29 June 2022
EMA/615349/2022
Data Analytics and Methods Task Force

# Workshop on Data Quality Framework for medicines regulation - Report

7 April 2022,
Virtual meeting, European Medicines Agency

**Official address** Domenico Scarlattilaan 6 ● 1083 HS Amsterdam ● The Netherlands
**Address for visits and deliveries** Refer to www.ema.europa.eu/how-to-find-us
**Send us a question** Go to www.ema.europa.eu/contact **Telephone** +31 (0)88 781 6000

# Background

Europe's healthcare sector is generating increasing amounts of raw data contained in electronic systems from primary, specialist and hospital care, drug and disease registries, clinical trials, genetic and genomic testing, manufacturing processes, spontaneous adverse event reports and mobile health systems. All these data are increasingly used for purposes of improving healthcare and thereby the lives of the citizens of the European Union.

EMA is looking to develop an European Data Quality Framework for characterising and measuring data quality in the context of regulatory decision-making. There are many potential use cases spread over different Member States' jurisdictions, public and private healthcare institutions, and a myriad of independent systems. Centralised large-scale databases are built and maintained through established production systems through aggregation and standardised quality control, but this model is acutely limited through lack of interoperability and strict patient data protection rules.

Federated models are gaining rapid traction, using the data while keeping them behind the firewalls of their originating organisations. This model requires a standardisation and systematic quality management that can be applied to such distributed systems.

The workshop was an initiative of HMA, EMA and TEHDAS meant to both share the current progress on building a data quality framework for medicines regulation and to solicit the comments and ideas of experts in this field to help shape the drafting process. Breakout sessions focused on particular use cases with fruitful discussions on the current data quality landscape and how the future of data quality in medicines regulation should look in the clinical and non-clinical areas.

# Welcome and introduction

Emer Cooke, EMA, opened the meeting by welcoming everyone to the workshop. Ms. Cooke thanked TEHDAS for the collaboration with EMA and HMA on this effort and expressed the Agency's vision to transform medicines regulations in a data driven process. For evidence-based decision making, there is a need for a good understanding of the quality of the data that underpins these decisions, which translates into a need for a European Data Quality Framework.

The experience with the COVID-19 pandemic proved to EMA and the world that there is a need for high quality data to support decision making. The response to the pandemic capitalised on this need, but also brought into focus a wider requirement for high quality data to support robust scientific assessment and regulatory decision making. This requirement is being addressed as part of the Big Data Steering Group workplan, in the drafting of the EU Data Quality Framework.

Peter Bachmann, Head of Unit International Liaison Office and Conferences for the Federal Institute for Drugs and Medical Devices, BfArM (Germany), presented the overarching vision of the Big Data Steering Group (BDSG) and the HMA-EMA workplan which includes the Data Quality Framework workshop.

The recommendations of the BDSG are:

- Establish a data quality framework for regulatory use of big data sources with associated data quality metrics

- Expansion of qualification advice process to establish renewable certification of datasets and big data methods and strategies
- Establish criteria for reliability of device based diagnostic and other vitro diagnostics
- Proactive external communication to promote adoption of data quality framework
- Promote use of ISO-IDMP standard

The applicability and intended use of the Data Quality Framework would ultimately be tailored for the intended regulatory purpose. It was noted that the data quality framework needs to be applicable to various types of databases used in the regulatory activities surrounding medicinal products for human and veterinary use, such as:

- Real world data
- Clinical trial data
- Spontaneous reporting
- Quality
- Manufacturing
- Genomic data
- Imaging data
- Non-Clinical Data
- Veterinary

The goal for the EU Data Quality Framework is included in the Network Strategy 2025 alongside enabling access to healthcare data (including subsequent analysis), analysis of individual patient data from clinical trials, and promote the standardisation of targeted data.

# Plenary session: Current Data Quality Landscape

Peter Arlett, EMA, chaired this session.

Paolo Alcini, Head of Healthcare Data at EMA, presented an overview of the landscape analysis of data quality frameworks, performed as a preparatory step in the drafting process. During the analysis it was noted that many of the proposed frameworks were still in a conceptual phase, as ideas or approaches, but rarely accompanied by an applicable process or tools.

The landscape analysis collected information focused around the following areas:

- Formalised Data Quality Framework
- Regulatory Compliance
- Experience
- Extensible Framework
- Product or Tool

Danica Marinac-Dabic, Associate Director at the Office of Clinical Evidence and Analysis at FDA, presented the maturity model for medical device registries and coordinated registry networks (CRNs). CRNs are the real-world data sources, started by the FDA in 2015, encompassing strategically partnered electronic health information systems serving one or more clinical areas. These CRNs build on the national and regional registries, harmonise data elements and link data to comparable data across systems (e.g. EHR, administrative claims, patient generated data)

The levels of maturity (Early Learner - Making Progress - Defined Path to Success - Well Managed – Optimized) are being developed and optimized by the FDA and are envisaged to be publish at a later stage. A number of US CRN Learning Community groups and foci concentrate on various clinical areas for each CRN. The FDA has launched this effort in Nov 2014, supported by MDEpiNet Analytic Center and presents a collaboration of 28 regional and national registries. These CRNs are already producing the regulatory-grade evidence needed for post market surveillance, mandated post approval studies, labelling expansions and ROI studies documenting up to 550% return of investment using this approach.

Enrique Bernal-Delgado, Head of Data Science for Health Services and Policy Research at the Institute for Health Sciences in Spain, provided an overview of TEHDAS (the joint action Towards the European Health Data Space), and their ongoing initiative towards building a data quality framework for the European Health Data Space (EHDS) for secondary use of data. TEHDAS joint action is split in a number of work areas; of particular interest, joint action work area 6, focuses specifically on excellence in data quality and building a data quality framework (DQF). The goal of this work area is to provide recommendations on the overarching concepts, structuring the recommendation in the context of the data life cycle, roles and responsibilities and guidance on how these measures can be implemented. Secondly, the initiative aims to provide recommendations on standards of interoperability for data discoverability, semantic interoperability and nodes communication.

Peter Berzin, Head of Data Quality at Odysseus Inc, presented an overview of quality control systems for secondary use data. The presentation included the reiteration of the difference between primary use data and secondary use data, and which use cases/data sources fit into each category. Primary use data is considered "controlled data" as it is generally regarded as highly regulated and governed by international standards - including those requiring Quality Management Systems. Secondary Use Data is considered "uncontrolled data" in the sense that it lacks quality measures, as the data is re-used for other purposes including regulatory decision making, instead of its initial purpose. In this case, the data quality needs to be measured using quality dimensions that determine if the data are fit for the regulatory purpose. Various groups focus on expanding QMSs in order to control the quality of data at the data entry point (e.g.: ASQ and CogentQI).

Various Data Quality Frameworks were presented along with the data quality dimensions that each of these frameworks focuses on. The more dimensions a framework checks, the more extensible it will be for other data sources.

# Breakout sessions and reports

The breakout sessions were designed to draw out ideas, recommendations, and possible solutions for establishing a robust data quality framework in each of the breakout session focus areas. The feedback and comments from these sessions will be used to shape the overall Data Quality Framework being designed by HMA-EMA for regulatory purposes.

## *Group 1: Secondary Use of Real-World-Data (EHR, claims, registries, and wearables)*

***Chair:*** *Enrique Bernal-Delgado, TEHDAS*

***Rapporteur:*** Luis Pinheiro, EMA

The group discussed how secondary use real-world data is currently governed, principles underlying high data quality, and what frameworks or guidelines are needed to ensure that the data is high quality and fit-for-regulatory purpose.

**Table 1.** Summary of Breakout Session 1 - *Secondary Use of Real-World-Data (EHR, claims, registries, and wearables)*

| Questions | Summary Notes |
|---|---|
| Should secondary use of real-world data be accessible to or governed by regulatory bodies, if so, how? | Support for access to RWD particularly for safety and drug utilization use cases. However, concerns around lack of harmonization, ethical approval of studies. **Consistency, standardization and ethics of access** are key words. |
| Are existing data quality frameworks sufficient to determine the quality and fit for regulatory purpose of secondary use real-world data? If not, what are the challenges? | Data Quality and Fit-for-Purpose related but distinct – need to be clear what is meant with fit-for-regulatory-purpose. DQ frameworks may need to be tailored to specific use cases. **Tailored** DQF **and flexibility** are key issues. |
| Are data standards needed and is there a need to transform the data for data quality to be assessed (i.e. standardised to common data model and language)? | Difficult creating an overarching data quality framework and difficult to assess DQ in a timely manner without standards. DQ standards are time dependent. CDMs have served well for medicines but not medical devices. **Interoperability** is key. |
| What regulatory guidance or influence is needed for secondary use data, if any? | Ensure alignment and organized deployment across EU. Clarity on what regulatory thresholds for quality are. **Regulate & Collaborate** and **Thresholds for decision-making** are key. |

## *Group 2: Primary Use Data (Pre-clinical, Clinical trial, and Manufacturing)*

***Chair:*** *Frank Pétavy, EMA*

***Rapporteur***: *Eftychia-Eirini Psarelli, EMA*

The group discussed how pre-clinical, clinical trial and manufacturing data is currently defined, collected, controlled, enforced, or influenced. The group considered how high quality, in terms of data, is defined in these areas and what barriers to high quality data exist.

**Table 2.** Summary of Breakout Session 2 - *Primary Use Data (Pre-clinical, Clinical trial, and Manufacturing)*

| Questions | Summary Notes |
|---|---|
| What measure of quality should pre-clinical, clinical, and manufacturing data conform to in the context of regulatory decision making? | • **Lack of guidance in the pre-clinical setting** when compared to the other areas<br>• Different measures per region apply<br>• Data quality is **context specific**; need to accept various levels of quality depending on the research area<br>• Fair interplay of data quality for medical devices needs to be ensured<br>• Quality of algorithms (AI, ML, NLP) should not be disregarded |
| How can the existing pre-clinical, clinical, and manufacturing regulatory requirements or processes be strengthened to produce higher quality data? | • Need for **increased interaction between regulators and manufacturers** at an **early stage of product development lifecycle**; impact is more acute on innovative products (e.g. ATMPs); patients and research investment<br>• Need to leverage clinical trial data for secondary use |

| Questions | Summary Notes |
|---|---|
| | • Need for **education and training** on existing guidelines in the clinical and manufacturing area<br>• Data quality goes hand-in-hand with **quality of experiment** |
| What barriers to high quality data exist today with these data types? | • Lack of **awareness of data quality standards** for clinical and manufacturing<br>• Data **fragmentation and heterogeneity**<br>• Lack of **interoperability** across all domains<br>• Lack of **harmonisation** in documentation process and selective reporting in pre-clinical setting<br>• Data access barriers for **inspection** purposes<br>• Lack of **Therapeutic Area guidance** in some regions |

## *Group 3: Considerations when augmenting standardised primary or secondary use data with study-specific data (Hybrid approach)*

***Chair:*** *Xavier Kurz, EMA*

***Rapporteur:*** *Lifang Liu, EMA*

Discuss approaches when existing secondary data is insufficient to reliably address a study question and needs to be complemented with primary study specific data or vice versa (e.g. RWE).

**Table 3.** Summary of Breakout Session 3 - *Considerations when augmenting standardised primary or secondary use data with study-specific data (Hybrid approach)*

| Questions | Summary Notes |
|---|---|
| What challenges and opportunities exist when secondary data for addressing a study question and needs to be complimented with primary data or vice versa (e.g. RWE)? | **Opportunities:**<br>• Collect complementary data on study exposures and outcomes (smoking, Healthcare status etc)<br>• Registry-based randomized studies (e.g. medical devices)<br>• Feasibility to collect secondary data from patients/identify the right patients and contact them later on<br>• Use secondary data to populate in the medical information, enrich and broaden medical history<br>• Follow up trial patients after trial completion for longer follow up schemes<br>• Learn from clinical trial data to extrapolation to real-world population<br>• Primary data can be used to validate secondary database (variables, outcomes)<br>• Additional primary data from the patient perspective (PRO) can be collected to supplement secondary data<br><br>**Challenges:**<br>• Identify the right patients in the database/linkage problems<br>• Access to patient data (patient consent etc.)<br>• Internal validity of the hybrid approach in terms of quality (primary vs secondary data)<br>• Prospective approach causes selection bias, align data collection, define variables<br>• Risk of conflicting data: having different diseases/status<br>• Data with different quality (controls), the acceptability/use of the data will be affected by the one with the POOREST quality |

| Questions | Summary Notes |
|---|---|
| Are there any standardised / harmonised approaches (beyond existing guidance) that can be suggested for the acquisition of study-specific data? | • The joint action of registries- PARENT- has recommendations of linking databases<br>• Guidelines for registry-based randomized studies, originated from Scandinavian countries<br>• Exact linkage vs probabilistic linkage—increase transparency: standard sharing, algorithms, accuracy<br>• Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: overview and implications (Padmanabhan S et al. 2019)<br>• Align/find the same patients across different databases which are of totally different structure through common data models<br>• What would be the data quality framework for the hybrid approach:<br>a. implementation from the initiation (clear about what data to be used in secondary use to design the initial trial data)<br>b. Same CDM should be applied in primary and secondary data usage (to avoid data collection redundant CDM clinical trial vs clinical practice)?<br>c. Formal procedures to select data from existing data sources for both primary and secondary use of data |

## Data quality framework – proposed structure

Christian Reich, VP Real World Solutions at IQVIA, presented a proposal on the structure of a future draft of the Data Quality Framework. The importance of defining a robust and consistent terminology related to data quality, harmonised with other similar initiatives where applicable and appropriate was highlighted. The landscape analysis also noted, that data quality dimensions and maturity models are a helpful tool to be used within the regulatory process. Lastly, the importance of establishing communication guidelines on clarity and transparency principles for data quality issues was highlighted.

The focus of the drafting process will initially be on general guiding principles previously defined while drafting the specific data quality dimensions and will build incrementally on further use cases within the identified priorities of the regulatory network. With regards to initial deep dive on particular use cases, the priority identified revolves around secondary use data, longitudinal data and clinical data of patients used for secondary use. In this context, for the purpose of regulatory decision-making, understanding the real world, utilisation of medicines, estimation of benefits/risk, supporting relative rather than absolute evidence, and supporting probabilistic evidence are some of the topics that need addressing.

## Concluding remarks

Jesper Kjær, Danish Medicines Agency, Co-chair of the Big Data Steering Group, on behalf of the Big Data Steering Group and Peter Arlett, thanked all the presenters and moderators of the workshop. He also gave a special thanks to the breakout session rapporteurs for their impressive work on summarising the discussions.

The general scope of the data quality framework will be drafted during 2022, taking the input from the workshop into consideration, as well as being consulted with stakeholders as part of the development. Emphasis was placed on defining concepts such as 'fit for purpose', especially in various therapeutical areas, considering the impact of introducing quality requirements and defining a harmonised terminology in relation to data quality aspects. Consistency, standardisation and ethics of access would need to be carefully considered in this context. The aspect of quality education and training should not be neglected, ensuring that the community is able to fully understand and embrace awareness about data quality standards.

Similar initiatives, such as the ones shown in the landscape analysis or maturity models proposed by international regulators need to be carefully considered in a European setting for both primary and secondary use of data. TEHDAS activity on data quality aspects is important particularly in the context of the data lifecycle from a regulatory perspective, in strengthening the quality by design and data capture.

The need for better guidance on data quality in the pre-clinical setting was brought forward; it was underlined that there is a need to align globally around this topic. It was also emphasised that this alignment should be coordinated as an interaction between regulators, manufacturers in early stage of production and development of life cycle of medicinal products. Data quality management in clinical trials is generally considered to be relatively mature, therefore is likely more beneficial to concentrate efforts in areas where the needs are greater.

With regards to international harmonisation, the Big Data Steering Group will endeavour to bring up these topics in their international outreach.

The closing note remarked the importance of a collaborative approach on the topic on a global scale and appreciated the engaging participation in this workshop.

# List of abbreviations

| Abbreviation | Definition |
|---|---|
| AI | Artificial Intelligence |
| ATMPs | Advanced therapy medicinal products |
| BDSG | Big Data Steering Group |
| BfArM | Bundesinstitut für Arzneimittel und Medizinprodukte |
| CDM | Common data model |
| CRN | Coordinated Registry Networks |
| DQAF | Data Quality Assurance Framework |
| DQF | Data Quality Framework |
| EHR | Electronic Health Record |
| EHDS | European Health Data Spaces |
| EMA | European Medicines Agency |
| EMR | Electronic Medical Record |
| FDA | Food & Drug Administration |

| | |
|---|---|
| HMA | Heads of Medicines Agency |
| ISO | International Organisation for Standardisation |
| ISO-IDMP | International Organisation for Standardisation for the identification of medicinal products |
| ML | Machine Learning |
| NLP | Natural Language Processing |
| PRO | Patient reported outcomes |
| QMS | Quality Management System |
| RWD | Real World Data |
| RWE | Real World Evidence |
| TEHDAS | The joint action Towards the European Health Data Space |
| UDI | Unique Device Identification |