Survey results on the Data source catalogue

Rebuild of Resource database catalogue

Presented by Ana Cochino – ENCePP plenary, 18-Nov-2021

Outline of the presentation

- Background information
- Survey’s results
  - Use cases
  - Suggestions for improvement
  - Data bank/data source
  - Institutions – fields prioritisation
- Next steps
- Discussion points

Background: survey structure and objectives

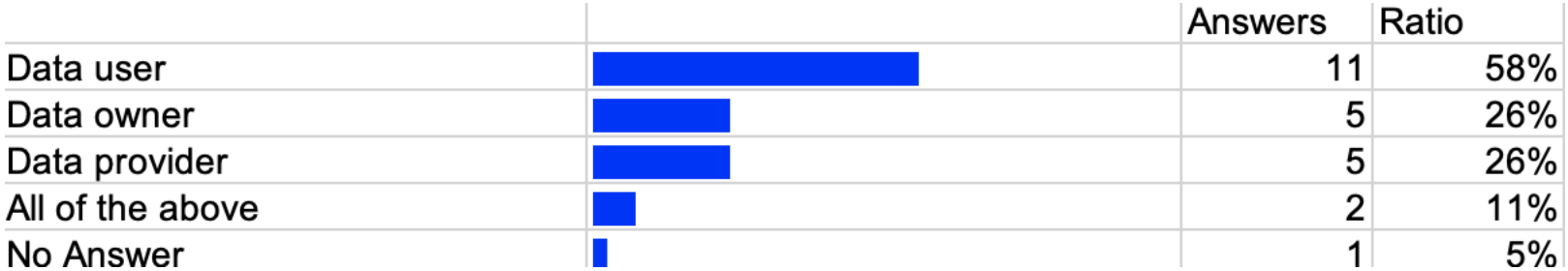
- The survey was sent out to all ENCePP members with the purpose to:
  - Understand further points for improvement of the current existing EU PAS Register and ENCePP resources database building up on existing experience
  - Get feedback on the usefulness and feasibility of some data elements proposed in the current 'metadata list' of the MINERVA project
  - Collect requirements related to functionalities and common use scenarios
- The questionnaire was structured in three sections:
  - Data sources catalogue, from the point of view of the data user
  - Data sources catalogue, as a 'data owner'
  - Studies catalogue
- The results presented here focus on more challenging areas rather than an extensive summary of all responses received

Survey results: responders

Responder roles were defined as:

- Data user - using the catalogue to search, view, export, consult the information
- Data owner - sustains the collection of records in a data bank (e.g., a healthcare payer); experience in managing databases, entering database information in the past (e.g.: in the current ENCePP resource catalogue)
- Data provider - authorised to obtain access to and/or receive extracts from one or multiple data banks (e.g., for the purpose of research or surveillance)

- 18 responders, multiple roles per responder were allowed



	Answers	Ratio
Data user	11	58%
Data owner	5	26%
Data provider	5	26%
All of the above	2	11%
No Answer	1	5%

Use cases for data source catalogue

The survey asked responders to share their view on **how they would use the catalogue**. Responses below:

<u>When reading a study:</u>

- I would need more information about the contents and coverage of a data source
- Ensure data sources used for a study are fit for purpose and reliable

<u>When planning to run a study:</u>

- Planning large scale multinational PAS studies. Details about data sources, particularly their coverage, limitations and governance would be of interest
- We would rather use it to know which data sources exist for specific diseases, and then be able to identify appropriate ones to be used for research questions, and then be able to connect with the data provider to learn more about it, discuss possible access and feasibility

Use cases for data source catalogue (continued)

When planning to run a study (continued):

- Most common use by our organization would be to find appropriate data sources for a study we plan to run, using criteria (simple and combined) such as:
  - disease and therapeutic area
  - type of data (registry, electronic medical records, claims, ...),
  - country coverage,
  - the research objectives (drug utilization study, safety, effectiveness)
  - the final stakeholder (regulatory (EMA/HMA/FDA) &/or HTA)..

General research on data sources:

- To see which data sources are available for my area of interest.
- Details related to acquiring access to a data source.
- To understand more details about the contents and coverage of a data source that I have read about or am planning to use, along with access conditions

Responders - Frequency of current use

The responders currently use the ENCePP resource database with the below frequency:

		Answers	Ratio
At least once a week		0	0%
At least once a month		5	26%
At least once every six months		7	37%
At least once a year		4	21%
Less than once a year		0	0%
Never		2	11%
Not applicable		0	0%
No Answer		1	5%

Suggestions for improvement on current tool (1/3)

Data collection and content:

- The coverage of countries; the data type is also extremely important (EMR, Disease registry, claims, surveillance data etc...)
- Improve completeness of data on the description of data sources - relevant field are missing (the information is provided on a voluntary basis and the level of detail varies)
- Expand with consistent information on content, coverage, linkage and access conditions across all registered data sources
- Some metrics could furthermore be included and derived by e.g. R-Packages

Data maintenance and sustainability

- Sustainability and update of information: need to ensure that catalogue entries are updated
- All PASS and PAES independent on the regulatory obligations should be collected

Suggestions for improvement on current tool (2/3)

Display of information:

- Have a "snapshot/table of main characteristics of data population" available for each data source
- Sections dedicated to specialties (e.g. pediatrics or rheumatology)
- More regular updates, more modern layout of platform
- Speed of access and navigation

Searchability – general functionalities proposed

- Advanced search functions could be added (e.g. by specific events of interest)
- Drop down menus are preferable and segmenting better the filters
- The browser should allow to link multi criteria of search
- Adapt the search capability to allow for a more efficient search (e.g.: including "and", "or"; data coverage, years of collection, specific disease coverage, simplification of search regarding the type of data (claims, clinical data, biomarker data...)...)
- The "data access" should be also on the front end (type of access, direct, indirect access to data, clouding system, is the data transferable to a central database or not)

Suggestions for improvement on current tool (3/3)

Specific filters and queries to be added:

- The very first filter to be added could be "therapeutic area / disease" instead of data source name, to accelerate the pre-screening of data of interest.
- Which data sources have paediatric data
- To be able to search for specific areas of interest, such as a specific disorder
- Filter: "data source that has been used for regulatory/HTA decision making"

Other functionalities:

- Data sources and studies catalogue should be functionally linked
- The web link of a specific resource should remain unchanged throughout its updates
- Usability - linkage and harmonisation with/to other catalogues / registries would reduce the burden of double data entry

Conceptual framework – ‘data bank’ vs ‘data source’

- Data sources are composed of data banks, which are data collections sustained by a specified organisation

  i.e. a payer in a healthcare system, such as an insurance company or government; a network of clinicians; a public health or another government institution

- Each data bank is defined by:
  - A specific *class of events* that prompts the creation of a record of an event (i.e. access to an emergency room; a visit to a primary care centre; the dispensing of a medication)
  - A specific *population* (i.e. the persons entitled to receive healthcare assistance funded by a specific payer; persons assisted by a specific network of primary care centres; legal inhabitants of a region or country) whose events prompt the creation of a record in the data bank
  - A *data model* and *data dictionary*

‘Data bank’ and ‘data source’ – examples

Data source: CPRD

Data banks: CPRD (GOLD or Aurum) primary care data; Hospital Episode Statistics (HES); Death Registration (Office for National Statistics. ONS); Cancer registration data, etc.

Data source: Danish National Registers

Data banks: Danish National Patient Register, Danish National Prescription Register, Cause of Death Register, Danish Medical Birth Register, etc.

Data source: ARS Toscana

Data banks: Registration with Healthcare System, Hospital Discharge Records, Exemptions from co-payment, Diagnostic Tests or Procedures Reimbursement, Pharmacy Dispensation Records

Data source: The Estonian Biobank

Data banks: The Estonian Biobank, Population register, Estonian Causes of Death Registry, Estonian Cancer Registry, Estonian Tuberculosis Registry, Estonian Health Insurance Fund

Data bank/data source – survey results and feedback (1/2)

		Answers	Ratio
Concept is clear		11	58%
Concept not clear		7	37%
Cannot say		0	0%
No Answer		1	5%

- It seems to be over-granular/unnecessary in some cases because redundancy between both
- It is not immediately clear - why do you want to distinguish between data bank and data source?
- I don't understand if "data source" is a search term or an existing data source. So, for example, if I enter "Danish National registers", do I get all national registers in Denmark listed, or is there already an existing data source where all registers are summarized. Same accounts for the Estonian biobank. Why are all those data banks listed? Are all data banks listed which contain "Estonian" in their name? Why are then not all listed with "biobank"? Or does already a data source "The Estonian biobank" exists which contains all the named data banks? But why is the Estonian Biobank then also listed as a data bank. It cannot be both, a data source and a data bank?

Data bank/data source – survey results and feedback (2/2)

- The distinction between both concepts would be useful. However, I doubt that the terms "data banks" and "data sources" will be easily understood by "naive researcher" or people outside in Europe. A suggestion my be to replace "data banks" by "datasets". Also, a careful consideration of the description of population covered in "data banks" compared to "data sources" will be required

- In this context, the terminology of data source would be too broad in most of the cases, as when you will use the "CPRD" or the "Danish registers", not all data banks will be used for a study. So you will still have to detail what has been included into the "data source" for your study, and maybe confusing as for another study with the same data source, it may include different data banks?

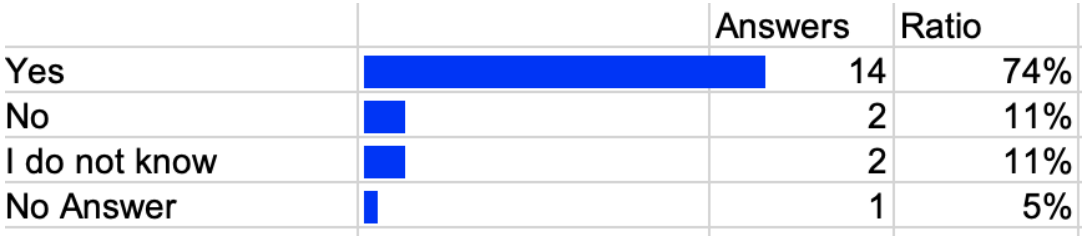
- Individual differences not fully considered, may not be a meaningful distinction for all data sources

- The concept to structure information in this way is helpful, however what is the difference then between data bank and data base (in some languages they might be identical?

- Yes, it is clear and it is helpful (disclosure: part of the MINERVA Consortium). We believe that this framework helps distinguish terms that in the past were often confused. For example, the term 'database' is a more technical ('IT') concept that refers to the tables (rows and columns) within a databank.

Survey results – Information on Disease specific registries

Do you have a particular interest in disease specific registries?



	Answers	Ratio
Yes	14	74%
No	2	11%
I do not know	2	11%
No Answer	1	5%

Suggestion of diseases of particular interest highlighted during the survey:

- Cancers (haematology, solid)
- Rare and ultra rare diseases
- Paediatric rheumatology diseases
- Multiple sclerosis
- Cystic Fibrosis

Therapies of special interest – values proposed

A new variable 'categories of therapies of special interest' is proposed. Initial values proposed (MINERVA/EMA):

- Advanced therapy medicinal products (medicines for human use that are based on genes, tissues or cells)
- Vaccines
- Contraception/emergency contraception
- Injectable/infusion medicines

Further proposals collected during the survey are:

- Medical devices
- Biological
- Oral oncology medicines
- Biosimilar(possibility of distinguishing between originator and biosimilars)
- Public health interventions
- Surgeries
- Pediatric medication
- Rheumatology therapies
- Chemotherapy
- oxygen therapy, physical therapy
- Anti-inflammatory biologics
- Oncological drugs
- Orphan drugs,
- Diagnostic imaging and contrast agents

Proposed data elements in 'Institutions' - prioritisation

Institution: an organisation that contributes scientifically to research studies that involve one or more data sources, such as a DAP, a data originator, or an institute with analytic expertise that contributes to studies in the catalogue

- A large amount of information is proposed to be collected on Institutions – it is likely that it needs prioritisation (at least in an initial development phase). The question below aims to establish a priority of information.

Proposals of information to be deprioritized (implemented at a later stage):

- Details on the role of an institution and expertise
- Specific details on the access of an institution to a given data source (data bank)

Likelihood of querying the following information with regards to 'Institutions':	Very Likely	Likely	Neither likely nor unlikely	Unlikely	Very Unlikely
The name, the address, the contact details and the type of **institution**	8	9	0	1	0
The **medical specialty, expertise in studies, institution's resources** (e.g.: pharmacoepidemiologist, legal experts etc)	5	6	2	3	1
Experience/expertise of the institution with particular data sources	5	5	4	2	1
Information related to th**e access available to a given institution** (e.g.: the reason or conditions for access to a specific data source)	7	5	3	2	1
View the **list of studies** conducted by an institution	4	10	2	2	0

Proposed data elements in 'Institutions' - prioritisation

Is the list of institutions (e.g. :universities, companies, contract research organisations etc.) that have been granted data access useful information in your activity?		Answers	Ratio
Yes		8	42%
No		8	42%
No Answer		3	16%

Further comments received on the use cases for information:

- The contact details are used to reach to institution to understand the model of scientific collaborations, the model of data management, transfer, use of private cloud or not, data protection and restrictions, data handled outside the institution and licensed
- This would be useful to identify potential collaborators with experience in the use of a specific data source for the conduct of studies
- This information is very useful in the case of study feasibility assessments.
- For framework contract tenders to know if contractors have access to particular data sources and for which purposes (research topic) these were used

Next steps – use cases, improvements, suggestions

Use cases:

- EMA taking the information on board and enhance business case documentation and high-level requirements; enhance understanding on user's need from the system

Improvements:

- Most suggestions are attainable and are part of migration to a new technology (e.g.: faster system, more user-friendly, advanced search criteria, additional fields and filters, connections between studies and data source catalogue etc.)
- Others (mostly process related) are more challenging to achieve:
  - Expand with consistent information on content, coverage, linkage and access conditions across all registered data sources
  - Sustainability and update of information: need to ensure that catalogue entries are updated
  - Linkage and harmonisation with/to other catalogues / registries would reduce the burden of double data entry
  - PASS and PAES in the risk management plan – a proposal for an amendment of the legislation has been put forward to make their registration in the catalogue mandatory

Discussion points – data banks vs data source

Slido question 1:

Following the discussion on concepts behind the data source/data bank concept – which terminology seems to fit best?

a) data source – data bank

b) data source – dataset

c) data source – database

Slido question 2:

Terminology aside – do you find the conceptual split:

a) Very useful (implement catalogue structure focusing around these concepts)

b) Useful, but in a small measure (collect minimal information on data sources)

c) Too granular (think alternative solutions)

Open discussion:

- Can it be done in a different way (any proposals from the group)?

Discussion points – data elements needing feedback

Slido question 3:

Would you like to have information on the vocabularies used in the new database catalogue for the following variables? (Note, this question does not refer to content – we will collect information on content (e.g.: if the data source contains information on 'cause of death'), the question asks about the need-to-know *which vocabulary* is used

a) Cause of death (e.g.: ICD, Read, SNOMED, MedDRA…)

b) Prescriptions (e.g.: RxNorm/ EphMRA/ ALT/ DrugBank)

c) Procedure results (e.g.: ICD, Read, SNOMED, MedDRA…)

d) Active substance information, additional to medicinal product information (e.g.: RxNorm, ATC, Art 57)

e) Common Data Model (e.g.: what vocabularies are used for events, such as diagnoses, procedures etc., specifically related to the common data model)

f) Genetic data (e.g.: OGG/ FG/ GO/ EGO/ SOPHARM/ PHARE)

Discussion points – data elements needing feedback

Slido question 4:

Would you like to have information on the following variables in the new database catalogue?

a) Population size by age group

b) Data source funding source (own institution, industry, European public funding,…)

c) Detailed information on data access (eg. reason for requesting access, access fee, time to process request for access, possibility to access individual patient level data, etc.)

d) Median observation time of the population in the data source (i.e.: median time for which unique individuals with records captured in the data source are observable)

e) Median age of the population in the data source

f) Distribution of population by gender of the population in the data source

g) Description of dispensing data (additional to capturing prescription data)

h) Details on ETL process and CDM (list of ETL specification documents, Version of CDM to which the data source has been ETL-d)

Any questions?

Further information

[Insert relevant information sources or contact details as applicable.]

Official address Domenico Scarlattilaan 6 ● 1083 HS Amsterdam ● The Netherlands
Telephone +31 (0)88 781 6000
Send us a question Go to www.ema.europa.eu/contact

Follow us on **@EMA_News**