IMI2 821520 - ConcePTION

ConcePTION

WP2 – Improving the collection, analysis and interpretation of pregnancy pharmacovigilance data

D2.2 Report describing the metadata model (variables) for data collection on pregnancy data sources

Lead contributor	E.P. van Puijenbroek; Pharmacovigilance Centre Lareb
	e.vanpuijenbroek@lareb.nl
Other contributors	B. te Winkel; Pharmacovigilance Centre Lareb F. Dumas-Sillan; Pfizer L. Yates; KRISP, University of KwaZulu-Natal

Due date	30 November 2019
Delivery date	8 November 2019
Deliverable type	R
Dissemination level	PU

Description of Work	Version	Date
	1	1 April 2019

Document History

Version	Date	Description	Non-contributor reviewers (if applicable)
V0.1	27 Aug 2019	Comments	M.A. Swertz, Molgenis-support F. de Andrade Molgenis-support R. Bromley, University of Manchester F. Dumas-Sillan, Pfizer M.S.F.E. Stellfeld, Novonordisk
V0.2	20 Sept 2019	Draft	B. te Winkel and E.P. van Puijenbroek, Pharmacovigilance Centre Lareb
V0.3	04 Oct 2019	Comments	M. Cunnington, GSK C Dodd, University Medical Centre Utrecht F. Dumas-Sillan, Pfizer L Greener, TEVA H.M.E. Nordeng, University of Oslo J.L. Richardson, UK Teratology Information Service , Newcastle upon Tyne Hospitals NHS Foundation Trust and Public Health England Peter Singleton, MHRA M.S.F.E. Stellfeld, Novonordisk L. M Yates, KRISP, University of KwaZulu-Natal
V0.4	10 Oct 2019	Final check	B. te Winkel and E.P. van Puijenbroek, Pharmacovigilance Centre Lareb
V1.0	24 Oct 2019	Final Version	B. te Winkel and E.P. van Puijenbroek, Pharmacovigilance Centre Lareb

Summary

This report describes the content and development of the metadata model that was designed within task 2.2. for the purpose of collecting and cataloguing data sources in which data have been specifically collected for the purpose of assessing the use of a medication or medications in pregnant and /or breastfeeding women. In this task an inventory of industry and publicly held primary data sources and handling processes will be created. Information detailing these datasets will be stored in an online searchable and updatable data catalogue that describes the nature of the data collected, including clinical variables and data handling, as well as the data strategy and management. The data catalogue will cover the collection of various aspects, among which relevant characteristics, variables, data formats enabling a proper assessment of the information present and data on the provenance of the data sources. Finally, details will be collected from the various sources in respect to data storage, handling processes and governance. The metadata model described in this report will also be used to build the survey for populating the data catalogue.

Contents

Introduction

A key component of ConcePTION is the establishment of a framework that includes a FAIR data catalogue (findable, accessible, interoperable and re-usable)[1] of data sources including standardized workflows and tools to guarantee analytical quality. The systematic identification and recording of primary data sources that have been established for the purpose of assessing medication use in pregnancy and/or lactation. The use of an online, searchable catalogue is thus key to ensuring that these datasets and their key characteristics are visible to researchers and regulators and can be accessed in a timely manner if required.

In task 2.2 a searchable catalogue of industry and publicly held primary data sources and handling processes will be created. Once identified, public institutions and pharmaceutical companies with relevant data sources will be approached, and asked to provide details of relevant data source architecture and data collection processes via a structured electronic survey. The information provided will be deposited in a data catalogue that describes the variables collected, the data collection structure, the company or organization's data strategy and management. The catalogue is intended to capture information on human pregnancy and lactation datasets held in regulatory, industry and teratology information service databases, as well as drug or clinical disease-based registries. These datasets contain a mixture of prospectively reported pregnancy exposures and retrospective case reports like spontaneous reports. In addition to pregnancy specific data collections, case reports of exposed pregnancies data that have been collected incidentally as secondary outcomes through various other sources will be captured. The data catalogue will be a fundamental step in the ConcePTION project towards identifying, acquiring, curating, storing, managing and governing the currently fragmented and possibly underutilized pregnancy and lactation datasets.

This report describes the metadata model for the survey underpinning the catalogue of pregnancy and breastfeeding exposure data sources, as part of task 2.2. The creation of the data catalogue and the survey itself will be carried out by BBMRI_ERIC within WP7.

Harmonization with other workpackages and tasks

The development of the metadata model as part of task 2.2. integrates with various tasks in several workpackages, including:

Task 1.1 in which population based databases with potential for secondary use of data for the purpose on assessing medication use in pregnant and/or breastfeeding women sources will be identified and added to the ConcePTION data source catalogue, via an aligned survey designed by workpackage 1 colleagues.

Task 2.1 in which the process needs, barriers and information requirements for various stakeholders will be described. As part of this tasks stakeholder meetings will be held and used to identify and engage potential data owners or data providers, in order to obtain detailed information about the data source.

Task 2.3 defines and describes the core data elements required for the prospective collection and follow-up of exposed pregnancies. Although the data elements for long term outcomes and more complex confounding maternal factors under consideration in this task are only due for completion in M36, information relating to the more routinely collected core data elements will be implemented in the survey and metadata model when possible.

Task 7.4 in which the data source catalogue structure will be created, based on the meta

datamodel described in this report with specific search features that will be defined based on the user requirements identified.

Existing data source catalogues

In the past several basic inventories have been made in which data collections have been described. Examples include the EUROmediSAFE inventory[2] and the EMA inventory[3]. Both are not recently updated and have no adequate search options. Details of the data sources which have already been recorded within these resources will be transferred to the ConcePTION data source catalogue to avoid duplication of information collection and to achieve the goal of having a single, recognized and trusted repository of information. However, the nature of the information collected in these earlier inventories differs from the variables identified for the data catalogue to be developed for ConcePTION. In developing the metadata model for the catalogue, consideration was given to the structure and content of the existing data collections.

- EUROmediSAFE, an inventory of available data sources in all 28 EU Member States (2018) for potential use when evaluating the perinatal and long-term childhood risks associated with in-utero exposure to medication.[2]
- EMA systemic overview of data sources for drug safety in pregnancy research.[3]

Elements to be captured in the metadata model

In order for the catalogue to be a useful reference source, the information collected in the metadata model will describe various characteristics of the pregnancy and/or lactation exposure datasets. [4,5] In addition to information describing the location and ownership of the dataset, details of key clinical variables and outcomes, along with the format in which the data are collected will be captured to enable accurate identification and qualification of the exposure and outcome information available. In addition, the provenance of the data should be catalogued. This refers to clearly documenting the input, systems, and processes that influence data of interest. Finally details will be collected of the various data sources in respect to storage, handling processes, access and governance. A concise overview of the topics to be addressed in relation to the aforementioned points is provided below. This overview is neither exhaustive nor prescriptive but gives guidance for the topics to be addressed in the meta data model.

A. Aspects describing the source as a whole and its datasets
    1. Details of the organisation like name and (website) address, and type (e.g. MAH, PV centre, research centre)
    2. Name, function and contact details of responsible contact persons
    3. Nature of the data collection (e.g. spontaneous reports, registry, medication and/or disease specific)
    4. Characteristics of the data collection (e.g. sample size, in- and exclusion criteria, time period and population covered, who provides the data?)
    5. For spontaneous reports: possibility to identify cases of exposure during pregnancy and lactation
    6. For pregnancy cohorts: number and points in time of data collection; comparison group
B. Variables and data formats enabling the identification of the information present in respect to
    1. Medication, e.g. dosage, timing of exposure, indication
    2. Type of foetal and pregnancy outcomes captured, such as: pregnancy-, delivery-complications and short- and long-time health related information of the child.

3. Type of outcomes for breastfeeding, e.g. production of breastmilk, and short- and long-time health related information of the child.
4. Miscellaneous information of the mother like social or economic aspects and health related information, e.g. comorbidities, medical-, family- and obstetrical history, concomitant mediation and exposure other possible teratogens.
5. Miscellaneous information of the pregnancy, e.g. nature of conception (natural, ART), LMP, EDOB and outcome of prenatal screening

C. Provenance of data
   1. Application of specific regulations/terminology/definitions in relation to the various variables that are collected and those derived (like EDOB and timing of exposure), and validation procedures in place
   2. Information on systems in place for coding e.g. medication and outcomes, co-variables
   3. Possibility to ask for additional information

D. Storage, Handling processes and Governance of the data of the various resources
   1. Information on the type of database (e.g. spontaneous reporting system, registry)
   2. Information on access, backup, security and storage
   3. Data cleaning procedures and duplicate detection
   4. Current signal detection methods and screening procedures
   5. For spontaneous reports: export to national PV centres, EMA (Eudravigilance) or WHO-UMC (Vigilyze)
   6. Possibility for linkage to other data sources; legal and technical aspects
   7. Ethics and consent
   8. Documentation on metadata of the data collection and data management plan

Development of metadata model

A first draft of the survey to capture key information from data owners or providers was developed by members of task 2.2. The content of this survey was decided by review of fields within existing catalogues, along with input from members of the ConcePTION network who are actively involved in pregnancy PV in order to collect sufficient details of data sources to enable end-users to efficiently assess the potentially utility of the dataset for a specific purpose. Since the metadata model was not only for the development of the data catalogue, but also for the development of the survey needed to collect the data, the format for building both questionnaire and data catalogue was discussed with BBMRI-ERIC. According to the instructions received and in collaboration with BBMRI-ERIC an MS Excel spreadsheet was created for a description of the metadata model requirements. In early September, a revised version of the questionnaire was sent to all members of task 2.2. All remarks and comments were taken into account, and the revised version of the spreadsheet was then circulated to the wider members of WP2, as well as WP1 and WP7 for additional comments. This resulted in the final version that is presented in this deliverable.

Description of metadata

The description of the variables follows the FAIR principles. This implies that in the development of the meta data model, it was ensured that data should be Findable, Accessible, Interoperable and Reusable. Based on the description of the metadata model a survey will be built that will be sent to relevant stakeholders.

The metadata model is presented as an Excel sheet which is part of the current report. The metadata model is described in two tables/sheets (Addendum 1). In the first table “attributes” the various attributes to be collected and their characteristics will be described. The second table

"categorial_mref" describes the various reference categories and their possible values. The metadata model itself is a separate excel file.

References

1 H2020 programme. Guidelines on FAIR DATA management in Horizon 2020.
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

2 Gorki M, Morris J. Inventory of available data sources in all 28 EU Member States for potential use when evaluating the long-term risks for children associated with in-utero exposure.
http://www.euromedicat.eu/content/EUROmediSAFE%20Inventory_Finalv2_2018_07_06.pdf

3 Charlton R, de Vries C. Systematic overview of data sources for drug safety in pregnancy research Consultancy EMA/2010/29/CN prepared for the European Medicines Agency, June 2012 updated for ENCePP June 2016.
ww.encepp.eu/structure/documents/Data_sources_for_medicines_in_pregnancy_research.pdf

4 Task 7.4: Creation and filling of a FAIR data source catalogue(s); Full proposal ConcePTION project task 7.4 page 221

5 WP7: Information and data governance, ethics, technology and data catalogue and quality support Full proposal ConcePTION project task 2.2 page 188

List of abbreviations

ART	Assisted Reproduction Techniques
EDOB	Estimated Date Of Birth
LMP	Last Menstrual Period
MAH	Marketing Authorization Holder
PV	PharmacoVigilance
EMA	European Medicines Agency
WHO-UMC	World Health Organization Uppsala Monitoring Center

Addendum I Description of tables used in the metadata model

Datasheet “attributes”

Section	Main section of the data-catalogue and survey	The data-catalogue consist of three sections: organisation; spont reporting data and pregnancy cohort.
Metadata type	One of the four main topics mentioned in “Elements to be captured in the metadata model ” shown on page 6 of this document	
Datatype chapter	Corresponding number of the subtypes mentioned in “Elements to be captured in the metadata model ” shown on page 6 of this report	
Name	Name of attribute	
Label_datacatalogue	Label to be used in the data-catalogue	
Questionnare section	Section to be used in the questionnaire, in which corresponding items can be grouped	
Label_questionnaire	Label to be used in the questionnaire	This is the actual question that will be shown in the questionnaire
Explanation	Additional explanation for the questionnaire	Will be shown when clicking on a button for additional information next to the actual question
Optional choices	Possible answers that can be chosen for a specific question	Collated topics of sheet “categorial_mref”
Single/multiple selection	Indication if one or multiple answers may apply for a question	
Datatype	Type of data allowed	
Ref entity	Refers to the reference category in datasheet “categorial_mref” to be used	
Conditions	Shows under which condition, the question is shown	Allows for convenient navigation based on answers given
Actions	Shows which condition should be followed by what action	Allows for convenient navigation based on answers given
Nillable	Allow for value nill?	Is the question mandatory or not?
Remarks		

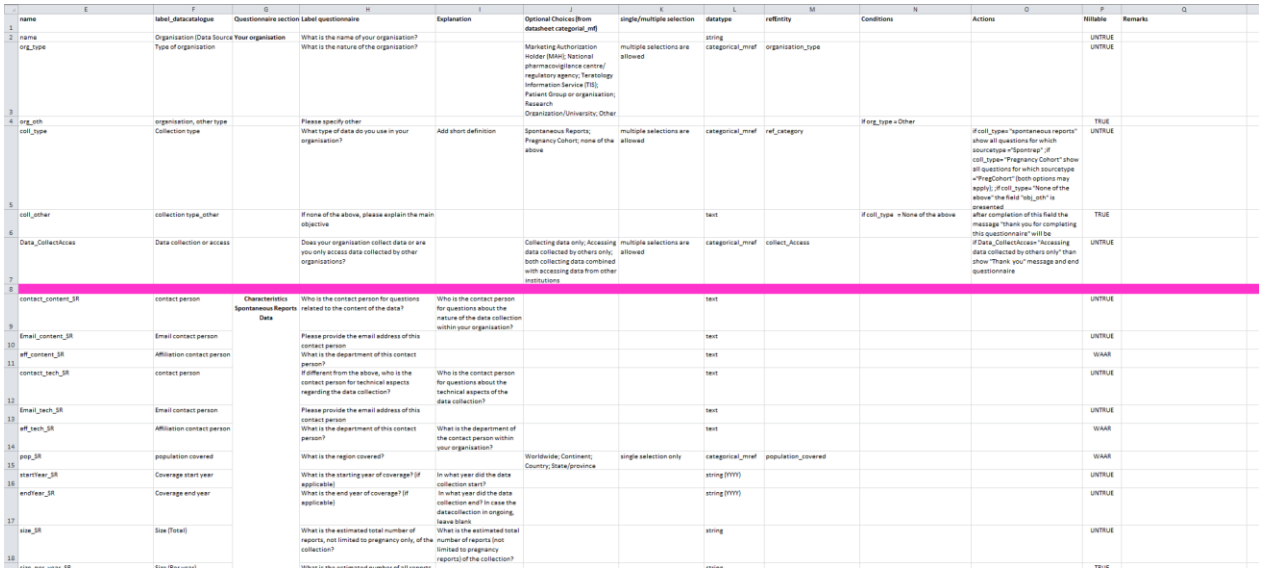
Datasheet catergorial_mref

Ref_entity	Name of the reference category	e.g. adverse_outcome or exposure_source
options	Describes the individual values that are allowed within these categories	

Addendum II: Metadata model

The data-catalogue and survey consist of three sections: Organisation; Spontaneous reporting data and Pregnancy cohorts. Only the latter section will be repeated once an organisation has multiple cohort studies for which data are needed.

The metadata model itself is a separate excel file: WP2.2Metadatamodel_datacollection.xlsx



	name	label_datacatalogue	Questionnaire section	Label questionnaire	Explanation	Optional Choices (from datasheet categorial_mf)	single/multiple selection	datatype	reflintity	Conditions	Actions	Nillable	Remarks
2	name	Organisation (Data Source	Your organisation	What is the name of your organisation?				string				UNTRUE	
3	org_type	Type of organisation		What is the nature of the organisation?		Marketing Authorization Holder (MAH); National pharmacovigilance centre/ regulatory agency; Teratology Information Service (TIS); Patient Group or organisation; Research Organization/University; Other	multiple selections are allowed	categorical_mref	organisation_type			UNTRUE	
4	org_oth	organisation, other type		Please specify other						If org_type = Other		TRUE	
5	coll_type	Collection type		What type of data do you use in your organisation?	Add short definition	Spontaneous Reports; Pregnancy Cohort; none of the above	multiple selections are allowed	categorical_mref	ref_category		if coll_type= "spontaneous reports" show all questions for which sourcetype ="Spontrep" ;if coll_type= "Pregnancy Cohort" show all questions for which sourcetype ="PregCohort" (both options may apply) ; if coll_type= "None of the above" the field "obj_oth" is presented	UNTRUE	
6	coll_other	collection type_other		If none of the above, please explain the main objective				text		if coll_type = None of the above	after compilation of this field the message "thank you for completing this questionnaire" will be	TRUE	
7	Data_CollectAcces	Data collection or access		Does your organisation collect data or are you only access data collected by other organisations?		Collecting data only; Accessing data collected by others only; both collecting data combined with accessing data from other institutions	multiple selections are allowed	categorical_mref	collect_Access		if Data_CollectAcces= "Accessing data collected by others only" than show "Thank you" message and end questionnaire	UNTRUE	
8													
9	contact_content_SR	contact person	Characteristics Spontaneous Reports Data	Who is the contact person for questions related to the content of the data?	Who is the contact person for questions about the nature of the data collection within your organisation?			text				UNTRUE	
10	Email_content_SR	Email contact person		Please provide the email address of this contact person				text				UNTRUE	
11	eff_content_SR	Affiliation contact person		What is the department of this contact person?				text				WAAR	
12	contact_tech_SR	contact person		If different from the above, who is the contact person for technical aspects regarding the data collection?	Who is the contact person for questions about the technical aspects of the data collection?			text				UNTRUE	
13	Email_tech_SR	Email contact person		Please provide the email address of this contact person				text				UNTRUE	
14	eff_tech_SR	Affiliation contact person		What is the department of this contact person?	What is the department of the contact person within your organisation?			text				WAAR	
15	pop_SR	population covered		What is the region covered?		Worldwide; Continent; Country; State/province	single selection only	categorical_mref	population_covered			WAAR	
16	startYear_SR	Coverage start year		What is the starting year of coverage? (if applicable)	In what year did the data collection start?			string [YYYY]				UNTRUE	
17	endYear_SR	Coverage end year		What is the end year of coverage? (if applicable)	In what year did the data collection end? In case the datacollection in ongoing, leave blank			string [YYYY]				UNTRUE	
18	size_SR	Size (Total)		What is the estimated total number of reports, not limited to pregnancy only, of the collection?	What is the estimated total number of reports (not limited to pregnancy reports) of the collection?			string				UNTRUE	

Figure 1 Screenshot of the metadata model excel file