Milestone 6.1

Identifying those data quality features that could be legally bound and providing advice to the European Commission

11 November 2021

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)
Persephone Doupi	Finnish Institute for Health and Welfare (THL)
Thomas Engsig-Karup	Central Denmark Region (RM)
Francisco Estupiñán-Romero	Aragon Institute of Health Sciences (IACS)
Ramón Launa-Garces	Aragon Institute of Health Sciences (IACS)
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 October 2021.

Disclaimer
The content of this deliverable represents the views of the author(s) only and is his/her/their sole responsibility; it cannot be considered to reflect the views of the European Commission and/or the Consumers, Health, Agriculture and Food Executive Agency or any other body of the European Union. The European Commission and the Agency do not accept any responsibility for use of its contents.

Contents

1 Executive summary

The purpose of this milestone document is related to the overall work in the TEHDAS Project Work Package (WP) 6: Excellence in data quality. The aim is to provide solutions for the trustworthy secondary use of health and health care data with a view to promoting the digital transformation of European health systems.

The findings and recommendations of this document are a result of a combination of literature studies, working groups and bilateral meetings with partners and stakeholders. Concretely, a number of Thematic Working Groups were formed, and especially the Thematic Working Group on "Data Quality Legislation" has provided input for this milestone document.

The Thematic Working Group on "Data Quality Legislation" was asked to deal with three questions. These were later discussed with the broader group to achieve consensus on the recommendations.

The first questions asked was "What aspects of the European Health Data Space (EHDS) should be regulated by law in terms of data quality?". Consensus was reached that the most important aspect of data quality is reliability, and this is best achieved by legislation on the data quality assurance processes. As the quality of a data set is relative to its purpose or use, the focus should be on the quality assurance performed by the different actors in the process, such as data collectors, data holders etc. It should be ensured that resources are not shifted from patient care to quality assurance aimed at secondary use of health data.

The second question posed to the thematic work group on data quality legislation was "In terms of data quality, what is the best approach to encourage coherence across member states in the implementation of new legislation, codes of conduct, best practice guidelines or other?". The recommendation is that it would be beneficial to establish an EHDS quality assurance governance structure (i.e. body, service, unit, committee) that could develop and implement the EU wide guidelines on data quality assurance. The preferred model of the EHDS is a federated model of national bodies responsible in each Member State.

The third question posed to the thematic work group on data quality legislation was "What are the boundaries of national legislation/regulation in terms of data quality?". The recommendations are that the quality dimensions of health care data should be defined at EU level as part of the EHDS Data Quality Assurance Framework and implemented on a national level by national authorities.

In addition, the milestone document has been tasked with looking at existing EU legislation regulating data quality, such as the European Statistical System (ESS) and INSPIRE to see whether some of the same structures could be used for the EHDS. The WP6 group broadly agree that the committees set up in both initiatives, with the view to assist the European Commission, could serve as inspiration for the establishment of an EHDS quality assurance governance structure (i.e. body, service, unit, committee).

Based on an analysis of the data user's "user journey", a number of steps have been identified where legal enforcement may be advisable in the context of data quality assurance. The features, which this document recommend being legally bound, are audits, processing procedures and meta data catalogues.

2 Context

This TEHDAS Milestone 6.1 document will identify features of an EHDS Data Quality Assurance Framework (DQAF) that could be legally bound. These features cannot be viewed isolated from the overall work in Work Package 6 or the TEHDAS Joint Action as a whole. The recommendations made in this document are based on conclusions drawn from a number of working meetings between the partners participating in work package 6. The dialogue and discussions will continue for the duration of the project period and, as such, the document represents a milestone in a continued process that will develop an EHDS Data Quality Assurance Framework. Going forward, the arguments will be refined, and the work will elaborate on the design of the DQAF.

It is important to underline some of the premises for developing a European DQAF within the healthcare domain. The recommendations in this document are made in a specific context that requires attention in order to understand the underlying rationale.

The first point to note is that collection, use and storage of healthcare data is organised differently across Member States. National healthcare sectors are organised differently across Europe and this is reflected in the way data is managed. The different setup of data sharing initiatives within the same Member States further increase the complexity. This makes it difficult to compare data between data sharing initiatives and between the Member States.

The second point is that data quality is multidimensional. Quality is relative to the need of the user and a particular data set can meet the quality requirements of one user, but not of the other. This will be explained in greater detail in part 3 on Data Quality. The implication is that you cannot apply generic metrics for quality measurements directly to the data sets.

The third point to keep in mind is that the task is to recommend a DQAF which can accommodate all relevant institutions in the 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. At the same time, the DQAF should ensure that the data quality is high enough for the data to be relevant for secondary use.

A fourth point to bear in mind when reading the document at hand, is that all references and recommendations on governance and legal matters in this document must be seen as relating to data quality and the introduction of a DQAF, which differentiates this from the overall work on legislation and governance done in TEHDAS Work Package 5: Sharing data for health.

The fifth and last point to note is that this document is not addressing the costs associated with the recommendations. These aspects are treated elsewhere in the TEHDAS JA Project.

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.

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

Task 6.1 will deliver on the first operational objective – the development of the EHDS DQAF - and the output of task 6.1 will contain two types of contents: 1) The requisites for the inclusion of data in the EHDS, including identification of those features that could be legally bound, and 2) guidance on how to get data included in the EHDS and recommendations for data quality governance.

The present document (Milestone 6.1) will identify features that could be legally bound and provide advice to the European Commission on the topic of legislation in the context of data quality assurance within the EHDS, specifically with regard to secondary use of health data. As such, the document represents a milestone and a subtask in the overall deliverable from Work Package 6.

3 Methodology

3.1 Thematic Work Groups

A number of thematic work groups has been established in order to break down the work into smaller pieces and groups. The aim is to focus the discussions and go deeper into the subject matter. The first thematic work group on “Data Quality Legislation” took place on 21 June 2021 and the purpose of this meeting was to collect input from a group of partners in Work Packages 6 and 5, who volunteered to join the discussions. In addition, the first thematic working group on “Data Quality Governance” took place on 13 September 2021. The input from these meetings and subsequent presentation and feedback from the wider group of partners in WP6 provides the main content of this milestone document.

3.2 Literature review

A review of relevant articles, reports and legislation was conducted. A list of documents is included in the references section.

To gain insight into the data quality aspects that could be legally bound, we searched the literature for (a) EU legislation relating to data quality (b) assessment methods, procedures and tools. Literature was considered relevant if it described the dimensions of data quality and the procedures for the control and the assurance of data quality, through all phases of data collection and utilization. Reviewed literature is included in the References section of this Milestone.

4 Definition of data quality

Work Package 6 will select or develop a definition of data quality to be used in the context of the EHDS DQAF, including which dimensions to use for quality assessment, as part of the work package deliverables. The definition of data quality found in ISO 25012 is used here to illustrate the concept of a data quality model.

ISO 25012 - Data Quality model - defines a general data quality model for data retained in a structured format within a computer system. It focuses on the quality of the data as part of a computer system and defines quality characteristics for target data used by humans and systems:

> *Data quality refers to the degree to which characteristics of data satisfy stated and implied needs when data is used under specified conditions.*

Moreover, data is deemed of high quality if it correctly represents the real-world construct to which it refers, in the way it has been designed to represent it.

There are many other definitions of data quality and most focus on utility or ‘fitness for use’.

The reason is that data regarded as useful for one purpose, and therefore perceived to be of high quality, can be useless for other purposes/needs and therefore perceived to be of low quality.

Nonetheless, it is relevant to speak about data quality in the context of sharing and delivering data sets without knowing exactly what the data requester will use the data for. It is important in the data discovery phase that the quality of data is known to the requester; not in absolute terms, but rather as measurements on a number of dimensions.

The Data Quality model defined in ISO 25012 is composed of 15 characteristics covering all types of data sets. The following 4 characteristics/dimensions are examples:

- *Completeness*: The degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use.
- *Accuracy*: The degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use. It has two main aspects:
    - Syntactic Accuracy: Syntactic accuracy is defined as the closeness of the data values to a set of values defined in a domain considered syntactically correct.
    - Semantic Accuracy: Semantic accuracy is defined as the closeness of the data values to a set of values defined in a domain considered semantically correct.
- *Currentness*: The degree to which data has attributes that are of the right age in a specific context of use. (This dimension is, outside the ISO 25012, mainly referred to as "Timeliness").

- *Portability*: The degree to which data has attributes that enable it to be installed, replaced or moved from one system to another preserving the existing quality in a specific context of use.

These dimensions have broad relevance and are similar to the dimensions mentioned as examples in the TEHDAS Project Plan: Accuracy, coherence, timeliness, consistency, trustfulness.

It is important that the data holder can describe the dimensions without input from data requesters. The highest possible degree of transparency is encouraged.

The ISO 25012 examples can be applied to all data sets across different domains. There may be other dimensions that are more relevant considering the specificities of health data sets. These have been explored in TEHDAS Work Package 5, Milestone 5.7, section 5.3 (Element 3: Research and specific data types) and will not be explored further here.

In TEHDAS, Data quality assurance has to be understood as a process to ensure the reliability of data (health and other sectoral data) for trustful reuse in policy making, regulation and research. Work Package 6 will take this perspective as a point of departure when developing an EHDS specific definition of data quality.

5 What aspects of the EHDS should be regulated by law in terms of data quality?

The first question posed to the thematic work group on data quality legislation was "What aspects of the EHDS should be regulated by law in terms of data quality?" There are many options to consider, such as methodology of quality assurance, data quality level or quality assurance bodies at EU or national level.

There is a high level of variation among Member States with respect to national legislation and interpretation of GDPR and differences on how we see the possibilities of GDPR. This reflects the diversity of the national health care systems in the Member States and the different models for treatment and reimbursement, which, in turn, determines the data collection processes.

It is easier to regulate access to data since the basic premise is clear: That everyone has the right to the protection of his/her personal data. The quality of data, on the other hand, is relative to the needs of the user, the researcher or the policy maker.

In terms of access to data for healthcare professionals there are significant differences between primary use (patient care and treatment), where broad, immediate and easy access is supported only for professionals involved in the individual care of the patient, and secondary use, e.g., clinical research or policy making, where permission to access health data can be complex and difficult to obtain.

5.1 Primary and secondary use

In the context of patient treatment, health data is collected for the purpose of diagnostics, provision of care etc. It is not an option to compromise the quality of patient treatment to improve the quality of data for secondary use. Care must be taken not to introduce standards, policies or guidelines that aim to improve secondary use of data if it negatively impacts patient treatment. Data recording/registering at the point of care should be as clinically relevant and meaningful as possible. For this reason, the range of regulation proposed is rather limited.

With regards to streamlining standards in terms of primary use of health data and secondary use of health data it should be an aspiration, not something to be enforced. Secondary use is another layer compared to primary use and secondary use across borders is yet another layer. The deciding factors should be what most efficiently supports each use. Care of patients will have different efficient solutions compared to secondary use (research, regulation or policy making). The primary objective is taking care of the patient and that takes precedence.

There are strong incentives for recording health data in a timely and accurate fashion in the healthcare sector, for example patient safety, planning and logistics, reimbursement etc. These are primary incentives. Fitness for secondary use is a secondary incentive.

The implication is that the cost of implementing data quality procedures *specific to the requirements of EHDS* should be covered by the stakeholders that benefits directly from the secondary use.

Health data is often understood as any personal data generated within healthcare systems (NIVEL Report, p. 14), but it should be noted that health data could stem from a number of sources. Data holders connected via EHDS nodes will hold different types of health data for secondary use, such as socio-economic data, that are not necessarily collected from patients' medical records – whether electronic or analogue.

5.2 Good practice (recommendation) vs. mandatory practice (legislation)

Legislation should be used to ensure that data quality assurance is achieved and maintained by the different actors in the data quality process.

Specific quality aspects of secondary use need to be regulated to ensure every Member State has an easy way to understand the standards that are being discussed, and how they can comply with them at national level.

Having strong regulatory tools to enforce data quality is paramount to making sure we advance as fast as we can at the pace of technological development. We need to push the data quality agenda forward, because data quality is more important than ever in the big data era with AI and the databases we use for public health and other purposes. This underlines the need for strong regulation on data quality and making sure data quality counts; also when speaking about data governance and legislation.

An important discussion is whether stronger regulatory measures are needed for data quality assurance, which most often rely on recommendations. In the areas where data is shared at EU level, there is an urgent need to move forward with a focus on the role of legislation on data quality.

Legislation on data quality does not mean regulation of every single semantic standard or other types of standards for every health information system. The digitalisation of medical records varies across Europe and the development/implementation of electronic medical records are closely aligned to local clinical workflows. Furthermore, the system design reflects the organisation of the hospital sector at national level. These things combined means that quality assurance is easier to apply to national registries, research databases, other health data repositories as well as other cross-sectoral data – socio-economic, education, environment, etc. that are established in the first place for the purpose of secondary use of health data.

We are putting quality at the forefront and we will, either at EU level or at national level, define the steps to have the appropriate legal framework and governance bodies. At the same time, special care must be taken not to interfere with the requirements and practices of patient care and treatment.

5.3 Reliability of data

There is general agreement that data reliability is one of the most important dimensions of data quality. Data should be true to the reality we are trying to capture and, to that end, the GDPR has harmonised national regulations to enable anyone to have personal data corrected, assuming there is proof of error.

To ensure reliability you must have a policy in place that allows you to look for errors and have the procedures to locate them and correct them etc. The reliability of data that the health system or health information system is capturing should be a cornerstone in a data quality assurance framework. Legally bound procedures and governance bodies should build on that. We have to ensure that data is reliable within the system.

Establishing a framework on data quality assurance requires institutions providing the EHDS with sources of data to have an established quality assurance policy and must have obligatory quality assurance procedures running. It is also a prerequisite to agree on certain interoperability standards. Using the EHDS requires a quality assurance analysis that is common to all the institutions. The methodology itself does not need to be enforced by law, but the establishment of a data quality assurance body within the governance framework of the EHDS should be legally bound.

The interconnectedness with the major data governance framework established under the EHDS is yet to be defined, as well as the relation to national authorities and nodes in the EHDS.

6 Coherent legislation across Member States

The second question posed to the thematic work group on data quality legislation was "In terms of data quality, what is the best approach to encourage coherence across member states in the implementation of new legislation, codes of conduct, best practice guidelines or other?"

There is a need for having a mixed approach and to use various mechanisms. A combination of new legislation, codes of conduct, best practice guidelines balancing legislation and recommendation approaches will be needed.

Data sharing organisations should have their own data quality assurance framework (i.e. data management policies, procedures, and guidelines) in place to ensure the reliability of the health data. The creation of an EU data quality assurance body needs to take into account what it is already in the EU arena, in order to limit the complexity in governance structure and bureaucracy.

The NIVEL report shows a preference for setting up “an EU level infrastructure to support access to data for secondary data use purposes" (NIVEL report p. 143). Based on this, the concept of an EU level Data Permit Authority (DPA) has been and will continue to be discussed with the view to present recommendations in the final Deliverable 6.1, in particular when discussing potential governance structures of the DQAF.

The Thematic Work Group on data quality legislation agrees that a dedicated EHDS quality assurance body should be established, and this body should be responsible for developing and implementing guidelines on data quality assurance. No further conclusions are drawn regarding the mandate.

On the linkage of the EHDS quality assurance body with national bodies, the preferred model is a federated one: An aligned network of national bodies responsible within each country.

7 Boundaries of national legislation on data quality

The third question posed to the thematic work group on data quality legislation was "What are the boundaries of national legislation/regulation in terms of data quality?"

There is a need for a solution embracing everyone, since all Member States will be part of the EHDS. Cross border collaboration, completeness, and accuracy of data needs to be addressed at the national level.

GDPR leaves a lot of room for diverging legislation at national level. However, the GDPR boundaries are boundaries on data protection, not on data quality. At the same time, forthcoming EHDS legislation may not solve technical issues (e.g., interoperability), but can facilitate the establishment of a governance framework that includes expert groups with a focus on improving data quality.

Cross border collaboration depends on policies regarding data quality elements, such as completeness and accuracy, to be implemented at national level to ensure the quality of the sources. Data holders should also be required to report the quality of the data; both from a clinical perspective and a research perspective.

For each database, each data sharing institution should ensure the technical and semantic interoperability and data quality information (meta data) needed to decide whether that database can be of use for their intended purpose.

This meta data should be a requisite for entering a database in the EHDS.

8 Assessment of existing EU legislation on data quality

8.1 European Statistical System (ESS)

At EU level, Eurostat is currently the only office that has legislative power to regulate the area of data quality. It has taken many years to establish this system. Even within a uniform framework of data quality assessment, the way that data is collected is different and the selection will be biased. It is crucial to consider that the legal framework around the EHDS also needs to take such aspects into account.

The regulation of the European Statistical System (ESS) regarding aspects such as the independence of the producer, data quality and basic rules that need to be observed has been considered. ESS has reached a high level of maturity in terms of data quality via legislation and adherence to guidelines on submitting data in a particular way.

The level of detail in the specific design of an EHDS DQAF with indicators and parameters, however, is not yet clear. There are different expectations, but there is general agreement that the specific design of the EHDS DQAF should not be determined by legislation.

The ESS Committee established through regulation (EC) No. 223/2009 of 11 March 2009 on European statistics could serve as inspiration. A Committee established in the context of EHDS, that provides professional guidance on health data quality assurance, would support best practice and facilitate collaboration between Member States.

For the same reason, a similar proposal on an EU-level data permit authority is put forward in the NIVEL Report (Assessment of the EU Member States' rules on health data in the light of GDPR, 2021):

> "European level legislation would have the distinct advantage of building a robust and transparent governance structure, which could be supported at EU level to ensure strengthened cooperation between Member States" (p. 134).

8.2 Infrastructure for Spatial Information in the European Community (INSPIRE)

The INSPIRE Directive (2007/2/EC of 14 March 2007) is an example of EU legislation that establishes a framework for managing data sets in a specific area; namely spatial information. The directive specifies what information should be included in metadata and requires Member States to establish a web site with specific services (the Inspire geo-portal).

Interoperability of data sets and services is ensured by adopting implementing rules covering 34 spatial data themes. One of the themes is *Human Health and Safety*, which for example describes the geographical distribution of dominance of pathologies (allergies, cancers, respiratory diseases etc.). However, no minimum data quality requirements or data quality recommendations are defined for this spatial data theme.

Inspire was built on existing infrastructures for spatial information operated by Member States. It is assumed that these infrastructures were established in similar contexts and serving similar purposes, but this has not been thoroughly investigated. However, it is unlikely that existing infrastructures in the health data domain will be equally comparable. Health data holders are very diverse organisations/entities and include hospitals, research infrastructures, national authorities, statistical institutes, universities etc.

This diversity must be taken into consideration when establishing legislation and governance in the area of health data quality.

The establishment of a Committee to assist the Commission, again, is a recurring element.

9 Identification of features that could be legally bound

This section aims at providing advice on data quality assurance processes that could require either legal enforcement or be subject to recommendation. For this purpose, we have analysed the different steps of the EHDS data and users' journey for any secondary use of data (i.e., policymaking, regulation, or research) and identified those where data quality may be significantly affected. In table 1 below, where the whole data journey cycle is described, we have identified five steps that are particularly relevant to data quality assurance: Data collection, data publication, data discovery, data delivery and data processing and analysis.

The steps in table 1 that are not directly related to data quality, for example how to set up a network of trustworthy institutions or how to access data, are within the scope of Work Package 5 and will not be addressed here.

9.1 Data collection

Data is collected from a number of different sources. Data collection does not necessarily mean the recording of health data at the point of care, i.e. recording/registering patients' information. Data collection in this context happens when public institutions collect and curate data from multiple sources – EHRs for example, registries or surveys.

The EHDS users will need to trust on accessing data collections that enable high-quality policymaking, regulation, and research. Data collections will be gathered, curated, and maintained by a number of institutions that, ideally, should shape a network of EHDS trustworthy institutions. These institutions, namely data holders, should be audited on procedures of quality assurance including:

- Whether they have a DQA system operational
- Whether they maintain high-quality data collections
- Whether they maintain semantic interoperability standards
- Whether they maintain meta-data catalogues for each data collection.

Data holder institutions need to be audited in accordance with the EU DPA or another EU data quality assurance body. In addition, given the evolving nature of data collections, a rating and promotion system should be offered beyond the mandatory auditing system. A consensus process fostered by the EU DPA should be taken into consideration to define the dimensions to assess and rate on. Once the system is agreed, a self-assessment tool available for data holders would inform the DPA on the rating value and/or the possibility of promotion.

When defining the items to be audited, there are elements already considered in Work Package 7, in particular in M7.1, suggesting a number of actions to be taken during data ingestion; e.g., data format validation, wrong units, temporal consistency, missing data, management of outliers.

9.2 Data publication

EHDS data holders are obligated to publish meta data about their data sets including for example information on:

- the provenance of the data collection (i.e. data collector, the identification of the data origin, whether data has been processed at origin, etc.),
- the relevance of the data collection (meaningful uses and potential “linkability”),
- the coverage (population, population subgroups and geography covered),
- the timeliness (production and refreshment) and
- how accurate and reliable the information is (potential sources of bias, missingness, completeness, consistency across data sources).

Data holders will be recommended to use machine readable standards.

The publication of meta-data should follow the DCAT standard of publication: https://www.w3.org/TR/vocab-dcat/

9.3 Data discovery

Although meta-data information will provide users with relevant information, they may need additional information on the quality of the data at variable level. For example, how much impact outlier values do have in the data set or what kind of data distribution the relevant variables have.

Data holders have to be encouraged to prepare synthetic datasets mirroring their data collections to allow the implementation of additional quality assessment tools (e.g., visual analytics in a standardised way). Otherwise, data holders will have to publish this information by default (for example, percentage of missing data, distribution graphs for all the variables in the data set, outlier values, etc.)

9.4 Data delivery

Once a data holder gets a query from a potential user, the institution services will have to prepare the datasets. Before effective delivery, either if data is transferred out or not (i.e. data remains in premises and remote access provided), the data preparation process is responsible for gathering all the information required and performing the necessary harmonization steps so as to provide a coherent data set or data sets to the data requester. Data preparation may need complex linkage operations of individual and aggregated data, when the research questions to be answered imply the use of a large number of data sets coming from multiple ecosystems (e.g., research projects data, real-world data, clinical trials data etc.). Data will be provided as pseudonymised or anonymised data and will be minimised for specific purposes.

Data holders should be obligated to publish their data preparation and delivery procedures as a legally bound requirement.

9.5 Data processing and analysis

Once data has been delivered to the requester, data will be processed to finally provide the answers to the research questions.

In terms of the data quality, users will be strongly recommended to publish their research protocol with detailed information on the methodology, and the whole analytical pipeline using OpenAire initiatives (e.g., Zenodo), so as to allow other actors in the EHDS to assess the process and reproduce their results. In the case of using AI intelligence analytics, the recommendation is to implement as a general approach explanatory artificial intelligence (XAI). Specific recommendations on this respect will be provided in the final deliverable.

KEY to Table 1 (Below)

Note:

No = No need; R = Recommended; M = Mandatory

Competent bodies: Bodies which grant access to the re-use of health data to support the public sector.

Data holder: Means a legal person or data subject who, in accordance with applicable Union or national law, has the right to grant access to or to share certain personal or non-personal data under its control; may or may not be a designated National or Regional DPA; may or may not act also as data collector.

Data collectors: Public entities that collect health data or any kind or data from other sectors relevant to health.

Data user: Natural or legal person who has lawful access to certain personal or non-personal data and is authorized to use that data for commercial or non-commercial purposes.

Table 1 Recommendations to follow

			Legal enforcement			
Item	**Governance (Who is liable?)**	**Matters of regulation**	**No**	**R**	**M**	**Sources**
Data collection	Competent bodies |DPA | Data hold's | Data coll's	Regular audits			M	
	Competent bodies | DPA | Data hold's | Data coll's	Rating system and promotion		R		HDRUK – practice area 1 and 3
Data publication	DPA | Data holders | Data collectors	Meta-data catalogues - DCAT			M	EUROSTAT | HDRUK | INSPIRE
	DPA | Data holders | Data collectors	Meta-data catalogues – machine readable		R		
Eligibility to the EHDS	EU DPA | Competent bodies | DPA	Bases for the development of the network of EHDS trusting institutions				EC
	EU & Nat'l DPAs	Communication protocols with/across DPAs				
	Competent bodies |DPA | Data holders	Communication with data collectors				HDRUK practice area 2
Data discovery	Competent bodies |DPA | Data hold's | Data coll's	Standard query language in place (ie, API)				
	Data holders | Data collectors	Building synthetic data sets mirroring data collections | publishing visual analyses of quality at variable level		R		Partly HDRUK -practice level 1 comprehensive level
Data access	DPA | Data holders | Data collectors	Clear access procedures (guidelines published)				BBMRI
	Competent bodies | DPA | Data hold's	Safe access to individual level data				BBMRI
	DPA | Data holders | Data coll's	Guidelines to comply with Ethical Principles				
	Users	Data management plan				
Data delivery	DPA | Data holders	Clear processing procedures (guidelines published)			M	HDRUK Practice area 4
	DPA | Data holders	Not hampering meaningful reuse – pseudonyms as preferred system		R		
	DPA | Data holders	Communication system for data delivery				
Data processing & analyses	Competent bodies | DPA | Data hold's | Data coll's	Access through Secure Computing Environment				BBMRI
	Data user	Auditable software		R		OPEN AIRE
Finalization & Devolution	Data user	Destruction of the datasets obtained				
	Data user	Open-source outputs	No			OPEN AIRE

10 Conclusions

A number of conclusions can be drawn with regards to features of an EHDS data quality assurance framework, that could be legally bound. This section summarises the conclusions based on the discussions and findings presented in this document.

The milestone document concludes that the processes around the data - from data collection to data delivery - are key to determining what questions a dataset may be able to answer, and this is why this milestone recommends focussing on specific steps in the data quality assurance process to ensure the reliability of data (health and other sectoral data) for trustful reuse in policy making, regulation and research.

To avoid impacting patient treatment negatively by implementing standards or legislation that draws resources away from the point of care to improve the secondary use of data, legislation should be aimed at the data quality assurance processes from the point of data collection at an institutional level, including mandatory quality assessment processes, meta data and auditing.

Data collection does not mean the recording of health data at the point of care, i.e. collecting data from patients. Data collection, in this context, happens when public institutions collect and curate data from multiple sources, for example electronic health records, registries or surveys.

The quality assurance specific to secondary use of data begins when data is compiled for purposes other than the primary purpose for which the data was collected. The cost of implementing data quality procedures specific to the requirements of EHDS should be covered by the primary users of the EHDS and those end beneficiaries of the secondary use of data (policy makers, regulation agencies and citizens).

There is consensus in Work Package 6 that it would be beneficial to establish an EU EHDS body, charged with developing guidelines on data quality assurance, and that the preferred model of the EHDS is a federated model of national bodies responsible in each Member State. The actual data quality assurance framework should be implemented on an institutional level in the different Member States. As such, national authorities are responsible for implementing the EHDS Data Quality Assurance Framework.

The proposed EHDS body could draw inspiration from the Committees established with a base in the legislation on ESS and INSPIRE. This option will be further explored with Work Package 5 as part of providing recommendations on governance models for the EHDS.

Furthermore, Work Package 6 recommends that an EHDS data quality assurance framework should be applied on an institutional level rather than the level of individual data sources.

Recommendation on legally bound audits

Data holders should be audited on procedures of quality assurance, and a national authority should audit the data holder institutions and their data sets in accordance with the EHDS Data Quality Assurance Framework.

A data quality assurance framework focussing 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.

Recommendation on legally bound processing procedures

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 and ensure the highest possible degree of transparency.

Recommendation on legally bound meta data catalogues

Data holders should be obligated to publish meta data about their collections including information on data provenance, the relevance of the data collection, the coverage of the data collection, the timeliness and how accurate and reliable the information is.

References

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

2. *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)*

3. "*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

4. "*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

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

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

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

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

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

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

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

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

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

    a. *Quality Assurance of Administrative Data: Setting the Standard* (January 2015)

    b. *Administrative Data Quality Assurance Toolkit* (February 2019)

14. *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

15. *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

16. 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

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

18. *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

Annex

List of Contributors

Author	Partner
Cátia Pinto	Shared Services of Ministry of Health, EPE (SPMS)
Hanne Louise Høimark	Central Denmark Region; Denmark (RM)
Christian Fynbo Christiansen	Central Denmark Region; Denmark (RM)
Maria Heilskou Pedersen	Central Denmark EU Office; Denmark (CDEU)
Philipp Schardax	Federal Ministry of Social Affairs, Health, Care and Consumer Protection; Austria (ATNA)
Roch Giorgi	Aix-Marseille Unversity; France (FR-HDH -AMU)
Jean-Charles Dufour	Aix-Marseille Unversity; France (FR-HDH -AMU)
Océane Xayakhom-Dauvergne	Health Data Hub; France (FR-HDH)
Zoltan Thinsz	National Board of Health and Welfare; Sweden (SEHA - NBHW)
Beatrice Kluge	Gematik; Germany (BMG - GEMATIK)
Herman van Oyen	Sciensano; Belgium
Shona Cosgrove	Sciensano; Belgium
Sarah Craig	Health Research Board; Ireland (DoH -HRB)