2 September 2021
EMA/139729/2021
Data Analytics and Methods Task Force

Technical workshop on real-world metadata for regulatory purposes

12 April 2021

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 An agency of the European Union

Disclaimer:

This document is a draft for consultation purposes only.
It should not be interpreted as representing the formal position of EMA or HMA.

Background

Identification of appropriate data sources to generate real-world evidence is of increasing need for regulatory decision making. Metadata are data that describe other data to create a clearer understanding of their meaning and re-use and to achieve greater reliability and quality of information. Access to a standard and electronic set of complete and accurate metadata information can contribute to identifying the data sources suitable for a specific study, facilitate description of the data sources planned to be used in a study protocol or research proposal and contribute to assessing the evidentiary value of the results of studies using multiple data sources. Data discoverability is a goal for the scientific community. The Heads of Medicines Agencies–European Medicines Agency (HMA-EMA) Joint Big Data Task Force recommended "*to promote data discoverability through the identification of metadata*" as part of its Recommendation III: "*Enable data discoverability. Identify key meta-data for regulatory decision making on the choice of data source, strengthen the current European Network of Centres for Pharmacoepidemiology and Pharmacovigilance (ENCePP) resources database to signpost to the most appropriate data, and promote the use of the FAIR principles (Findable, Accessible, Interoperable and Reusable).*" This goal is also included in the 2020-2021 Work Plan of the HMA-EMA Big Data Joint Steering Group.

The goal of this technical workshop was to review and gather stakeholders' feedback on the preliminary list of metadata required for characterising real-world data sources and their definitions to fulfil regulatory use cases, the proposed options for the process to collect and maintain metadata from real-world data sources, and the proof-of-concept catalogue of data sources and metadata currently being developed.

Welcome and introduction

Paolo Alcini, EMA, opened the meeting by welcoming everyone to the workshop.

Nikolai Brun, from the Danish Medicines Agency and co-chair of the Big Data Steering Group, presented the background of the metadata project and how it fits within the framework of the HMA-EMA Big Data Taskforce priority recommendations and the DARWIN EU project. Dr Brun emphasised the objectives of the Big Data Initiative Recommendation 3, which focuses on the need to gather input from stakeholders on the preliminary list of metadata required for fulfilling regulatory use cases, the definition of the metadata list, and the proof-of-concept catalogue. He also highlighted the need for automation of the data processing of algorithms to create and maintain metadata.

Stefania Simou, Data Scientist from the EMA Data Analytics and Methods Taskforce, provided an overview of the HMA-EMA Big Data Taskforce priority recommendation to enable data discoverability and the increasing importance of real-word evidence generation to inform regulatory decisions. Acknowledging the increasing complexities of data needs, Stefania Simou underlined the importance of the identification of appropriate real-world data sources and the need for a comprehensive assessment of their characteristics and contents and establishing a framework for data quality and representativeness. Attention was also given to the regulatory use cases of metadata to support the selection of a research network for analysis, conduct of studies to address research questions, evaluation of research proposals, and assessment of the evidentiary value of the results of studies. Stefania Simou also introduced the Metadata for data dIscoverability aNd study rEplicability in obseRVAtional studies (MINERVA) project that EMA launched in November 2020 to deliver on recommendation 3 of the HMA-EMA Big Data Taskforce.

Susana Perez-Gutthann, from RTI Health Solutions and Lead of the MINERVA Consortium, provided an overview of the project. As of the date of the workshop, the Consortium had developed a preliminary list of metadata and proposed options for the process to collect and maintain the metadata from real-

world data sources, and a preliminary a proof-of-concept catalogue, all of which were presented in this workshop. The feedback collected through the survey before the workshop, as well as the workshop discussions, will be integrated into the next steps: finalising the list of metadata with definitions, describing the options of metadata collection and maintenance, piloting the metadata collection with the participating data sources, and populating the proof-of-concept catalogue with the metadata obtained. The final deliverable will be a good practice guide summarising the experience of this pilot and providing recommendations. The MINERVA Consortium is a partnership of 18 research centres in 12 countries covering 15 population data sources and patient registries that are specialised in regulatory-driven post-authorisation safety studies. The metadata and proof-of-concept catalogue intend to be the basis upon which to build a system that will make working with data sources easier and that should allow regulators and prospective investigators to identify data sources meeting the requirements for specific studies. The metadata in the catalogue tool should also assist throughout the process of data source discovery and feasibility assessment, design of the protocol and statistical analysis plan, data processing, and development of the computational analysis script. The catalogue tool should also support the dissemination and evaluation of the resulting evidence, all under FAIR guidance principles. This effort should also strengthen the ENCePP Resource Database.

Session 1: Conceptual framework, definitions, and use cases

Xavier Kurz, EMA, chaired this session.

Romin Pajouheshnia, MINERVA Consortium, Utrecht University, introduced a brief video to present the use case of the proof-of-concept catalogue. The primary use case for the metadata list and concept catalogue (i.e., supporting data discoverability in the context of conducting a new study for regulatory purposes) was presented. It was explained that the preliminary metadata list was designed around this use case, as well as to support the (1) transparency of data and evidence generated to support regulatory decisions and (2) assessment of the evidentiary value of results from analyses conducted using real-world data. The video showed a preliminary "mock-up" of a proof-of-concept catalogue, with a user navigating the proof-of-concept catalogue to identify data sources that could be used in an example regulatory study to investigate the safety of vaccines for COVID-19. During the video, one possible workflow was demonstrated as an example, where relevant data sources and the respective data banks were identified by first identifying similar existing studies and navigating the metadata descriptions associated with the data banks, data sources, and institutions involved.

Rosa Gini, MINERVA Consortium, ARS Toscana, described the underlying conceptual framework and definitions. EMA staff authored a paper published 3 years ago describing electronic healthcare databases in Europe for regulatory purposes (Pacurariu et al., 2018). This publication laid the foundations for the conceptualisation further developed within the IMI ConcePTION project, and now in MINERVA with additional definitions for very specific concepts. In the ConcePTION study, 20 data sources were analysed through interviews with researchers with expertise in the data sources. The initial unit of observation to describe a data source was a table. Content analysis of the documents revealed latent concepts: *data banks* are routine collections of data sustained by an organisation that is normally not a research institution (e.g., a healthcare payer). The *prompt* for creating a new record in a data bank is also an important concept and a determinant of which records are collected, e.g., still births would be included in a birth registry where records are prompted at birth, whereas they are not included in a birth registry in which records are prompted at the first home visit. In the ConcePTION study, four data sources included data from only a single data bank, while 16 data sources included data from multiple linked data banks on the same or on overlapping populations.

Susana Perez-Gutthann summarised the process to review and address the stakeholders' feedback received prior to the workshop, which was followed by a discussion chaired by Xavier Kurz focussed on

clarifications regarding the conceptual framework, key definitions and criteria, and selection of data sources. Specific discussion points included that, in this conceptual framework, a data access provider extracts data (or receives permission to access extracted data) from data banks to conduct a given study. A data source needs to include at least one data bank. While the underlying population of a data source might include the entirety of the population of a given region, persons who do not contact the health care system may not be included in the existing data banks. Discussion points around the data source selection criteria included that criteria were meant to be broad at this stage so that a large number of data sources could meet them; metadata to be collected later during this project will be informative on the characteristics and quality of the data sources and could be used to develop quality-based criteria in future projects beyond this pilot. The list of proposed metadata includes metrics to assess whether data capture has been continuous and complete, although there might be a degree of subjectivity in the assessment; the intention was that data sources that collected data for only 1 or 2 years would not be eligible. Discussion points around the six domains presented in the proof-of-concept catalogue included that the six domains are highly interconnected, as there are many paths to navigate the tool; this may depend on the final decisions on catalogue implementation and on how the user may want to use the catalogue.

Session 2: Preliminary set of metadata and definitions

Xavier Kurz, EMA, chaired this session.

Romin Pajouheshnia, MINERVA Consortium, Utrecht University, presented an overview of the preliminary list of metadata to describe real-world data sources for regulatory purposes.

The primary goal of the task to derive a metadata list was first presented: "*To define a set of metadata that should be collected from real-world data sources. Metadata should be relevant to regulatory needs; agreed with the Agency; and provide detailed information on source, spectrum, and quality of data sets*."

The term "metadata," broadly defined as "a set of data that describes and gives information about other data," was broken down into subcategories including the *generation*, *location*, *ownership,* and *governance* of the data; processes for *storing*, *handling,* and *accessing* of data; the *origin* and *time span* of data; descriptors of variables captured in the data including quality measures; and data format.

The method to derive the preliminary metadata list was explained to be based on incorporating aspects from a number of existing initiatives to collect metadata, including the ENCePP Resources Database, EMIF catalogue, and ConcePTION catalogue. The catalogue was developed based on (1) a web-based search for publicly available resources and (2) a series of interviews with experts from key organisations. The result was a preliminary metadata list based on information gathered from 57 documents and eight interviews.

The preliminary metadata list was presented within a set of six highly interconnected catalogue domains:

- Institutions (contributors to the metadata list, e.g., research institutions with study-based access to one or more data sources that have contributed to the current metadata list and/or with expertise to write protocols, reports, and scientific papers),
- Data sources (one or more data banks covering the same underlying population or whose underlying populations overlap),
- Data banks (data collections with an originator—the organisation that sustains the collection of records in a data bank, a prompt for record creation, and a data model),

- Common data models (tools for data harmonisation),
- Networks (collaborations of institutions), and
- Studies (research questions that are addressed using data sources made up of data banks).

The remainder of the presentation provided details of the metadata tables and individual variables included in the preliminary metadata list, including metadata on data access and permissions and metadata to quantitatively describe the content and quality of data.

Following the presentation, Romin Pajouheshnia gave a summary of questions from stakeholders, and additional questions were raised by Xavier Kurz.

- The potential for the catalogue format to allow entry of information by other systems, such as the EU PAS Register was considered. Romin Pajouheshnia stated that this should be considered and may require an automated process to update catalogue information. It was also discussed that if the EU PAS Register were updated, allowing capture of metadata from studies, the metadata catalogue should be linked to it.
- The value, feasibility, and potential duplication of work (capture of studies in both the EU PAS Register and the metadata catalogue) of capturing metadata from all studies carried-out based on all registered data banks was questioned. Rosa Gini explained that studies are often based on different data banks from the same data source, each providing useful information, and a specific description of the instance of the data source involved in a study is useful for inferring study replicability. Once the metadata catalogue is initialised, every subsequent update (regardless of whether or not it is study specific) will build on the existing metadata, possibly adding further detail if needed (e.g., due to a new validation study), and/or updating outdated information if necessary. Data curation by means of a central quality control process is envisioned, and it was highlighted that collection of metadata on quality control processes in place at the level of each data bank is needed.
- It was asked whether the pharmaceutical industry would be considered to be a data access provider in the metadata catalogue. Rosa Gini clarified that yes, this could be the case.
- It was noted that the metadata list and proof-of-concept catalogue will need to adhere to international standards and incorporate existing ontologies. This was agreed upon by members of the MINERVA Consortium, and it was explained that this implementation will take place in the next stage of the development of a proof-of-concept catalogue.
- Xavier Kurz asked whether a minimum set of quality indicators could be identified. Romin Pajouheshnia explained that while a number of measures of data quality are included in the preliminary metadata list, the availability of this information will depend on the presence of tools to support machine readability and automation, as well as expert review of these metadata. The challenge of defining a single set of indicators for characterisation of quality was noted. Members of the MINERVA Consortium explained that data quality is not included as a criterion for eligibility for data sources or data banks to be represented in the catalogue, in order to remain as inclusive as possible.
- It was explained that the focus of the catalogue is on European data sources, but no exclusion of data sources outside of Europe is planned.
- It was asked whether the catalogue could include information on the outcomes of regulatory use of the studies in the catalogue. Rosa Gini explained that ENCePP Working Group 3 is trying to obtain this kind of information, which requires manual review. While this would be very useful, it is currently not envisioned within the scope of the MINERVA project.

Session 3: Process to collect metadata and sustainability

Jesper Kjær, Danish Medicines Agency, chaired this session.

Miriam Sturkenboom, MINERVA Consortium, University Medical Center Utrecht, presented and addressed stakeholder questions regarding the metadata collection process and its sustainability.

Several initiatives have started the creation of metadata catalogues, but sustainability, harmonisation, and public access is limited. The MINERVA Consortium believes that sustainability of metadata catalogues can be achieved only when the financial and structural embedding is guaranteed. For structural embedding, it is necessary that the catalogue is compliant with existing European Commission guidance and requirements. In this context, the FAIR initiative is the key foundation for the creation of a sustainable metadata catalogue. Based on FAIR principles, the catalogue should be hosted by a FAIR data node, with a clear data management plan; digital objects (metadata, including information on databanks, persons, institutions) that are essential for research should be in standard formats and with persistent identifiers. This is the basic condition for sustainability. There are several choices to be made (e.g., which identifiers to use, and how to assign them to databanks) before this can be realised, which will be part of the ongoing MINERVA work. With regard to the financial sustainability of a FAIR metadata catalogue, different scenarios for the process to collect and maintain metadata from real-world data sources need to be explored: (1) a basic funding to collect, maintain, and update metadata on databanks independently of studies and (2) study-specific funding. It must be noted, as specified earlier, that even if a study-independent maintenance of the catalogue is structured, the study-specific sections may still include additional detail, because the study-specific section of a data source description is the responsibility of the scientists who conduct the study and must later remain frozen for study replicability. Cumulative information about the databank provided by all the different institutions will allow study replicability and provide the most up-to-date historical and study-specific information about the databank to support discoverability, transparency, and replicability. The possibility of duplicate entries of the same data bank was raised as a concern to be addressed. The MINERVA Consortium will investigate different solutions to this concern in the second half of the project.

The discussion included what type of institution should sustain this kind of catalogue infrastructure: it should be an organisation with a European mandate. Also, this infrastructure should not be related to a time-limited project. Metadata can be populated and maintained by "data stewards," as well as at the study level, which triggers the need for reconciliation. Any efforts need to be integrated with other European health data space initiatives, and regulations to support this effort may require enabling new regulations.

Session 4: Proof-of-concept catalogue

Jesper Kjær, Danish Medicines Agency, chaired this session

Morris Swertz, University Medical Centre Groningen, presented the proof-of-concept tool, or metadata catalogue, due later in the project. The proof-of-concept catalogue will enable access to the metadata discussed above. A vision of how that tool might look was presented to the workshop participants, and clarifying questions were addressed.

Background information was first given on the development process. FAIR principles will be followed, and a rapid prototyping approach will be used using "scrum" methods, reusing existing building blocks from the open-source project MOLGENIS, to benefit from this software's current usage for similar catalogues including the BBMRI-ERIC Directory of biobanks, the Horizon2020 projects EUCAN-Connect and LifeCycle, and the IMI project ConcePTION, amongst others. An important feature of this

methodology is the integration of user evaluation at successive stages of development; this will be sought to ensure bespoke user interfaces, learning from existing catalogues. Automation of data ingest will be explored during the data collection task. FAIR data interfaces should enable automated uploads and downloads. Dashboards will be added to enable catalogue users to compare contents of different databanks or time points, heightening the interoperability of the data.

A short demo of the first iteration of the proof-of-concept tool was given. In this very early version, for example, the way in which a new data model can be set up was demonstrated. The interface through which the user might approach the metadata through multiple different channels (institution, data source, network, etc.) was also shown, receiving positive reactions from the participants.

Stakeholder feedback regarding the functionality of the proof-of-concept catalogue focussed on four main areas: Visualisation/presentation, e.g., ability to quickly compare different sources; Search & filter capabilities, e.g., ability to filter using and/or conditions; Data stewardship, e.g., to ensure metadata can be updated without duplicated entry; and FAIRness, e.g., collaborate on standards with other catalogue creators. Comments were very much in line with the project plan and yielded many new insights; for example, to stratify data by levels of validation; to clarify what individual/institution is responsible for the contents; to describe populations in different cross-sections to understand representativeness of the data; and that the catalogue will enable comparison of data sources and identification of gaps that might exist.

Concluding remarks

"…regulators want to discover data quality characteristics and have high levels of transparency, to explore the evidentiary value of data to support stakeholders, and to contribute to building a European health data space."

Peter Arlett, Big Data Steering Group Co-chair, EMA, emphasised that regulators want to discover data quality characteristics and have high levels of transparency, to explore the evidentiary value of data to support stakeholders, and to contribute to building a European health data space. Peter Arlett highlighted the rich discussion and feedback on the preliminary metadata list, sustainability, and a proof-of-concept catalogue that will help to take forward this work in data discoverability at MINERVA, EMA, and regulatory networks. He thanked EMA and EU network colleagues who contributed to the workshop, the MINERVA Consortium, panel speakers, and those watching. He closed the workshop announcing that further developments on catalogues on real-world metadata, as well as PAS and observational studies, will be posted on the Big Data webpage on the EMA website and wishing for a future with discoverable and impactful real-world data.

Reference

Pacurariu A, Plueschke K, McGettigan P, Morales DR, Slattery J, Vogl D, et al. Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation. BMJ Open. 2018 Sep 5;8(9):e023090. doi:10.1136/bmjopen-2018-023090.

List of abbreviations

ARS Toscana	*Agenzia Regionale di Sanità della Toscana* (Regional Health Agency of Toscana), Italy
BBMRI-ERIC	a European research infrastructure for biobanking
COVID-19	coronavirus disease 2019
DARWIN EU	Data Analytics and Real World Interrogation Network project in the European Union
EMA	European Medicines Agency
EMIF	European Medical Information Framework
ENCePP	European Network of Centres for Pharmacoepidemiology and Pharmacovigilance
EU	European Union
EU PAS Register	European Union electronic register of post-authorisation studies
EUCAN-Connect	a project that brings together science from Europe and Canada to improve the quality of health care through a more efficient use of data
FAIR principles	findable, accessible, interoperable, and reusable
HMA-EMA	Heads of Medicines Agencies–European Medicines Agency
IMI	Innovative Medicines Initiative
MINERVA	Metadata for data dIscoverability aNd study rEplicability in obseRVAtional studies