



Preliminary list of metadata



Technical workshop on real-world metadata for regulatory purposes
Virtual meeting, April 12, 2021

Presented by Dr. Romin Pajouheshnia
Utrecht University



Outline

- 1 Definition and scope of metadata
- 2 Deriving the metadata list – search & interviews
- 3 Preliminary metadata catalogue – tables and variables
- 4 Summary

A preliminary list of metadata to describe European data sources



Objective

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**

Defining the initial scope

Basic definition of metadata

A set of data that describes and gives information about other data

Scope

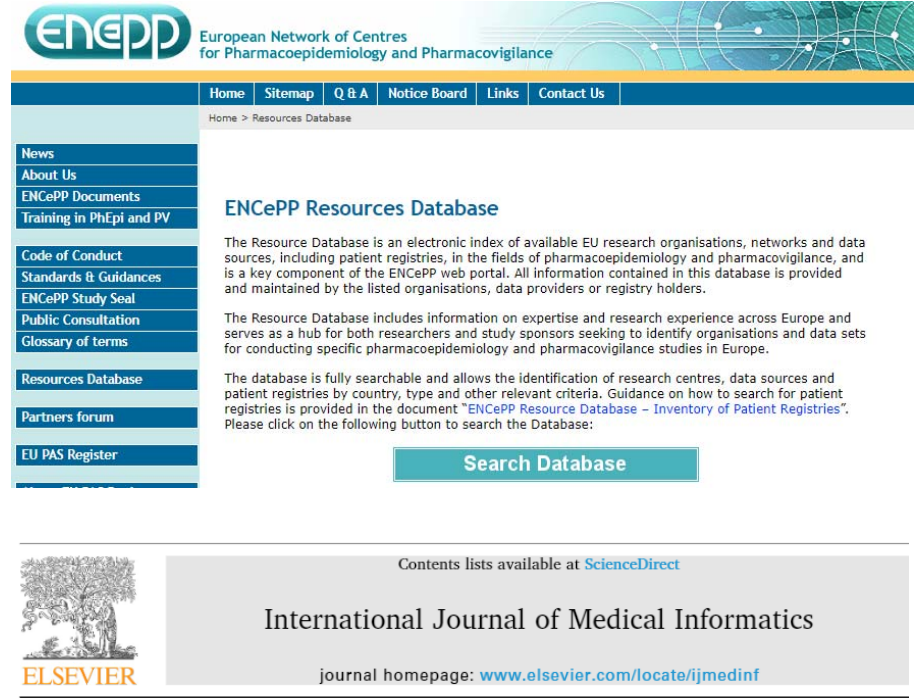
- Data describing the *generation, location, ownership, and governance* of the data (set)
- Data describing the *processes of storing, handling, and accessing* of data
- Data describing the *provenance/origin and time span* of the data
 - Including the input, systems, and processes that define data of interest
- **Descriptors of variables** captured in the data
 - Descriptors of the *underlying population, including disease populations*
 - Indicators of *quality and completeness*
- The *format (structure, model, coding)* in which the data are collected

Existing initiatives

- A number of related catalogues/ metadata databases exist
(e.g., ENCePP resources, EMIF catalogue)
- Need to build upon existing catalogues
- Design should support envisioned catalogue design and function

<http://www.encepp.eu/encepp/resourcesDatabase.jsp>

Oliveira JL, Trifan A, Bastião Silva LA. EMIF Catalogue: A collaborative platform for sharing and reusing biomedical data. Int J Med Inform. 2019; 126: 35-45.



The screenshot shows the ENCePP website interface. At the top, the ENCePP logo is displayed alongside the text "European Network of Centres for Pharmacoepidemiology and Pharmacovigilance". A navigation menu includes links for Home, Sitemap, Q & A, Notice Board, Links, and Contact Us. Below the menu, a breadcrumb trail reads "Home > Resources Database". A sidebar on the left contains a list of menu items: News, About Us, ENCePP Documents, Training in PhEpi and PV, Code of Conduct, Standards & Guidances, ENCePP Study Seal, Public Consultation, Glossary of terms, Resources Database, Partners forum, and EU PAS Register. The main content area features the heading "ENCePP Resources Database" followed by a descriptive paragraph. Below this, there is a "Search Database" button. At the bottom of the page, there is a banner for Elsevier's "International Journal of Medical Informatics", including the Elsevier logo and the journal's homepage URL: www.elsevier.com/locate/ijmedinf.

EMIF Catalogue: A collaborative platform for sharing and reusing biomedical data

José Luís Oliveira^a, Alina Trifan^a, Luís A. Bastião Silva^b

^a University of Aveiro, DETI/IEETA, Portugal

^b BMD Software, Aveiro, Portugal

Overview of methodology to define the metadata list

Diverse information available online, e.g., publications, websites, white papers

Additional information expected not to be in the public domain – unpublished (current) works, institutional or project-based tools, expert knowledge and experience

Two-part strategy to identify and collect information to derive the preliminary list:

1. Targeted web search

2. Structured interviews

1. Search for publicly available resources

Derived a list of organisations and consortia with expertise with multiple data sources for pharmacoepidemiologic research (study protocol)

Conducted a targeted web search for publicly available resources

- Organisational web pages
- Supplemented with PubMed searches

Supplemented with materials shared via interviews

Organisations and consortia

- FDA Sentinel (US)
- CNODES Canadian Network for Observational Drug Effect Studies (Canada)
- PCORnet (National Patient-Centered Clinical Research Network) (US)
- ISPE database special interest group working task DIVERSE
- Aetion
- Optum (US)
- IQVIA (Global)
- OHDSI (Observational Health Data Sciences and Informatics) (Global)
- ConcePTION, ACCESS, CONSIGN (Europe)
- EH DEN (European Health Data & Evidence Network) (Europe)
- EMIF (European Medical Information Framework) (Europe)
- FAIRplus (Europe)
- GetReal initiative (Europe)
- BD4BO (Big Data for Better Outcomes) (Europe)
- AsPEN (Asian Pharmacoepidemiology Network) (Asia)
- ISPE/ISPOR/Duke Margolis/National Pharmaceutical Council Real-world Evidence Transparency Initiative Partnership (Global)
- ENCePP (European Network of Centres for Pharmacoepidemiology and Pharmacovigilance) (Europe)
- ISoP (International Society of Pharmacometrics) (Global)
- EU netHTA (European Network for Health Technology Assessment) (Europe)
- 28 networks listed in the ENCePP database of pharmacoepidemiologic research networks will be included [<http://www.encepp.eu/encepp/search.htm>], including disease registries

Search results and extraction

57 resource items identified for extraction



Derived a “long list” of metadata variables and descriptive information

Items extracted

- *Type of resource*
- *Definition of metadata*
- *Processes for data collection, storage, handling, and access*
- *Origin of the data described; how/why it was generated*
- *General descriptors of the data or data source*
- *Data quality indicators*
- *Information on the format in which data are collected*
- *Description of any tool for access, analysis, or visualisation*
- *Any other metadata or types of metadata*
- *Additional references/resources that may be relevant*

- The items to be extracted were reviewed by coauthors
- The form was piloted by two researchers to improve consistency of information extraction

2. Structured interviews

Structured 60-minute interviews were conducted with representatives from 8 expert organisations or consortia

Interview template was piloted and revised with coauthors from IMI ConcePTION project consortium

- **FDA Sentinel**
- **CNODES**
- **IMI EHDEN**
- **IMI ConcePTION**
- **AsPEN**
- **IMI FAIRplus**
- **Maelstrom**
- **Aetion**

Interview template:

1. Definition and scope of metadata
2. Metadata collected within representative's organisation
3. Specific information on quality indicators
4. Materials describing metadata collected by organisation
5. Description of process for metadata collection and sustainability
6. Tools to record, maintain, process, or visualise metadata
7. Additional recommendations for the metadata catalogue

Synthesis of the preliminary metadata list

Minimum basic set derived based on *ENCePP Resources Database*, *EMIF catalogue*, *IMI [ConcePTION catalogue](#) design*



Metadata list formatted in the envisioned catalogue table structure



Initial scope (slide 4) of metadata was reviewed based on input from interviews



Items within scope added to the metadata list from the data extraction long list and interviews



Consortium and agency review

Preliminary metadata list – overview of catalogue tables

INSTITUTIONS

Contributors to the catalogue such as universities, companies, medical centres and research institutes

DATA SOURCES

Collections of data banks covering the same population

DATA BANKS

Data collections such as registries or biobanks

COMMON DATA MODELS

Common Data Element models and Harmonisation models

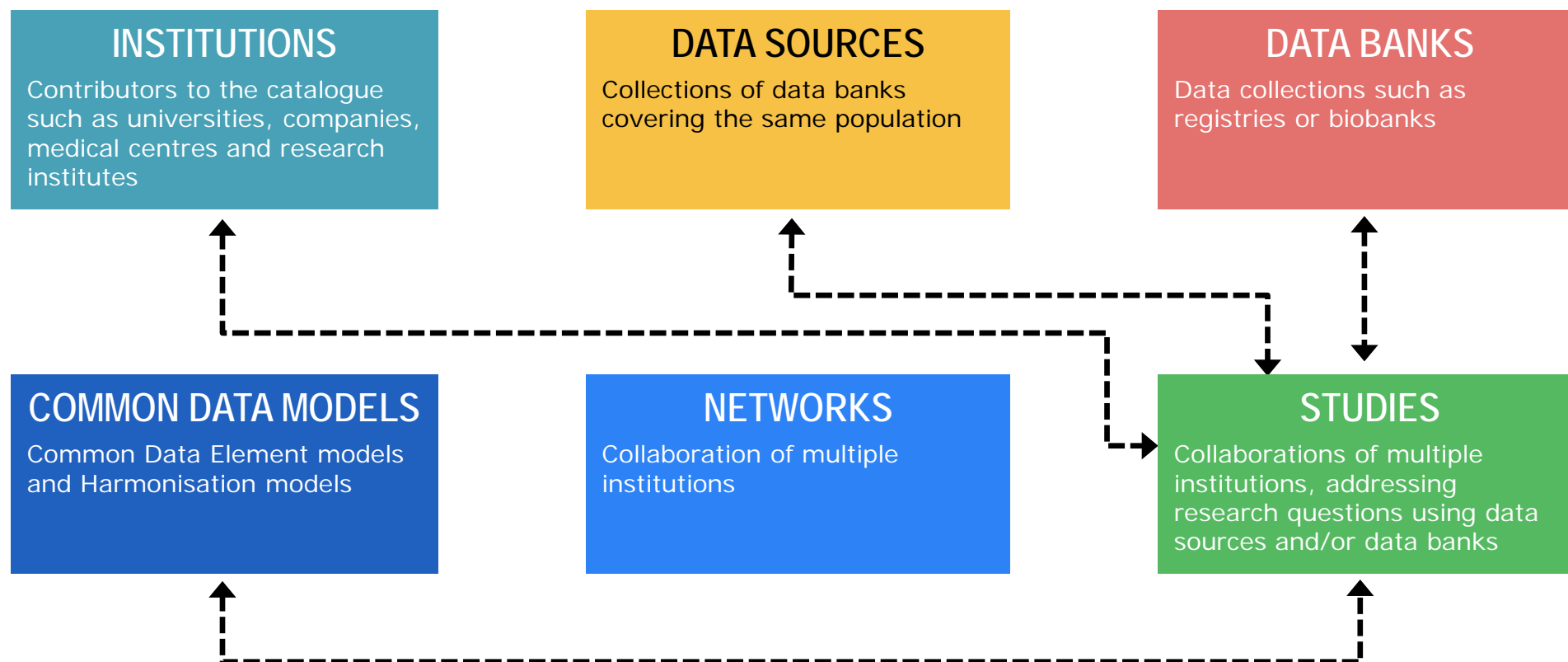
NETWORKS

Collaboration of multiple institutions

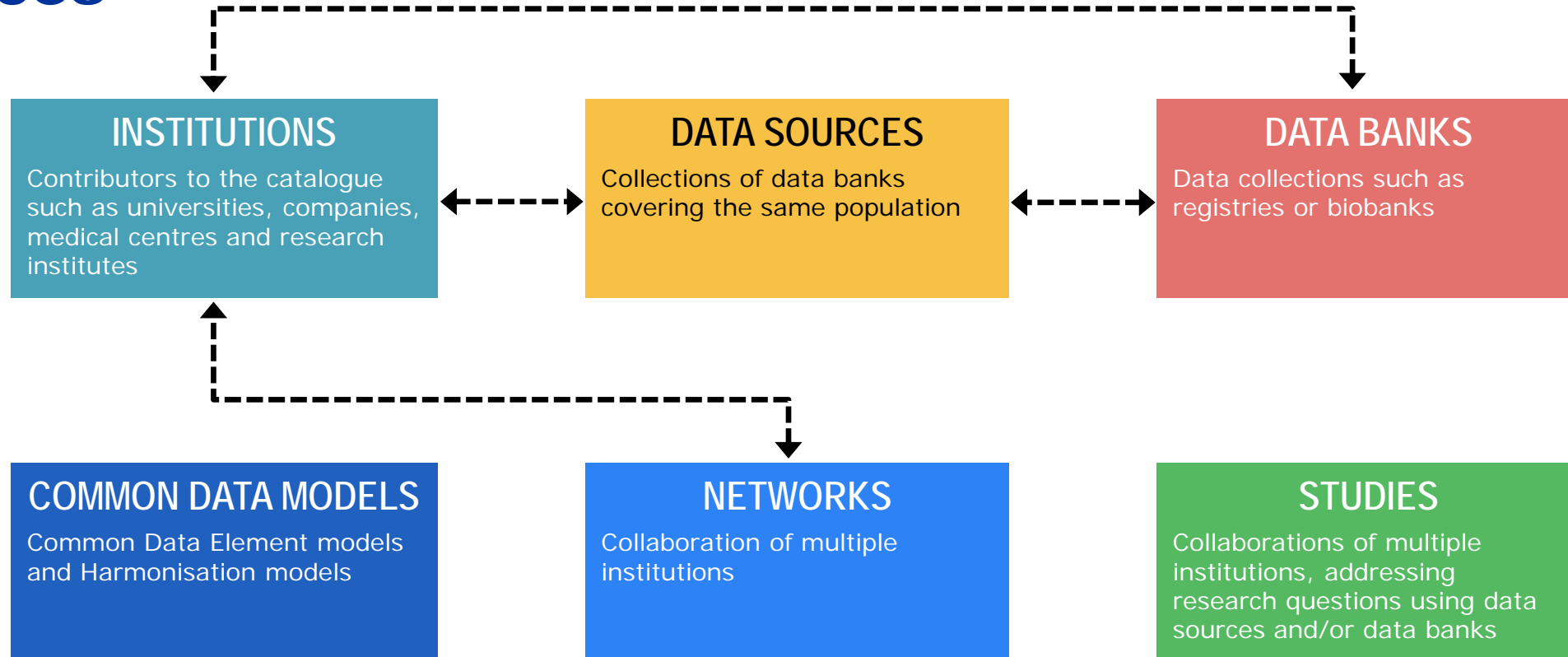
STUDIES

Collaborations of multiple institutions, addressing research questions using data sources and/or data banks

Preliminary metadata list – metadata connected to **studies**



Preliminary metadata list – metadata connected to **data sources**



0. Catalogue

Catalogue - entry: Metadata describing entry of information in the catalogue.

This will not in itself be displayed as a domain on the home page but will be used to support addition of metadata on metadata entry in each of the other tables.

E.g., *Metadata information – Date of last update: 01-01-2022*
– Name of contact person: R Pajouheshnia
– Institution/affiliation: Utrecht University

1. Institution

Institutions – role: Metadata describing an institution, its role in studies, and expertise.

Institutions – data sources: Metadata describing the data sources or data banks to which an institution has access, as well as details on whether data from only a subset of the underlying population can be accessed.

Institutions – reason for access: Metadata describing why an institution is able to access data sources, if the institution is a DAP.

Institutions – access contracts: Metadata describing how access to data sources is permitted for the institute and the time it takes between applying for access to a data source or an extract of the data source and obtaining the access, if the institution is a DAP.

Institutions – studies: List of relevant studies conducted by the institution.

INSTITUTIONS

Contributors to the catalogue such as universities, companies, medical centres and research institutes

1. Institution – metadata on data access and permissions

Institutions that act as data access providers can provide additional details on the extent and conditions of data access

- Data source/data bank name, data source subset
- Reason for access (plus description)
- Access permissions
- Access process, process time, and fees

INSTITUTIONS

Contributors to the catalogue such as universities, companies, medical centres and research institutes

2. Data source

Data source – access: Metadata describing the institutions that are able to access the data source, as denoted in the Institutions domain of the catalogue. This table links to directly to metadata provided in the *Institutions – data sources* table.

Data source – underlying population: Metadata describing the population that potentially can be captured in the data source.

Data source – quantitative descriptors: Numerical summaries of the data source population.

Data source – data banks: Metadata describing the data banks that make up the data source and summary information of their contents.

Data source – ETL: Metadata describing existing extract-transform-load specifications for mapping of data source to a CDM.

Data source – studies: Links the data source to any studies that listed it as a data source in the *Study – data sources* table.

Data source – publications: Metadata describing publications relevant to the data source, such as those that describe the contents of the data source.

DATA SOURCES

Collections of data banks covering the same population

2. Data source – content and quantification

Population(s) covered by a data source

- Data source countries, regions
- Causes of entry to and exit from the data source population
- Quantification: measurements per unit time (e.g., overall, yearly intervals)
 - (Active) population size, follow-up
 - Age, sex, and ethnicity distributions
 - Capture of biological samples

Content captured by the data source

- Data banks included
 - Types of information captured (sociodemographic, diagnoses, cause of death, prescriptions/dispensings, procedures and tests, genetic information, family linkage, health care centre/provider, free text, and processing tools)
- Linkages between data banks and their completeness

DATA SOURCES

Collections of data banks covering the same population

3. Data bank

Data bank – access: Metadata describing the list of institutions able to access the data bank, as denoted in the Institutions domain of the catalogue. This data bank links directly to the *Institutions - data sources* table.

Data bank – originator: Metadata describing the institution or body that sustains or maintains the collection of records in the data bank.

Data bank – population: Metadata describing the population that potentially can be captured in the data bank.

Data bank – quantitative descriptors: Numeric summaries describing data bank population

Data bank – prompt: Metadata describing the event(s) that trigger(s) the creation of a record in the data bank.

Data bank – data model: Metadata describing the data model of the data bank, including vocabulary.

Data bank – updates and lag time: Metadata describing the regularity of updates and time lags of the data bank.

Data bank – quality: Metadata describing qualitative descriptions of quality and qualitative and quantitative descriptors of completeness of the data bank.

Data bank – studies: Links the data bank to any studies that listed it as a data bank in the *Study – data sources* table.

Data bank – publications: Metadata describing publications relevant to the data bank, such as those that describe the contents of the data bank.

DATA BANKS

Data collections such as registries or biobanks

3. Data bank – content and quantification

Metadata variables aim to capture details of the coverage and content of information in data banks

Originator

- Type of organisation that sustains the collection of data in the data bank and the reasons why data are collected

Population covered

- Including disease subpopulations, pregnant women, and neonates
- Time span and setting of care of the data bank

Prompt

- The event(s) that trigger creation of a record in the data bank
- Including unit of observation in the data bank

Contents

- The dictionary or model of a data bank (description of the variables/fields captured)
- As with *Data sources*, the type of content captured

DATA BANKS

Data collections such as registries or biobanks

4. Common data model

Common data model – general: Metadata describing common data models utilised within studies captured in the catalogue.

Common data model – tables: Metadata describing tables included in the common data model.

Common data model – vocabulary: Metadata describing standard vocabularies used in the common data model.

COMMON DATA MODELS

Common Data Element models
and Harmonisation models

5. Network

Network – overview: Metadata describing networks/consortia linking to institutions and studies in the catalogue.

NETWORKS

Collaboration of multiple institutions

6. Study

Study – institutions: Metadata describing institutions involved in the study.

Study – protocols: Metadata describing protocols related to the study.

Study – data sources: Metadata describing data sources included in the study and software used for extraction and processing of data

Study – mappings to CDM: Metadata describing mappings that have been done to CDMs in the study.

Study – data characterisation: Quantitative descriptors of the completeness of data and numeric quality indicators, if generated within the study (e.g., level 1-3 checks on the CDM instance).

Study – results: Support standard tables in which study results could be entered.

Study – publications: Metadata describing publications generated by the study.

STUDIES

Collaborations of multiple institutions, addressing research questions using data sources and/or data banks

6. Study – study-specific quantitative descriptors

Study – data characterisation

- Could permit institutions to upload the results of rich descriptions of the completeness and quality of data
- And/or support standard tables with data characterisation metadata variables
 - E.g., based on framework of Kahn et al. (2016): conformance, completeness, plausibility

Study – results

- Descriptors of the study population (e.g., study baseline table)
- Could be supported by standard tables for entering descriptive information

Kahn MG, Callahan TJ, Barnard J, 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.

STUDIES

Collaborations of multiple institutions, addressing research questions using data sources and/or data banks

Summary

Preliminary metadata list derived, consisting of **six core metadata domains**

- Provide detailed description of sources of routinely collected electronic health data
- Content builds on existing initiatives
- Support **rich description of data sources and data banks**, potentially captured at various levels (data source, data bank, or study)
 - Includes quantitative descriptors of content and completeness
 - Implementation will be piloted in the proof-of-concept catalogue later in the project

Thank you!

For any question on this presentation, please contact: Malgorzata.Durka-Grabowska@ema.europa.eu

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