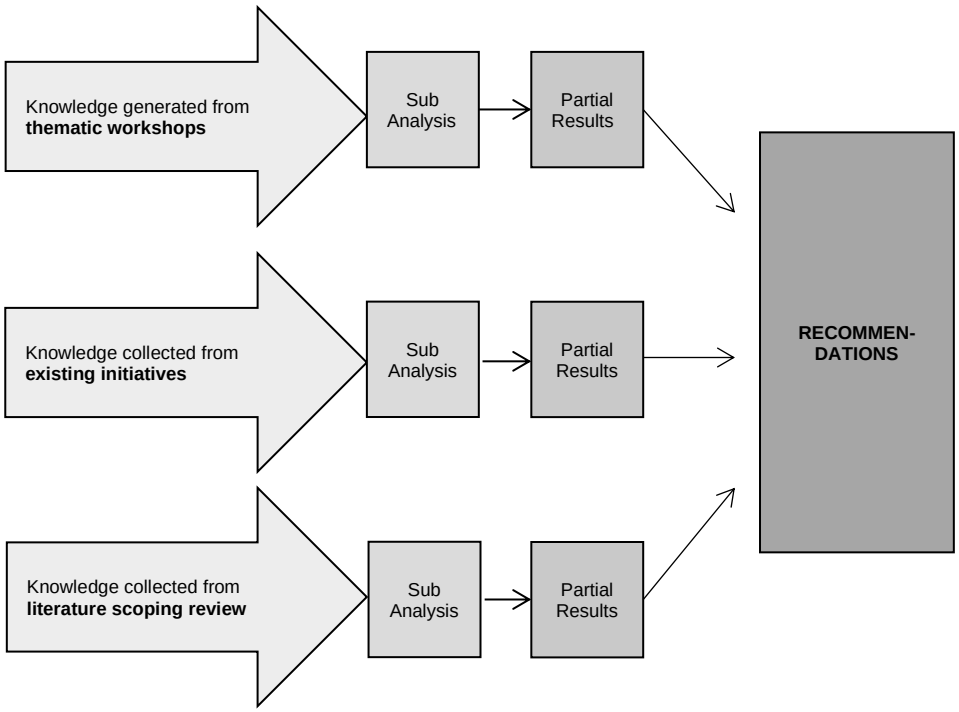
3.1 Overall Methodology

The methodologies used throughout the work in Task 6.1 have been a combination of workshops and knowledge collection through study of existing initiatives and literature reviews. The aim has been to ensure the scientific quality and objectivity through the review of the existing initiatives and the literature, whereas the thematic working groups and partner meetings have functioned as arenas for discussions, partner input from the allocated experts, and as consensus building exercises.

In practical terms, this has led to a three-strand approach, as illustrated in Figure 1. The first strand is the ongoing discussions in the thematic working groups. These are tasked with framing the discussion on specific themes, such as governance, data definition etc., and these outputs serve to both scope and focus the work as well as delivering a direct output in terms of recommendations. The second strand is the identification and analysis of relevant existing initiatives, from which knowledge on existing DQFs can be extracted and serve as an example. The third strand is the literature scoping review, which has been an ongoing process alongside the thematic working groups.

All three strands pass through an analysis phase, from which the final recommendations can be drawn.

Figure 1: Overview of methodology



3.2 Thematic Work Groups and Workshops

In order to best utilize the expertise of Work Package 6 partners as well as ensuring manageable discussions, a number of thematic work groups were established, which allowed for longer discussions on very specific subject matters. Two thematic working groups on “Data Quality Legislation” and “Data Quality Governance”, respectively, convened from June 2021 and through the process towards finalising Milestone 6.1 in September 2021. The input from these meetings and subsequent presentation and feedback from the wider group of partners in Work Package 6 provided the main content of the milestone document. Furthermore, the thematic working groups and partner meetings played a large part in the scoping of the work in Task 6.1 as explained in the introduction.

The next steps in the process were another framing of the scope and the initiation of the Scoping Review explained below. A partner meeting was held on 16 December 2021 where the process towards finalising this deliverable was presented and discussed. This included the setting up of two additional thematic working groups, to continue dividing the work into manageable tasks. One thematic group was established to look at definitions and dimensions of data quality and assessment and benchmarking mechanisms. Its work focused on a series of workshops which were set up to invite Member States (MS) to discuss experiences in relation to data quality and mechanisms for assessment and benchmarking. Three workshops were held on 8 February 2022, 22 February 2022 and 14 March 2022. The input from these workshops helped to inform the thinking of the wider group about these key issues of data quality. The workshops provided input on international practice in relation to health data quality issues and inputs and suggestions from work already under way in different countries were sought. At each workshop there was a recap of issues raised at previous

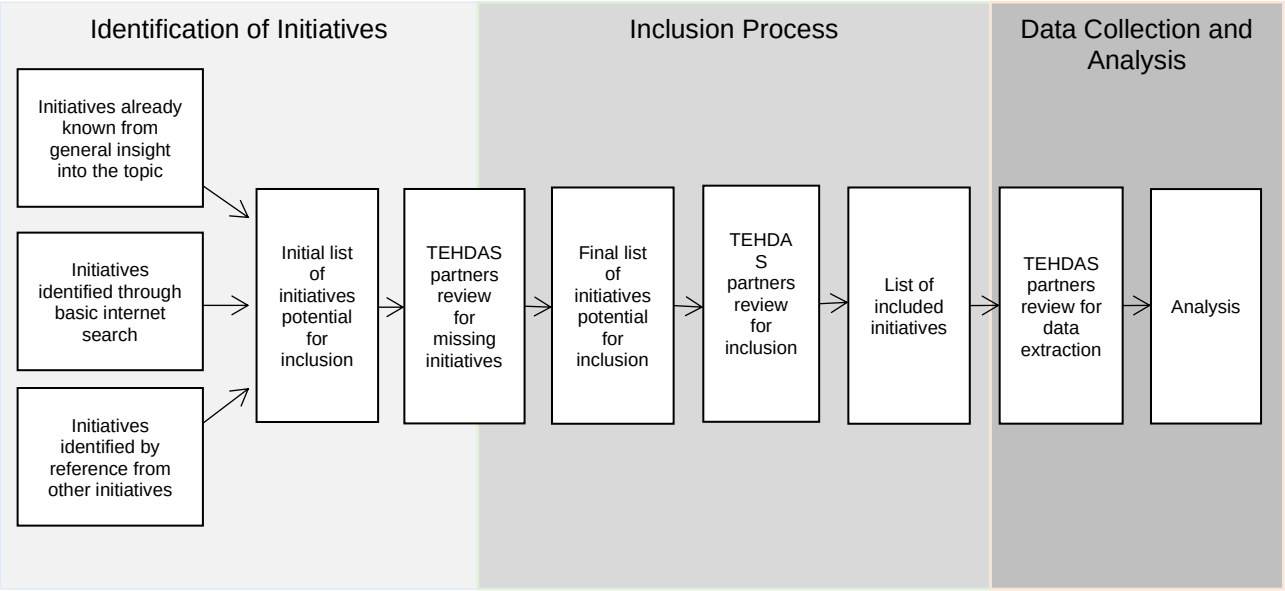
workshops to ensure that the process reflected a consensus among the participating MS. In addition, notes from the meetings were circulated for further comment or clarification.

3.3 Review of existing initiatives

The scope of this review was to identify already existing initiatives (collaborations on data-infrastructures) that apply DQFs in their programs and extract knowledge based on their experiences.

As presented in Figure 2 the strategy was to maintain a systematic and objective approach throughout the review process.

Figure 2: Overview of the methodology for collecting knowledge from initiatives



3.3.1 Identification of Initiatives

The first step in the identification process was to devise an initial crude list of initiatives (See Annex Table 6) possible for inclusion as a convenience sample. This crude list consisted of initiatives known from general insight into the topic, initiatives found through basic internet searches for international health data collaborations, and initiatives that were mentioned, or used as references, at the sites of already identified initiatives (recursive search). In order to ensure a comprehensive crude list of all initiatives that should be considered for further inclusion the second step was to invite the partners to review the list and note if any possible relevant initiatives were missing.

3.3.2 Inclusion Process

In order to maintain a systematic and objective selection process we conducted a pre-defined inclusion questionnaire presented in the table below. The questionnaire was designed to reflect the questions embedded in the purpose of the TEHDAS Task 6.1 (to explore and synthesize the existing knowledge and experiences on data quality frameworks (DQFs) in

the context of cross-border sharing of federated secondary use health data, and thereby identify good practice within this area and make recommendations).

Each partner was assigned a number of initiatives from the crude list for which they were asked to fill out the inclusion questionnaire. The assignment of initiatives was completely random. This first round of vetting was done to identify and select the initiatives most eligible for further analysis. It was a quick fact check and not a detailed examination.

As the final question in the inclusion questionnaire, the partners were asked to make a recommendation on whether or not to include the initiative for further analysis based on initial impressions and the partners' experience.

Based on the answers of the inclusion questionnaire, and particularly taking into account the partners' opinions regarding further inclusion, a final list of included initiatives was conducted.

Table 1: Initiatives Inclusion Questionnaire

Is exchange of healthcare data the main focus of the initiative? (Yes/No)
If so, what type of healthcare data (e.g., specific diseases, population health)?
Is data collected for secondary or primary use or both?
What is the source of data (e.g., clinical studies, medical records, national registries)?
Funding: Is the initiative operated under a public or private funding scheme? (State the source of funding)
Does the initiative share data across borders? (Yes/No)
If so, is data only shared within the EU? (Yes/No)
Is a Data Quality Assurance Framework (DQAF) implemented/operational? (Yes/No)
If so, is the DQAF documented and publicly available? (yes/no)
Is a Metadata Catalogue available? (yes/no)
Is the initiative described/analysed in publication? (peer reviewed paper, report, protocol, white paper etc.)
Please briefly state reasons why this initiative should be included or excluded from further analysis? (Main inclusion criteria should be whether or not the initiative demonstrates excellence in data quality assurance)

3.3.3 Data Collection and Analysis

Data from the included initiatives was extracted by selected partners, using a pre-defined data extraction questionnaire presented in Table 2 below.

The extracted data was summarized and the experiences were analysed using a qualitative approach and with relevance to the wording of the TEHDAS Task 6.1 purpose.

Table 2: Initiatives Data Extraction Questionnaire

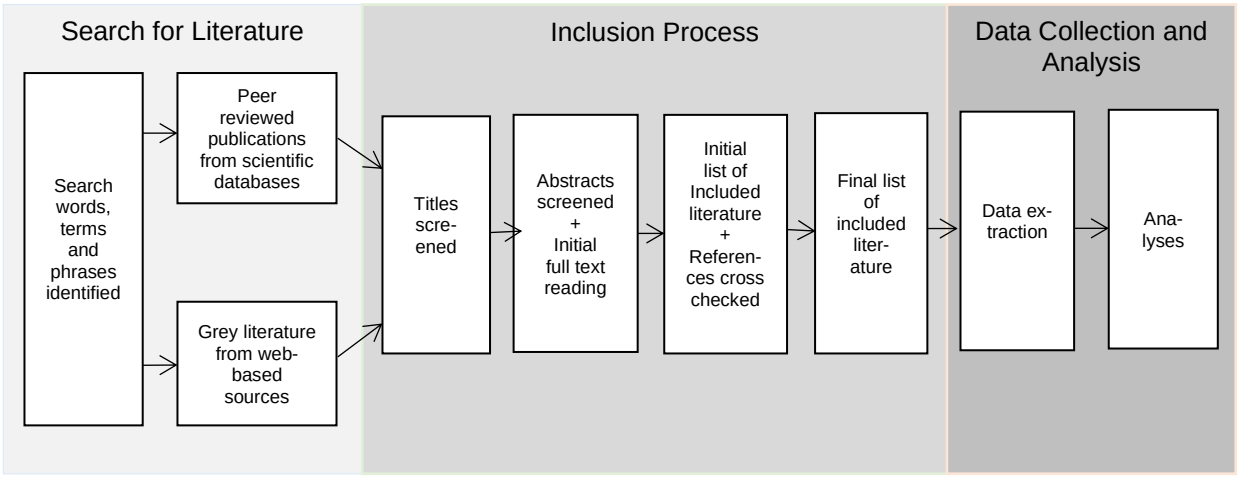
Name *What is the name of the initiative?*
[Name]
Domain *In which domain does the initiative operate?*
[Specific diseases, population health, general data sharing, genomics etc.]
Overall Framework *What is the overall framework of the initiative?*
[Project, joint action, research network, national health data authority, association, private company, foundation etc.]
Organisation *How is the initiative organised? Explain the governance model.*
[Steering committee, expert groups, independent members, national authorities etc.]
Semantic Framework *Has the initiative implemented a semantic interoperability framework?*
[Is the framework documented? Which standards are used?]
Data use *Secondary or primary use of data?*
[For which purpose was data initially collected and for which purpose is data now shared?]
Data sharing *Which entities are sharing data through this initiative?*
Data Quality Definition *Does the initiative operate with a specific Data Quality Definition? If yes, which? If no, are there any considerations as to why not?*
Data Quality Dimensions *Does the initiative operate with a specific set of Data Quality Dimensions (ex. accuracy, missingness, timeliness etc.)? If yes, which? If no, are there any considerations as to why not?*

Implementation of Data Quality Dimensions *If dimensions are used, how are the dimensions operationalised/implemented?*
Please explain data quality assessment at the following steps in the data lifecycle (corresponding with the EHDS users journey)
1. Data collection *Are audits or validation rules implemented?*
2. Data publication *Is a metadata catalogue available?*
3. Data discovery *Is a data dictionary or code book available?*
4. Data access *(This step may not be relevant for data quality assessment)*
5. Data delivery *Are processing procedures published?*
6. Data analysis *Does the initiative use auditable software?*
7. Finalisation *(This step may not be relevant for data quality assessment)*

Rationale for the overall Data Quality Assessment *Please distinguish between institutional level, data source and variable level*
Institutional level: Data source level: Variable level:
Legal aspects *What could be legal barriers to ensuring data quality?*
Governance aspects *Who is responsible for managing or enforcing data quality in the initiative?*
Auditing and/or rating system *Is an auditing or rating system implemented? If yes, please elaborate*
Semantic and syntactic interoperability *Is a data model specification available?*
Technological services and tools *Please refer to services and tools used by the initiative specifically aimed at data quality*
Anonymisation/pseudonymisation *Are there tools used for anonymisation/pseudonymisation?*
Commonalities and differences across experiences *What sets this initiative apart from other comparable initiatives?*
Please add any other relevant information below

3.4 Literature Scoping Review

Figure 3. Overview of the Methodology for Collecting Knowledge from Literature Scoping Review



3.4.1 Data Sources and Search Strategy

To ensure a systematic and objective collection of knowledge from the literature the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) – Extension for Scoping Reviews (1) served as the methodologic guide for this scoping review.

The starting point for the literature search was the purpose of TEHDAS Task 6.1: "to explore and synthesize the existing knowledge and experiences on data quality assurance frameworks (DQAFs) in the context of cross-border sharing of federated secondary use health data, and thereby identify good practice within this area and make recommendations".

From the defined purpose a comprehensive list of search words, phrases, and related MeSH-terms was conducted. This list is presented as Table 1 in the Annex (the MeSH terms included in the presented list are derived from the PubMed/MEDLINE database).

The strategy included searches for both peer reviewed scientific publications and for grey literature. Thus, comprehensive non-restricted searches on the Google Scholar Database, the PubMed/ MEDLINE Database, the Scopus Database, and the Web of Science were performed. The full search strategies are presented in Table 2 and 3 in the Annex (the search strategy for the scientific databases is exemplified by the PubMed/MEDLINE search).

The methodology for the searches on the scientific databases was to first build up blocks for ”Data”, ”Data purpose”, ”Quality”, and ”Dimensions/Quality measures”. Including both search

words, phrases (enclosed in double quotes) and MeSH-terms, advanced search builders were used to create the individual search blocks which were afterwards combined by Boolean operators (AND, OR and NOT) as relevant. All search results were lastly restricted to English language, pertaining to humans, and the last 10 years prior to the date of search.

3.4.2 Inclusion Process and Data Extraction

All studies of an acceptable quality that covered the purpose of the TEHDAS Task 6.1 were eligible for inclusion. All citations were exported to a citation manager (EndNote 20, Thomson Reuters) and checked for duplicate publications. Titles and abstracts were screened for relevance to the TEHDAS Task 6.1 purpose and publications clearly not relevant were excluded without further attention. The abstracts of the remaining publications were further reviewed and also an initial full text read was performed. Publications excluded hereafter were recorded with a short explanation for the exclusion. As a final step, the references of the included studies were crosschecked for any relevant publications not captured by the search.

Data from the included publications were extracted, following a pre-defined data extraction template covering study characteristics, study findings, and quality assessment of the publication. The data extraction template is presented in table 4 in the Annex, exemplified with the earliest (by year) of the included publications. Quality assessment followed a pragmatic qualitative reasoning, because no formal quality checklists for the main study type (conceptual research) were found.

4 Results

4.1 Thematic Work Groups: Findings and recommendations

This section will provide recommendations for the European Commission and Member States (MS) with regard to the inclusion of a Data Quality Framework (DQF) in the EHDS. The section presents views of WP6 experts from several MS and recommendations on:

- A definition of data quality
- Key dimensions of data quality as they apply at the level of data sources and at institutional level
- Mechanisms and models for benchmarking and assessment of data quality.

4.1.1 Definition of data quality

There are varying definitions of data quality and within the workshops a number of these were considered. For example, definitions used by the OECD and by the ISO refer to the ‘fit for purpose’ aspects of the data. The definition included in the TEHDAS glossary was also reviewed for its relevance. Most definitions focus on how data fits the purpose for which they are intended. In addition, the W3C Data Quality Vocabulary (DQV), which is linked to DCAT (https://www.w3.org/TR/vocab-dqv/) provides valuable background material on data quality and quality measures. The DCAT provides a framework in which the quality of a dataset can be described, whether by the dataset publisher or by a broader community of users. It does not provide a formal, complete definition of quality, rather, it sets out a consistent means by

which information can be provided such that a potential user of a dataset can make his/her own judgment about its fitness for purpose.

In the context of TEHDAS, the emphasis of data quality is ensuring that the data are fit for purpose for making decisions, supporting health research and population health. There are three main usages highlighted in the EHDS: Health research, policy making and regulation. It was agreed that the definition should incorporate the secondary use of data for research purposes.

Interoperability is seen as a requisite for the high-quality reuse of data but not as a feature consubstantial to quality. Interoperability for data discovery, for the developments of common data models, and for communication of digital objects across nodes will be developed in a dedicated section, in a future version of this DQF.

The definition of data quality proposed for the purposes of the EHDS then is as follows:

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

The difficulty with this definition is that it does not address the question of what level the quality is measured on, e.g., variable, data source or institutional level. This is addressed below, in the dimensions of data quality.

4.1.2 Key Dimensions of Data Quality

In this section, our considerations are set out on how to identify the core dimensions of data quality at:

- Institutional level
- Data source level

We also consider here whether data quality dimensions should be legally bound or encouraged based on guidance, education and by using standards.

A definition of data quality has little value without a definition of the dimensions of data quality that are measured and the level at which they are measured. For example, timeliness is important at data source level but the need to be transparent is important at institutional level. Through the examination of a number of data quality frameworks it is possible to extract a list of the dimensions that feature most often, and which have been subject to international comparison. For this report, the dimensions used by the Canadian Institute for Health Information (CIHI, https://www.cihi.ca/), the European Social Survey (ESS, https://www.europeansocialsurvey.org/), the United Nations Statistics Division (UNSTATS, https://unstats.un.org/), the Organisation for Economic Co-operation and Development (OECD, https://unstats.un.org/), and Health Data Research UK (HDR UK, https://www.hdruk.ac.uk/) were examined. From this examination, the dimensions that featured most were reliability, relevance, timeliness, coherence, coverage and completeness. Other dimensions considered included accessibility but it was agreed to omit accessibility as a dimension as it falls more appropriately under the remit of work package 5.

4.1.3 Dimensions at Data Source Level

Workshop participants were asked to consider which of the dimensions are key dimensions of data quality in the context of TEHDAS and to share any experiences of working with specific dimensions in their own countries or organisations.

In some countries, like Norway, completeness or coverage is measured while in Denmark quality assessment is often done at the level of data variables. It was also noted that data quality is often related to incentives, e.g., procedures or diagnosis recorded for reimbursement may have better coverage/completeness because of the financial incentive. The example of the Data Quality Framework developed by Health Data Research UK was considered for its identification of a broad range of categories and dimensions with definitions and a rating system going from bronze to platinum.

The table below sets out the dimensions that were deemed to be most important at the data source level and how they might be defined, as agreed by the Work Package 6 partners.

Table 3: Data Quality Dimensions at the Data Source Level

Dimension	**Definition**
Reliability	How closely it reflects what it was designed to measure and whether this is consistent over time.
Relevance	Meets the needs of users of the EHDS.
Timeliness	Collected within a reasonable period of time and collected/reported on dates agreed, e.g., close to decision makers' time of decision.
Coherence	Consistent over time and across data holders and can be combined and compared with other data sources.
Coverage	The degree to which the data adequately covers the population/event (i.e. representativeness)
Completeness	How complete are the variables?

It was noted that the list closely resembles how quality is measured at cancer registries and other health information systems internationally.

4.1.4 Data Quality Dimension at the Institution level

For the purposes of TEHDAS it was agreed that a broad range of health data is within scope. Transparency is key. At the level of the institution, it was considered appropriate to revisit the matters of regulation that were set out in the Milestone 6.1 report. From the items set out in that report, those relevant to data quality are listed below.

Table 4: Data Quality Dimensions at the Institutional Level

		Legal enforcement	
Item	**Matters of regulation**	R	M
Data collection	Regular audits		√
	Rating system and promotion	√	
Data publication	Meta-data catalogues		√
	Building synthetic data sets mirroring data collections | publishing visual analyses of quality at variable level	√	
Data delivery	Clear processing procedures (guidelines published)		√
	Not hampering meaningful reuse – pseudonyms as preferred system	√	
	Auditable software	√	

R=Recommended, M=Mandatory

This explains the elements of data quality at institutional level and it requires clarity on the dimensions that data holders are to be audited on. Therefore, dimensions are critical. From the perspective of a data quality assessment framework the dimensions were synopsised as key questions that need to be asked when assessing data quality.

1. Does a Data Quality Assessment Framework exist?
2. Are there regular audits on procedures of quality assurance?
3. Are clear data processing procedures operational and guidelines published?
4. Is a meta-data catalogue published?

A further suggestion was to incorporate data user feedback into the dimensions of data quality, especially as a tool to correct errors in the data. Users need to know about the different dimensions of quality in the data collections. This may apply to linked data sources or individual data sources.

The inclusion of anonymisation/pseudonymisation as a dimension was considered, but it was agreed that even if this affects the fitness for purpose, anonymisation is not a dimension as such and should rather be specified as an element of data quality.

4.1.5 Data quality assessment and benchmarking

The third workshop on assessment and benchmarking heard the experiences of both Norway and Ireland in relation to the collection and assessment of health data for secondary purposes and for standardisation. In Norway a national health metadata specification has been developed and in use since 2019. Metadata is stored in a repository that can be accessed through a search portal. At present, only two properties at data source level are directly linked to data quality: “Coverage” and a “Quality note”. The Data Quality Vocabulary (DQV) of W3C (World Wide Web Consortium) referred to earlier has been used as a reference standard in the work.

In Ireland, the data quality dimensions implemented by the Health Information and Quality Authority (HIQA, http://www.higa.ie), the national regulatory body for health information, closely align with those identified above. HIQA has developed guidance on a data quality framework and is looking at the full data quality cycle from e-learning modules to a programme of review focussed at the level of the institution.

In both Ireland and Norway, there has been considerable focus on self-assessment as the mechanism for raising awareness of data quality. It was acknowledged that any assessment of data quality requires resources and investment at a national level. In Ireland, information management standards are in place, which cover national health and social care data collections. These cover national registries, administrative data sources, national census national surveys. One of the standards focuses on data quality and asks of data holders: Do you have a data quality framework? Are you undertaking audits? Are policies and procedures in place in relation to data quality? These are the types of things assessed during reviews/audits.

Some of the learnings from the programme to date around data quality are clarity around roles and responsibilities, having a data quality strategy and the need for ongoing audit and assessment.

The Health Data Hub in France is also working on a metadata catalogue with descriptions of data at the database, tables and variables levels (https://catalogue-metadonnees.health-data-hub.fr/)

4.1.6 Level of assessment

The question as to at **which level** (data sources, institutions, nodes or other) should **data quality be assessed** was also considered. There was general agreement that data quality should be assessed at the institutional level and that the EHDS nodes should be responsible for ensuring transparency and implementation of data quality assessment procedures. There was also agreement that data quality assessment procedures should be closely linked to national data quality improvement efforts.

4.1.7 Minimum criteria for a data quality framework (DQF)

The question was also posed with regard to the minimum criteria for a data quality framework (DQF) that can be implementable; both at EU and national level. It is important to acknowledge the role that institutions play in making improvements, so institutions need feedback. It was acknowledged that regardless of the minimum criteria, the focus on implementing data quality initiatives has to be on a soft law approach. Much of the work to date in relation to data quality focuses on good practice guidelines and draft standards. This is largely in recognition that it would be too difficult to make data quality management mandatory.

It was agreed that there should be focus on driving quality at the front-line, at the point where information is first gathered from the patient or service user. Regulation should - at the national level - be directed towards education and information and promoting best practice through self-assessment and audit.

4.1.8 Benchmarking model/system

There was general agreement that data quality assessment comes first and then benchmarking can come later. Benchmarking could be a medium to long term goal, but the initial focus must be on getting the assessment up and running and then maybe the benchmarking will come as other work is done around standardisation. It was agreed that benchmarking would be challenging to include in the first phase. The need to focus on operationalising the dimensions for use in self-assessment was seen to be a crucial first step.

4.2 Review of Existing Initiatives

The review of initiatives described in section 3.3 resulted in a crude list of 47 initiatives. Each partner was assigned 1-3 initiatives from the crude list and was asked to fill out a brief questionnaire for each (see section 3.3).

As the final question in the inclusion questionnaire, the partners were asked to make a recommendation on whether or not to include the initiative for further analysis based on initial impressions and the partners' experience.

Answers were received regarding 31 of the 47 initiatives and the following 8 (out of 31) were selected for more detailed analysis/data extraction (see section 3.3.3.):

- OHDSI (Observational Health Data Sciences and Informatics)
- ECHO – ECHOAtlas (European Collaboration for Health Optimization)
- HRIC (Health Research and Innovation Cloud)
- Orphanet
- PHIRI (European Health Information Portal)
- Research Data Alliance
- EHDEN (European Health Data and Evidence Network)
- HIQA (Health Information and Quality Authority)

4.2.1 Characteristics of initiatives

One third of the 31 analysed initiatives operate in specific health care domains with the majority focusing on cancer, infectious diseases (incl. Covid-19), rare diseases, genomics and population health. The rest are not focusing on specific types of health data.

There are three main data sources out of 23 (out of 31) initiatives that collect data. Note that some initiatives use more than one data source:

- Medical records (5)
- Research data (10)
- National Registries (12)

In terms of funding, only three initiatives are operated under a private funding scheme and two in public/private partnership. The vast majority of initiatives are funded by the EU or Member States.

Data are shared across borders in 14 initiatives (45%), but only within the EU with the exception of two global initiatives.

Only five out of thirty-one initiatives have a Data Quality Framework implemented and operational and four of those had the DQF documented and publicly available. However, metadata catalogues were available at a third of the initiatives (10/31) and all initiatives except four has been described/analysed in a publication.

4.2.2 Semantic interoperability

Semantic interoperability frameworks are not explicitly mentioned on the websites or publications on the initiatives selected for detailed analysis, but references to OMOP, HL7 FHIR, SNOMED-CT and CDISC are frequently mentioned in the returned questionnaires and compliance with international health information standards is encouraged. Semantic interoperability frameworks are implemented through the adoption of a Common Data Model, e.g. OMOP or ECHO CDM.

4.2.3 Data Quality Definition and Dimensions

Data quality is defined by the Research Data Alliance as "a dataset's fitness to serve its purpose in a given context."

The Irish Health Information & Quality Authority (HIQA) defines quality data as data that is "fit-for-purpose" and uses the following five Data Quality Dimensions:

1. relevance;
2. accuracy and reliability;
3. timeliness and punctuality;
4. coherence and comparability;
5. accessibility and clarity.

HIQA has developed guidelines with a quality assessment tool that can be used to assess the quality of data against the dimensions, including a template for a *data quality improvement plan*.

Observational Health Data Sciences and Informatics (OHDSI, https://ohdsi.org/) work off the dimensions defined by Kahn et al. (2016). The categories are conformance, completeness, plausibility and two data quality assessment contexts: Verification and validation. These dimensions/categories are implemented in the OHDSI Data Quality Dashboard (DQD). The tool applies and evaluates 3,000+ checks and *reports the results to the user.*

ECHO (European Collaboration for Healthcare Optimization) operates with the Data Quality Assessment Framework established by Eurostat, which defines data quality as *data fit-for-purpose that yield high-relevant high-value statistical products as a result of their analysis,*

achieved through reliable and reproducible statistical production process. The following dimensions are used: Coherence, coverage, relevance, internal reliability and accuracy.

4.2.4 Quality Assessment in the Data Lifecycle

1. Data collection

With the exception of HIQA, the national authority in Ireland, all the initiatives that were analysed in-depth were federated infrastructures where data quality assessment is done at the point of collection by data holders. It is a common feature that the initiative provides guidelines and tools, but the assessment is decentralised, either at individual data holders or by a national/regional coordinator.

2. Data publication

Metadata catalogues are generally available, but in different formats. The entities/partners associated with an initiative can locate relevant assets, but a researcher working across communities will find it more difficult. There seems to be a need for generic metadata standards.

3. Data discovery

Data dictionaries or code books are not widely available, but an example can be found in the ECHO Data Model Specification

4. Data access

This step in the EHDS data lifecycle is not relevant in terms of data quality assessment.

5. Data use

Processing procedures are not publicly available for any of the initiatives analysed, except ECHO, where procedures are published on the echo-health.eu website.

6. Data analysis

None of the initiatives use auditable software, except OHDSI, that offer a number of publicly available open-source software packages for quality assurance and assessment. ECHO performed original data analysis using Stata12© statistical software. Scripts have not been published, but are available for auditing upon request to ECHO-Health project coordination.

7. Finalisation

ECHO and OHDSI archive digital objects using software solutions.

4.2.5 Legal aspects and governance of data quality management

The analysis has not uncovered any legal barriers to ensuring data quality.

Governance models are similar in all EU projects with a Project Coordinator, supported by a Steering Committee or Executive Board for the project/initiative. Advisory Boards or working groups are often set up to focus on specific activities or areas, e.g. data quality.

4.2.6 Technical services or tools

OHDSI has developed a suite of technical tool and services aimed at quality assurance, but the analysis has not identified any "industry standard" solutions that are used to assess the quality of health data.

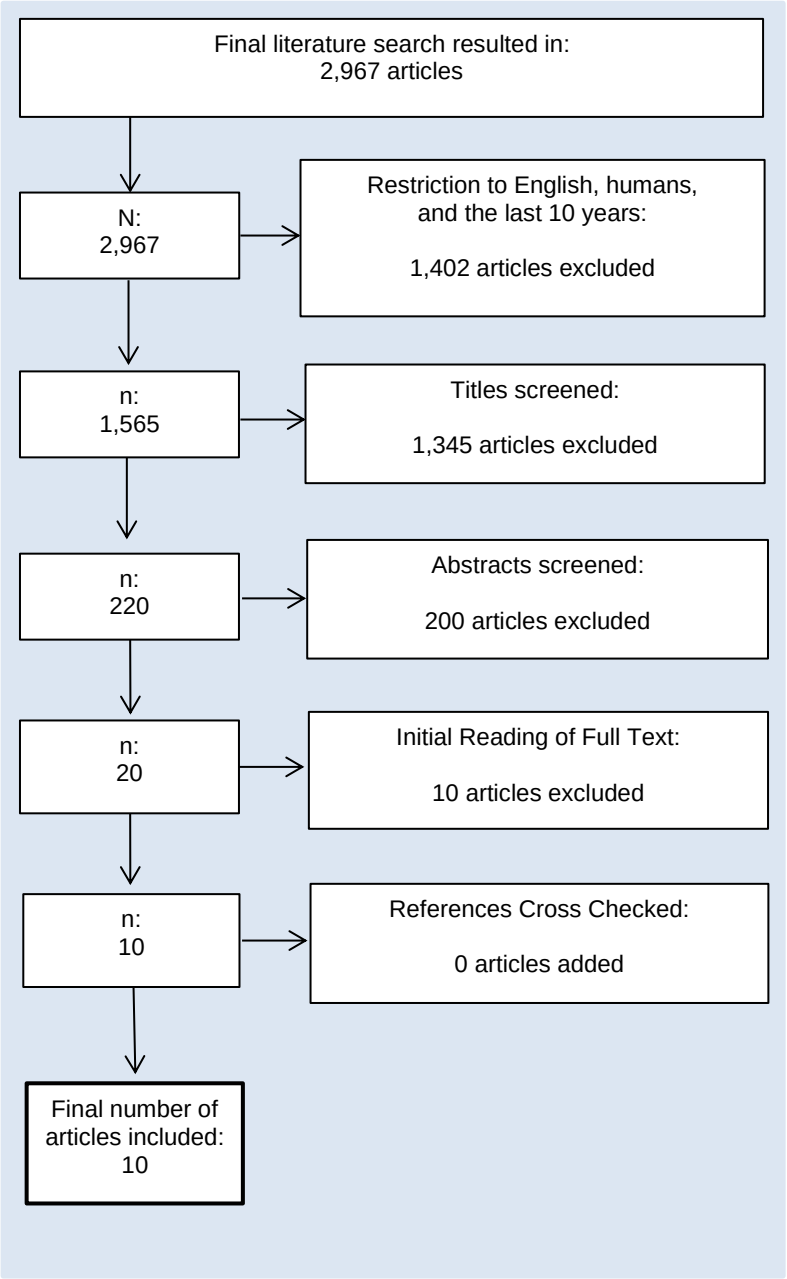
The Irish Health Information and Quality Authority offer e-learning modules on how to improve data quality for health and social care services.

None of the initiatives analysed has implemented methods for anonymisation or pseudonymisation.

4.3 Literature Scoping Review

The search for grey literature (Google Scholar) resulted in a total of 739 publications. Titles and abstracts (the latter when available) were screened and 10 publications were selected for further review. Of these, five publications were also found through the scientific databases search and the remaining five were excluded because they were not based on health data. Thus, the results presented in the following relates only to the search for peer reviewed scientific publications.

Figure 4: Flow Chart of the Literature Inclusion Process



The characteristics and finding of the included publications are summarized in table 5 below.

Table 5: Characteristics and Summary Findings of the Included Publications

First author, year, country	Domain (health area)	Study design	Secondary use data	Federated data	Data sharing	DQF dimensions	DQA methods	Recommendations / comments on DQF / DQA methodologies	Publication quality
Weiskopf et al.(2) June 2012 US	Generic	Literature review	Yes	No	No	▪ Completeness ▪ Correctness ▪ Concordance ▪ Plausibility ▪ Currency	▪ Gold standard ▪ Data element agreement ▪ Element presence ▪ Data source agreement ▪ Distribution comparison ▪ Validity checks ▪ Log review	▪ Consistency in discussion of dimensions. ▪ Development of validated systematic DQA methods.	Acceptable
Kahn et al.(3) July 2012 US	Generic	Conceptual research	Yes	No	Yes (multi-sites within the US)	***Intrinsic*** ▪ Accuracy ▪ Objectivity ▪ Believability ***Conceptual*** ▪ Timeliness ▪ Appropriate amount	▪ Attribute domain constraints ▪ Relational integrity rules ▪ Historical data rules ▪ State-dependent objects rules ▪ Attributes dependency rules (Specific methods to support the rules are not listed in this table).	▪ Conduct DQA within and across sites ▪ Consider the dimensions ▪ Use the suggested methods to assess the variability in multi-site datasets ▪ Do not jump straight to “stage 2” DQA if “stage 1” DQA is not performed	Acceptable
Brown et al.(4) Aug. 2013 US	Generic	Conceptual research	Yes	No	Yes (distributed data networks)	Builds on Kahn et al, 2012	▪ Review adherence to common data model - Syntactic consistency - Table structure	▪ Report DQA metrics and results	Acceptable

							- Expected relationships between tables ▪ Review each data domain (e.g., diagnoses or medication) separately and extend any model to also evaluate: - Frequency and proportions for categorical variables - Distributions and extreme values for continuous variables - Missingness - “Out-of-range” values - Expected relationships between variables within the domain - Normalized rates (e.g., per person) - Temporal trends (e.g., weekly or monthly) ▪ Assess expected clinical relationships		

Salati et al.(5) Dec. 2015 European	Thoracic surgery	Model development	Yes	No	Yes (multi-sites and cross-border)	No	DQ indicators: ▪ Site completeness (COM) ▪ Site reliability (REL) ▪ Rescaled COM ▪ Rescaled REL ▪ Summary measure aggregate data quality score (ADQ)	▪ Use analytic models during database management as well as before analytic studies.	Acceptable
Johnson et al.(6) Feb. 2016 US	Urinary catheter removal procedure	Model development	Yes	No	No	No	Ontology based DQA: ▪ References separate DQ domain- and task ontologies to compute measures based on proportions of constraints that are satisfied. ▪ These quantities indicate how well the data conforms to the domain and how well it fits the task.	▪ The advantage of the DQ ontology is that it provides a vocabulary for aspects of DQ and also defines a process to quantify it. ▪ Metrics can be shared, sites compared, and the DQ development followed over time. ▪ This particular model needs further validation	Acceptable
Reimar et al.(7) Oct. 2016 US	Patient transportation	Conceptual research	Yes	No	Yes (within one health care system)	Builds on Weiskopf et al, 2012 with a deeper evaluation of completeness and concordance through: ▪ preliminary analysis ▪ longitudinal concordance ▪ breadth ▪ data element presence	▪ Data element agreement ▪ Element presence ▪ Data source agreement ▪ Distribution comparison	▪ DQ metrics for benchmarking acceptable levels	Acceptable ↓ Methodology not clear and generalizability of framework not considered.

						▪ density			
Kahn et al.(8) Nov. 2016 US	Generic	Conceptual research (expert panels, literature review, and workshops)	Yes	No	Yes	A set of categories and sub-categories divided into a Verification and Validation context : ▪ Conformance - Value Conformance - Relational Conformance - Computational Conformance ▪ Completeness ▪ Plausibility - Uniqueness Plausibility - Atemporal Plausibility - Temporal Plausibility		▪ Standardized and validated methods for DQ are crucial ▪ The present intrinsic framework should be validated ▪ A harmonized operational framework that includes reusable DQA, visualization, and reporting tools is needed.	Acceptable
Sáez et al.(9) Feb. 2017 Spain	Generic (metrics tested on data from UCI heart disease data set)	Model development	Yes	No	Yes	No Relates the developed metrics to The concordance dimension, and the data source agreement and distribution comparison methods from Weiskopf et al, 2012.	Two metrics for the detection of undesired variability between data sources: ▪ the degree of global multi-source variability –(GPD) ▪ the degree of outlyingness of single sources –(SPO)	The stability metrics permit measuring the degree of data set concordance without requiring an additional gold standard data set.	Acceptable
Henley-Smith et al.(10) Aug. 2019 Australia	Generic (tested on GP data)	Conceptual research Model development	Yes	No	Yes	**Level 1** ▪ Revised Kahn et al, 2016 ▪ Added categories that reflect DQA in	Metrics developed to test: ▪ Level 1 – intrinsic features ▪ Level 2 – analytic implications	▪ Further development and test of the present DQF in a real world setting needed.	Acceptable

						data warehouse context **Level 2** ▪ Revised Kahn et al, 2016 ▪ Added categories that reflect DQA in research question context		▪ Ware-house data should be tested for fitness for **secondary** use.	
Liaw et al. (11) Jan. 2021 Australia	Generic	Literature review	Yes	No	Yes	▪ Kahn et al, 2016 - Added a contextual DQ category (Data organisation) with subcategories: - Timeliness - Trust - Relevance - Accessibility - Reusability - Governance ▪ Added a technical DQ category with subcategories: - Operating platform - Interoperability	Intrinsic, contextual and technical DQ indicators such as (not fully listed in this table): ▪ Reputation ▪ Missingness ▪ Reliability ▪ Applicability ▪ Common Data Model ▪ Fragmentation ▪ Traceability ▪ Data capture	▪ Comprehensive DQA requires a culture of reciprocity, transparency, and interoperability across the data production and curation life cycle. ▪ Effective DQ assessment is underpinned by rigorous documentation at point of care, good management, and appropriate governance across the RWD production and curation life cycle.	Acceptable

As shown in table 5 all ten included publications were rated to be of an acceptable quality. Six of the publications originated from the Unite States, all focused on secondary use of data and all but one on data sharing. No publication covering the use and sharing of federated data was found. Seven of the publications describes DQFs, and of these three are the original work of Weiskopf et al., 2012. Kahn et al. 2012, and Kahn et al. 2016, respectively, and the remaining four builds upon these original works. Nine of the publications includes methods for DQA. Three out of these publications present metrics and one an ontology based approach to specific quality measurement.

4.3.1 Knowledge Summary

Exact conclusions on which dimensions and measurement methodologies to be included in a DQF cannot be derived based on the included publications. However, the publications indicate consensus regarding the following: DQFs are vital for sharing of secondary use data; All aspects of a DQF must be clearly defined (e.g., the specific meaning of dimensions); DQFs should include intrinsic as well as contextual categories; DQA methods should be validated in a real-world setting; DQA should be conducted both within and across sites; and DQA results should be reported and followed over time. Additionally, the most recent publications points towards a shift in the focus from "fitness for purpose" to "fitness for secondary use" while emphasizing that DQFs need to take account of the full data life cycle.

The literature scoping review shows a lack of documented knowledge on secondary use of federated data and publications from a European setting are sparse. The experiences are few and time-limited. It can also be concluded that one size does not fit all.

5 Recommendations

The following recommendations are based on a synthesis of the results of each of the three methodological approaches explained in Figure 1. This includes recommendations on what should be considered by the European Commission as part of the continuing work with EHDS to promote the development of a DQF, including features that should be legally bound, based on knowledge generated from thematic workshops and collected from existing initiatives as well as from a literature scoping review. The recommendations suggest that we should not look for an existing, ready-made model to copy, but to draw inspiration from the articles and initiatives analysed in this report.

The recommendations are:

- The adoption of a working definition of data quality that focus on data "fitness for purpose" and how well data reflects the reality it represents.
- Reliability, relevance, timeliness, coherence, coverage and completeness should be adopted as measurable dimensions of data quality.
- Promote a focus on transparency at the level of institutions across Member States in relation to regular audits, a well-developed DQF and clear procedures in relation to processing the data.

- National competent institutions should audit data holder institutions on procedures of quality assurance and assessment and their data sets in accordance with the EHDS DQF.
- A data quality framework focusing on quality at the institutional level would rely heavily on transparency and auditing to ensure that quality standards are met. Implementing this framework at an institutional level would have the added benefit of tapping into the current local, regional, and national data collection and auditing systems, which makes the implementation of such a framework less complicated.
- As data requests may entail data linkage, data harmonization, and data transformation processes before delivery, data holders should be obligated to publish their data preparation procedures, metadata about their collections, including information on data provenance, relevance and coverage of the data collection and ensure the highest possible degree of transparency.
- Initiatives should focus on continuous improvement, encouraging good practice, design, development and implementation of toolkits for quality assessment and allocate resources to support data quality-focused work.
- In the medium to longer-term promote the development of a benchmarking process which will assist data managers and institutions with alignment against a Europe-wide approach to measuring data quality.

6 References and documentation

6.1 Background Materials and Documentation

Data quality review: a toolkit for facility data quality assessment. Module 1. Framework and metrics, World Health Organization 2017

ISO Standard 25012 - Data Quality model
https://iso25000.com/index.php/en/iso-25000-standards/iso-25012

Quality Assurance Framework of the European Statistical System, version 2.0, 2019
https://ec.europa.eu/eurostat/web/quality

Handbook on Data Quality Assessment Methods and Tools, European Commission EUROSTAT, Manfred Ehling and Thomas Körner (eds.), Wiesbaden 2007.

Data Utility Framework, Health Data UK (HDRUK)
https://www.hdruk.ac.uk/helping-with-health-data/ways-to-improve-data-quality/data-utility-evaluation/

UK Statistics Authority
https://osr.statisticsauthority.gov.uk/guidance/administrative-data-and-official-statistics/

- *Quality Assurance of Administrative Data: Setting the Standard* (January 2015)
- *Administrative Data Quality Assurance Toolkit* (February 2019)

Background paper to support guidance for a data quality framework for health and social care data collections, Health Information and Quality Authority, Ireland 2018
https://www.hiqa.ie/sites/default/files/2018-10/Background-to-support-guidance-on-data-quality-framework.pdf

The Challenges of Data Quality and Data Quality Assessment in the Big Data Era, Cai, L and Zhu, Y 2015, Data Science Journal, 14: 2, pp. 1-10, DOI:
http://dx.doi.org/10.5334/dsj-2015-002

DCAT https://www.w3.org/TR/2016/NOTE-vocab-dqv-20161215/

Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics

Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)

"Shaping Europe's digital future" (Communication from the Commission to the European Parliament, The Council, The European Economic and Social Committee and the Committee of the Regions), Brussels 19.2.2020

"Towards a common European data space" (Communication from the Commission to the European Parliament, The Council, The European Economic and Social Committee and the Committee of the Regions), Brussels 25.4.2018

Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act), Brussels 25.11.2020

Impact assessment report accompanying the proposal for a Data Governance Act (Commission staff working document), Brussels 25.11.2020

Assessment of the EU Member States' rules on health data in the light of GDPR (NIVEL Report), European Commission 2021

BBMRI-ERIC Quality Policy: Standardisation
https://www.bbmri-eric.eu/services/standardisation/

Process Flow: Q-Assessment Scheme for Biobanks and Sample Collections, 09.03.2021
https://www.bbmri-eric.eu/wp-content/uploads/Q-Assessment_Scheme_for_Biobanks_and_Sample_Collections_web.pdf

Access principles to BBMRI-ERIC self-assessment surveys (BBMRI-ERIC SAS)
https://www.bbmri-eric.eu/wp-content/uploads/Access_principles_BBMRI-ERIC_SAS.pdf

BBMRI-ERIC Quality Policy: Standardisation
https://www.bbmri-eric.eu/services/standardisation/

The Challenges of Data Quality and Data Quality Assessment in the Big Data Era, Cai, L and Zhu, Y 2015, Data Science Journal, 14: 2, pp. 1-10, DOI: http://dx.doi.org/10.5334/dsj-2015-002

6.2 Literature Scoping Review References

1. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med. 2018;169(7):467-73.

2. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013;20(1):144-51.

3. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care. 2012;50 Suppl(0):S21-9.

4. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. Med Care. 2013;51(8 Suppl 3):S22-9.

5. Salati M, Falcoz PE, Decaluwe H, Rocco G, Van Raemdonck D, Varela G, et al. The European thoracic data quality project: An Aggregate Data Quality score to measure the quality of international multi-institutional databases. Eur J Cardiothorac Surg. 2016;49(5):1470-5.

6. Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. Application of An Ontology for Characterizing Data Quality For a Secondary Use of EHR Data. Appl Clin Inform. 2016;7(1):69-88.

7. Reimer AP, Milinovich A, Madigan EA. Data quality assessment framework to assess electronic medical record data for use in research. Int J Med Inform. 2016;90:40-7.

8. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). 2016;4(1):1244.

9. Sáez C, Robles M, García-Gómez JM. Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances. Stat Methods Med Res. 2017;26(1):312-36.

10. Henley-Smith S, Boyle D, Gray K. Improving a Secondary Use Health Data Warehouse: Proposing a Multi-Level Data Quality Framework. EGEMS (Wash DC). 2019;7(1):38.

11. Liaw ST, Guo JGN, Ansari S, Jonnagaddala J, Godinho MA, Borelli AJ, et al. Quality assessment of real-world data repositories across the data life cycle: A literature review. J Am Med Inform Assoc. 2021;28(7):1591-9.

Annex

Annex Table 1: Search Words, Phrases, and related MeSH-terms

Free text search using quoted phrases	
"Data quality assessment"	
"Data quality assessment framework"	
“Secondary us of health data”	
Block search using free text and MeSH terms	
Search Words and Phrases	**Related MeSH-terms**
What is the type of data?	
Data	"Data Science"[Mesh] "Electronic Data Processing"[Mesh]
Health data	"Health Information Systems"[Mesh] "Medical Informatics"[Mesh] "Health Information Exchange"[Mesh] "Informatics"[Mesh]
Dataset	"Datasets as Topic"[Mesh] "Common Data Elements"[Mesh]
Database	"Databases as Topic"[Mesh] "Database Management Systems"[Mesh]
Federated data	No MeSH-terms
Aggregated data	No MeSH-terms
Metadata	"Metadata"[Mesh]
Big data	"Big Data"[Mesh]
What is the purpose with the data?	
Data sharing	"Information Dissemination"[Mesh]
Data flow	No MeSH-terms
Secondary use data	No MeSH-terms
What to know in relation to data?	
Quality	"Quality Control"[Mesh] "Quality Improvement"[Mesh] "Quality Assurance, Health Care"[Mesh] "Quality Indicators, Health Care"[Mesh] "Benchmarking"[Mesh]
Data quality	No MeSH-terms
Framework	No MeSH-terms
Assessment	No MeSH-terms
Evaluation	"Program Evaluation"[Mesh]
Monitoring	No MeSH-terms
Data quality framework	No MeSH-terms

Data quality assessment	No MeSH-terms
Data quality evaluation	No MeSH-terms
Data quality monitoring	No MeSH-terms
Data quality profile	No MeSH-terms
What to know in relation to quality?	
Governance	No MeSH-terms
Data governance	No MeSH-terms
Data quality definition	No MeSH-terms
Data quality features	No MeSH-terms
Data quality metrics	No MeSH-terms
Data quality indicators	No MeSH-terms
Data quality dimensions	No MesH-terms
Data quality score	No MesH-terms
Data quality rules	No MesH-terms
Data access	No MesH-terms
Data collection	"Data Collection"[Mesh] "Data Management"[Mesh] "Data Curation"[Mesh] "Data Warehousing"[Mesh]
Data processing	"Electronic Data Processing"[Mesh]
Data management	"Data Management"[Mesh]
Missing data	No MeSH-terms
Data completeness	No MeSH-terms
Heterogeneity	No MeSH-terms
Data heterogeneity	No MeSH-terms
Anonym	"Anonyms and Pseudonyms"[Mesh]
Anonymization	"Data Anonymization"[Mesh]
Data anonymization	"Data Anonymization"[Mesh]
Pseudonym	"Anonyms and Pseudonyms"[Mesh]
Pseudonymization	No MeSH-terms
Data pseudonymization	No MeSH-terms
Timeliness	No MeSH-terms
Accuracy	"Dimensional Measurement Accuracy"[Mesh]
Data accuracy	"Data Accuracy"[Mesh]
Validity	"Reproducibility of Results"[Mesh] "Program Evaluation"[Mesh]
Legislation	"Legislation as Topic"[Mesh]
Legal	"Liability, Legal"[Mesh] "Legislation as Topic"[Mesh]
FAIR principles	

Annex Table 2: Search Strategy for Grey Literature using the Google Scholar Database

("data quality management framework" OR "data quality assurance systems" OR "data quality assessment framework") AND ("health" OR "health information systems" OR "healthcare data")
("data quality management framework" OR "data quality assurance systems" OR "data quality assessment framework") AND ("health" OR "health information systems" OR "healthcare data") AND filetype:html
allintitle:"data quality management framework" OR "data quality assurance systems" OR "data quality assessment framework"

Annex Table 3: Search Strategy for Scientific Databases Exemplified by the PubMed/MEDLINE database search

((("data quality assessment") OR ("data quality assessment framework")) OR (("data quality assessment"[All Fields] OR "data quality assessment framework"[All Fields]) AND "secondary use of health data"[All Fields])) **OR** ((((((((((((((((((((data) OR ("Data Science"[Mesh])) OR ("Electronic Data Processing"[Mesh])) OR ("Health Information Systems"[Mesh])) OR ("Medical Informatics"[Mesh])) OR ("Health Information Ex-change"[Mesh])) OR ("Informatics"[Mesh])) OR ("health data")) OR (dataset)) OR (dataset)) OR ("Datasets as Topic"[Mesh])) OR ("Common Data Elements"[Mesh])) OR ("database")) OR ("Databases as Topic"[Mesh])) OR ("Database Management Systems"[Mesh])) OR ("federated data")) OR ("aggregated data")) OR ("meta da-ta")) OR ("Metadata"[Mesh])) OR ("big data")) OR ("Big Data"[Mesh])) **AND** ((("data sharing") OR ("Information Dissemination"[Mesh])) OR ("data flow")) OR ("second-ary use data")) **AND** ((((((((((("Quality Control"[Mesh]) OR ("Quality Improve-ment"[Mesh]))) OR ("Quality Assurance, Health Care"[Mesh])) OR ("Quality Indica-tors, Health Care"[Mesh])) OR ("Benchmarking"[Mesh])) OR ("data quality")) OR ((((((("quality") **AND** ("framework")) OR ("assessment")) OR ("evaluation")) OR ("Program Evaluation"[Mesh])) OR ("monitoring"))) OR ("data quality framework")) OR ("data quality assessment")) OR ("data quality evaluation")) OR ("data quality monitoring") **AND** ((((((((((((((((("governance" OR "data governance") OR ("data quality features")) OR ("data quality metrics")) OR ("data quality indicators")) OR ("data quality dimensions")) OR ("data quality score")) OR ("data quality rules")) OR ("data access")) OR ("data collection" OR "Data Collection"[Mesh] OR "Data Man-agement"[Mesh] OR "Data Curation"[Mesh] OR "Data Warehousing"[Mesh])) OR ("data processing" OR "Electronic Data Processing"[Mesh])) OR ("data manage-ment" OR "Data Management"[Mesh])) OR ("missing data")) OR ("data complete-ness")) OR ("heterogeneity" OR "data heterogeneity")) OR ("anonym" OR "Ano-nyms and Pseudonyms"[Mesh] OR "data anonymization" OR "Data Anonymiza-tion"[Mesh] OR "pseudonym" OR "pseudonymization" OR "data pseudonymiza-tion")) OR ("timeliness")) OR ("validity" OR "Reproducibility of Results"[Mesh] OR "Program Evaluation"[Mesh])) OR ("legislation" OR "Legislation as Topic"[Mesh] OR "legal" OR "Liability, Legal"[Mesh])) OR ("fair principles")))

Annex Table 4: Data Extraction Template

Study Characteristics	
Author:	N. G. Weiskopf and C. Weng
Title:	Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research.
Journal:	J Am Med Inform Assoc
Year:	2012
Country:	US
Study type:	Literature review
Purpose:	To review the methods and dimensions of data quality assessment in the context of electronic health record (EHR) data reuse for research.
Methods:	Literature review – close to fully systematic
Context relevant to TEHDAS	
Health data (domain):	Generic
Secondary use:	Yes
Federated data:	No
Data sharing:	No
DQF:	Yes
DQA methods:	Yes
If DQF – state dimensions:	Completeness Correctness Concordance Plausibility Currency
If DQA methods – state type:	Gold standard Data element agreement Element presence Data source agreement Distribution comparison Validity checks Log review
DQF /DQA methods recommendations:	The clinical research community needs to develop validated, systematic methods of EHR data quality assessment. We encourage researchers to be consistent in their discussion of the dimensions of data quality, systematic in their approaches to measuring data quality, and to develop and share best practices for the assessment of EHR data quality in the context of reuse for clinical research.
Quality assessment	
Compliance with TEHDAS questions	Moderate
Clear methodology	Lacking for research question prior to search, data extraction procedure, and quality assessment of study**.**
Agreement between purpose, methods, and results	Overall good
Peer review	Yes
Impact factor	Apr. 3.9 in 2013
Overall quality	Acceptable

Annex Table 5: Overview of Literature Excluded after Initial Reading of Full Text

	Publication	Reason for exclusion
1	Bouzillé G, Westerlynck R, Defossez G, Bouslimi D, Bayat S, Riou C, et al. Sharing Health Big Data for Research - A Design by Use Cases: The INSHARE Platform Approach. Stud Health Technol Inform. 2017;245:303-7.	Describes considerations and steps taken to build a platform for sharing of health data. No specific recommendations regarding DQF.
2	Dyke SOM, Linden M, Lappalainen I, De Argila JR, Carey K, Lloyd D, et al. Registered access: authorizing data access. Eur J Hum Genet. 2018;26(12):1721-31.	Addresses issues regarding access to shared data, not DQ.
3	Endler G, Schwab PK, Wahl AM, Tenschert J, Lenz R. An Architecture for Continuous Data Quality Monitoring in Medical Centers. Stud Health Technol Inform. 2015;216:852-6.	Does not consider DQ in relation to data sharing and/or secondary use.
4	Khatami R, Luca G, Baumann CR, Bassetti CL, Bruni O, Canellas F, et al. The European Narcolepsy Network (EU-NN) database. J Sleep Res. 2016;25(3):356-64.	Describes a European database, but does not cover DQF sufficiently to answer our questions.
5	Laberge M, Shachak A. Developing a tool to assess the quality of socio-demographic data in community health centres. Appl Clin Inform. 2013;4(1):1-11.	Well described DQ in relation to fitness for purpose in a primary use context. Does not consider DQ in relation to secondary use and/or data sharing.
6	McDonald SA, Mardis ER, Ota D, Watson MA, Pfeifer JD, Green JM. Comprehensive genomic studies: emerging regulatory, strategic, and quality assurance challenges for biorepositories. Am J Clin Pathol. 2012;138(1):31-41.	Does not address DQF.
7	Rahimzadeh V, Dyke SO, Knoppers BM. An International Framework for Data Sharing: Moving Forward with the Global Alliance for Genomics and Health. Biopreserv Biobank. 2016;14(3):256-9.	Does not address DQF.
8	Sarafidis M, Tarousi M, Anastasiou A, Pitoglou S, Lampoukas E, Spetsarias A, et al. Data Quality Challenges in a Learning Health System. Stud Health Technol Inform. 2020;270:143-7.	Provides a narrative review of DQA in healthcare and presents a cloud platform. This platform provides a QA module based om ML methods and the framework of Weiskopf et al. However, it is unclear if this article represents intended work or if the platform is in fact up and running, and if so how the models used are tested etc.
9	Schmidt BM, Colvin CJ, Hohlfeld A, Leon N. Definitions, components and processes of data harmonisation in healthcare: a scoping	Addresses data harmonisation definitions and does not include DQ aspects.

	review. BMC Med Inform Decis Mak. 2020;20(1):222.	
10	Scobie HM, Edelstein M, Nicol E, Morice A, Rahimi N, MacDonald NE, et al. Improving the quality and use of immunization and surveillance data: Summary report of the Working Group of the Strategic Advisory Group of Experts on Immunization. Vaccine. 2020;38(46):7183-97.	Touches only briefly the aspect of DQF as one of nine recommendation points for vaccine surveillance. No specifics regarding the content of a DQF.

Annex Table 6: Crude list of initiatives

ECIS (European Cancer Information System)
EU RD (European Platform on Rare Disease Registration)
GA4GH (Global Alliance for Genomics and Health)
OECD Health
Closer (Home of longitudinal research)
CONCORD (Global Surveillance of Cancer Survival at LSHTM)
NWB (Neurodata Without Borders)
NHGRI (National Human Genome Research Institute)
X-eHealth
IRDiRC (International Rare Diseases Research Consortium)
INSPIRE
TriNetX
WHO-GCO (Global Cancer Observatory)
DO>IT
ROADMAP
GAIA-X
NordForsk
ByCovid
INCF (NeuroScience)
INFACT
OpenEHR
PARENT – EUnetHTA
EATRIS (European Infrastructure for Translational Medicine)
EUROCARE (Survival of cancer patients in Europe)
EuroHOPE (European Health Care Outcomes, Performance and Efficiency)
PHIRI (European Health Information Portal)
Research Data Alliance
1+M Genomes
ECHO – ECHOAtlas (European Collaboration for Health Optimization)
HRIC (Health Research and Innovation Cloud)
OHDSI
Orphanet
BBMRI (European Research Infrastructure for Biobanking)
HARMONYplus
ECRIN (European Clinical Research Infrastructure Network)
BigData@Heart

CDISC
PIONEER
EJP RD (The European Joint Program on Rare Diseases)
ELIXIR
EOSC (European Open Science Cloud)
HBM4EU (European HBM Platform)
DARWIN EU
EHDEN
BRIDGE
GHDx (Global Health Data Exchange)
GO FAIR
HEALTHYCLOUD (European Health and Innovation Cloud)