IMI2 821520 - ConcePTION

ConcePTION

WP7 – Information and data governance, ethics, technology, data catalogue and quality support

D7.16 Report with publications on results of component algorithms

Lead contributor	Rosa Gini (10 – ARS Toscana)
	Email rosa.gini@ars.toscana.it
Other contributors	Giorgio Limoncella (Org number – UNIFI)
	Sue Jordan (23 – Swansea University)
	Sandra Lopez (Org number – Novartis)
	Giuseppe Roberto (10– ARS Toscana)
	Anna Girardi (10 – ARS Toscana)
	Olga Paoletti (10 – ARS Toscana)
	Claudia Bartolini (10 – ARS Toscana)
	Leonardo Grilli (Org number – UNIFI)
	Emanuela Dreassi (Org number – UNIFI)
	Carla Rampichini (Org number – UNIFI)
	Daniel Thayer (23 – Swansea University)
	Hywel Turner Evans (23 – Swansea University)
	Alex-Ioan Coldea (23 – Swansea University)
	Miriam Sturkenboom (1 – UMCU)
	Judit Riera Arnau (1 – UMCU)
	Carlos Duran Salinas (1 – UMCU)
	Caitlin Dood (1 – UMCU)

Due date	31 August 2024

Delivery date	31 August 2024
Deliverable type	R
Dissemination level	PU

Description of Work	Version	Date
	V1.0	24 May 2024
	V2.0	19 August 2024

Document History

Version	Date	Description	Non-contributor reviewers (if applicable)
V1.0	24 May 2024	First Complete Draft	
V1.1	12 June 2024	Comments by contributors	All contributors
		Consistency review	Constanza Andaur Navarro (1-UMCU)
V1.2	28 June 2024	Consolidated comments	Olga Paoletti (10-ARS) Giuseppe Roberto (10-ARS) Giorgio Limoncella (10-ARS) Anna Girardi (10-ARS)
V2.0	14 August 2024	Final Version	Rosa Gini

Summary

In this deliverable, the ConcePTION framework is utilized to develop strategies to fully exploit data diversity, in four areas.

First, identifying the full list of pregnancies that occurred in the population represented in an instance of a data source. The work on this topic has profited from collaboration with organizations external to ConcePTION. The operationalisation of this work has been stored in the ConcePTION Pregnancy Algorithm, an open-source tool that has been applied already in multiple studies inside and outside the project itself (ConcePTION Pregnancy Algorithm wiki), and a manuscript is undergoing finalisation that collects results from 8 European data sources.

Second, designing and developing a tool to allow investigators to extract from data sources the number of days of treatment associated with prescribing or dispensing a medication. This work derived an open-source function, named CreateDoT (CreateDoT wiki) and a manuscript is under development.

Third, analysing strengths and limitations of the scarce information on breastfeeding available in the data sources participating in ConcePTION. This work is being use in the Demonstration Project 2 of ConcePTION WP1.

Fourth, developing tools to address misclassification, particularly lack of sensitivity in algorithms used to indicate occurrence of a healthcare condition. A manuscript on this work has been accepted for publication in American Journal of Epidemiology (*Limoncella et al, 2024).*

Table of Contents

Chapter 1. Data diversity in ConcePTION

1.1 Introduction

ConcePTION aims to contribute to filling the knowledge gap regarding the effects of medicines in pregnancy and lactation, by developing a system across European Data Access Partner (DAPs) that transforms existing and routinely collected healthcare data into evidence in a robust and transparent manner.

Diversity across such data sources in Europe poses challenges, that can only be addressed if diversity is firstly, acknowledged, and secondly, embraced to convert it into opportunities.

In previous work (*Deliverable 7.5, Thurin et al, 2022*), the project investigated diversity across European data sources and set up a conceptual framework to describe it. The framework has been used to design and populate the ConcePTION Catalogue (Deliverable 7.10), which has been used by the MINERVA Project to create recommendations (*Pajouheshnia et al, 2024, Gini et al. 2024a*) for the recently launched HMA-EMA Catalogues of data sources and studies (HMA-EMA Catalogues), and for the VAC4EU Catalogue (https://vac4eu.org/catalogue/). The framework has been compared with other similar frameworks and methods that represent primary data collection datasets, where it has been found complete and compatible (*Swertz et al, 2023*).

Finally, a framework to represent data diversity has been recently introduced in the context of the DIVERSE initiative funded by the International Society of Pharmacoepidemiology (*Gini et al, 2024b*). In this work, it is argued that, to ensure reproducibility of study findings, representation of data diversity is necessary and complementary to an accurate representation of study design and to transparent implementation. Indeed, in multi-database studies based on common data models and common analytics, where study design and implementation are identical, results across different data sources are often heterogeneous, even when population differences are implausible. The DIVERSE framework proposed nine dimensions to represent diversity across data sources: organization accessing the data source, data originator, prompt, inclusion of population, content, data dictionary, time span, healthcare system and culture, and data quality.

The ConcePTION framework is fully compatible with the DIVERSE framework and is mentioned in the DIVERSE manuscript as a valuable source for possible ontologies on several dimensions.

In this deliverable, the ConcePTION framework is utilized to develop strategies to fully exploit data diversity, in four areas. First, identifying the full list of pregnancies that occurred in the population represented in an instance of a data source. The work on this topic has profited from collaboration with organizations outside the ConcePTION consortium. The operationalisation of this work has been stored in the ConcePTION Pregnancy Algorithm, an open-source tool that has been applied already in multiple studies inside and outside the project itself (ConcePTION Pregnancy Algorithm wiki), and a manuscript is undergoing finalisation that collects results from 8 European data sources.

Second, designing and developing a tool to allow investigators to extract information from data sources regarding the number of days of treatment associated with prescribing or dispensing a medication. This work originated from an open-source function, named CreateDoT (CreateDoT wiki) and a manuscript is under development. Third, analysing strengths and limitations of the scarce information on breastfeeding available in the data sources participating in ConcePTION. This work is being used in the Demonstration Project 2 of ConcePTION WP1.

Fourth, developing tools to address misclassification, namely, lack of sensitivity in algorithms used to indicate occurrence of a healthcare condition. A manuscript on this work has been publication in American Journal of Epidemiology (*Limoncella et al, 2024)*

1.2 References

(Deliverable 7.5) Dodd C, Gini R, Sturkenboom M., et al. Report on existing common data models and proposals for ConcePTION (D7.5). Zenodo. 2020. https://zenodo.org/records/5829417

(Deliverable 7.10) Swertz, M, & Hyde, E. Test report for FAIR data catalogue 2nd (D7.10). Zenodo. 2023. https://zenodo.org/records/7568799

(Gini et al, 2024a) Gini R, Pajouhesnia R, Gutierrez L, et al. Metadata for data dIscoverability aNd study rEplicability in obseRVAtional studies: lessons learnt from the MINERVA project in Europe. Pharmacoepidemiology Drug Saf. 2024

(Gini et al, 2024b) Gini R, Pajouheshnia R, Gardarsdottir H, Bennett D, Li L, Gulea C, et al. Describing diversity of real world data sources in pharmacoepidemiologic studies: The DIVERSE scoping review. Pharmacoepidemiology and Drug Safety. 2024;33(5):e5787.

(Limoncella et al, 2024) Limoncella G, Grilli L, Dreassi E, Rampichini C, Platt R, Gini R. Addressing bias due to measurement error of an outcome with unknown sensitivity in database epidemiological studies. AJE, 2024.

(Pajouheshnia et al, 2024) Pajouheshnia R, Gini R, Gutierrez L, et al. Metadata for data dIscoverability aNd study rEplicability in obseRVAtional studies: definition and recommendations of use from the MINERVA project in Europe. Pharmacoepidemiol Drug Saf. 2024

(Swertz et al, 2022) Swertz M, Enckevort E van, Oliveira JL, Fortier I, Bergeron J, Thurin NH, et al. Towards an Interoperable Ecosystem of Research Cohort and Real-world Data Catalogues Enabling Multi-center Studies. Yearb Med Inform. 2022 Aug;31(1):262–72.

(Thurin et al, 2022) Thurin NH, Pajouheshnia R, Roberto G, Dodd C, Hyeraci G, Bartolini C, et al. From Inception to ConcePTION: Genesis of a Network to Support Better Monitoring and Communication of Medication Safety During Pregnancy and Breastfeeding. Clin Pharmacol Ther. 2022 Jan;111(1):321–31.

Chapter 2. Pregnancy algorithm and Mother-children linkage

2.1 Introduction

The ConcePTION Pregnancy Algorithm (PA) is a meta-algorithm that aims at identifying list of pregnancies experienced by the instance population in the most comprehensive manner, along with their start and end dates from diverse European data sources. The algorithm is stored in a publicly accessible GitHub repository (ConcePTIONAlgorithmPregnancies).

As detailed in Section 1 of the present Deliverable, the participating data sources differ in terms of available data that are pertinent to the purpose of identifying pregnancies, for instance, birth register, congenital anomalies register, hospital admission and discharge records and primary care medical records. Some sources are more informative and accurate regarding pregnancy-related data, for example, the birth registries are specifically designed to collect information on such events. However, birth registries do not comprehensively reflect all the pregnancy episodes in the data source, as they only registered pregnancies when completed (*Campbell et al. 2022, Bertoia et al. 2022, Margulis 2022, Nordeng et al. 2024*).

This aspect represented somehow the starting point of the ConcePTION PA development, as researchers initially focused on Italian studies on underestimation of maternal mortality (*Donati S et al., 2011, Donati S et al., 2018*). According to such studies, use of a single data bank could lead to massive underestimation of this important indicator. Other sources of inspiration were studies conducted by several groups in Europe (*Shink T et al, 2020*) that had highlighted how data on pregnancy end could be partial or even inconsistent. The algorithm of Matcho (*Matcho et al., 2018*) provided the base for code lists that were then expanded with the support of multiple research groups within and outside of the ConcePTION consortium.

Several algorithms exist that can identify pregnancies in health data sources. The ConcePTION PA shares some of its characteristics with some of them, in terms of expanding the inclusion of pregnancies beyond those marked as complete (*Bertoia et al. 2022, Chomistek et al 2023, Nordeng et al. 2024*) and in terms of including more than one data source (*Charlton et al. 2014, Matcho et al. 2018, Cohen et al. 2020*). However, it also introduces several novel aspects, that are described in the Discussion session of the present Deliverable.

In the present Deliverable we first describe the general strategy of the algorithm and then we detail the specific ingredients that are used in each step of the algorithm.

2.2 Structure of the ConcePTION PA

In section 2.2 the structure of the ConcePTION PA is presented in an overview, and all the components are then extensively described in section **2.3 Ingredients of the algorithm**.

The purpose of the ConcePTION PA is to identify both ongoing and completed pregnancies and estimate the start date (last menstrual period), end date and type of pregnancy end. It can be used in different data sources, that may have different data provenance. Based on the type of available information and the provenance, a quality indicator is created for each identified pregnancy episode. The conceptual design is that any record indicating a pregnancy is retrieved from available data banks in the data sources. The algorithm labels records with tentative information on when that pregnancy started, when it ended, and which type of end that pregnancy had (see **Table 2.1** for the classification of types of pregnancy ending).

Table 2.1. Type of pregnancy end assigned by the algorithm

Type of end	Description	Details
LB	Live birth	the pregnancy ended in a live birth
BUNSP	Birth unspecified	the pregnancy ended in a delivery with unspecified outcome of the baby (including live birth), after gestational week 22
UNSP	End Unspecified	the pregnancy ended at the record date, but outcome of the pregnancy is unspecified (including live birth) and gestational age at the record date is unspecified (before or after gestational week 22)
SB	Stillbirth	baby loss before or during the delivery, after gestational week 22 or week 24 in the UK
SA	Spontaneous abortion	pregnancy loss before 22 weeks' gestation (24 weeks in UK)
T	Elective termination	legal termination of pregnancy /medical abortion
ECT-MOL	Ectopic or molar pregnancy	the fertilized egg implants outside the uterus or there is evidence of abnormal product of conception
ONGOING	Pregnancy ongoing	the estimated date of end of pregnancy is after the date on which data are extracted
UNK	Unknown	the imputed or observed date of end of pregnancy is before the cutoff date of the data, therefore the pregnancy has surely ended, but the type of end could not be established
UNF	Unfavorable Unspecified	pregnancy with observed end date, but outcome unspecified, except live birth
LOSTFU	Lost to follow-up	the estimated date of end of pregnancy is after the end of the observation period (i.e. a continuous period of inclusion in the underlying population of the data source) of the pregnant person e.g. the woman leaves the country.

Retrieval of records

All records that imply that a pregnancy is observed on the date of the record are retrieved from multiple data streams:

- CONCEPTSET (see section **2.3.2 CONCEPTSET**): this is the stream that retrieves records with a diagnostic code or a procedure code implying that the person is experiencing an ongoing pregnancy or an end of pregnancy, such as a diagnosis of preeclampsia or of spontaneous onset of labour, or a procedure of amniocentesis or of a Caesarean section

- ITEMSET (see section **2.3.3 ITEMSET**): this is a manner to retrieve records carrying other, non-diagnostic coded observations collected during routine healthcare data, , such as the recording of a positive results from a pregnancy test; this is also used to retrieve records that do not imply pregnancy at the record date, but that help assessing characteristics of a pregnancy that has been retrieved by other records (e.g., last menstrual period)

- PROMPTSET (see section **2.3.4 PROMPTSET**): this is a manner to retrieve records of birth registries, terminations registries, and/or spontaneous abortion registries recorded in the CDM table SURVEY_OBSERVATIONS
-
- EUROCAT (see section **2.3.5 EUROCAT**): this retrieves records of a congenital anomaly notification in the EUROCAT table

Quality assignment to all retrieved records

Records are retrieved from multiple streams that may vary significantly in terms of information carried about the pregnancy observed. This information includes the pregnancy start date, pregnancy end date, type of pregnancy end, and gestational age, and can either be found in the record itself or imputed by the algorithm.

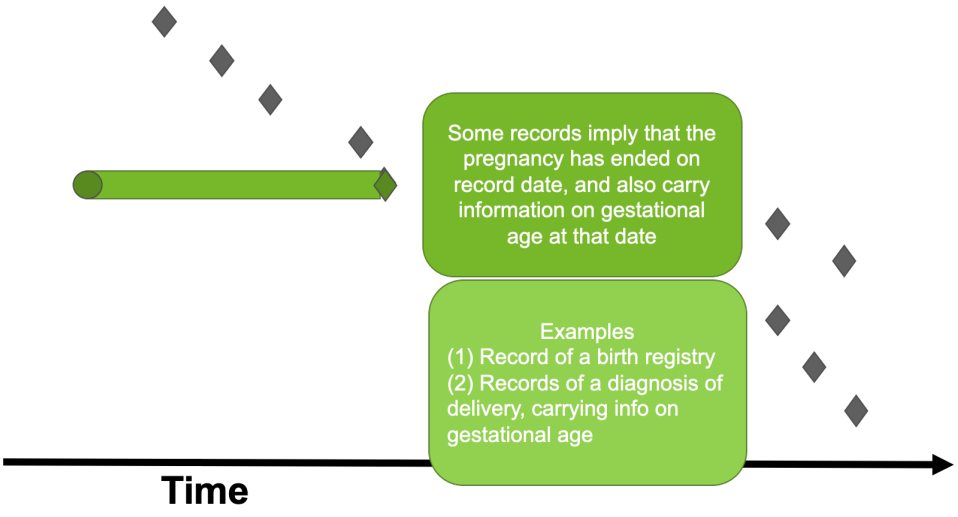
Based on the type and availability of this information, records are assigned a quality color among the followings:

- Green, if both, pregnancy start date and pregnancy end date are recorded (**Figure 2.1**, **Table 2.2**),
- Yellow, if pregnancy end date is recorded as the record date and pregnancy start date is imputed (**Figure 2.1**, **Table 2.2**),
- Blue, if pregnancy start date is recorded and pregnancy end date is imputed (**Figure 2.1**, **Table 2.2**),
- Red, if both, pregnancy start date and pregnancy end date are imputed (**Figure 2.1**, **Table 2.2**).

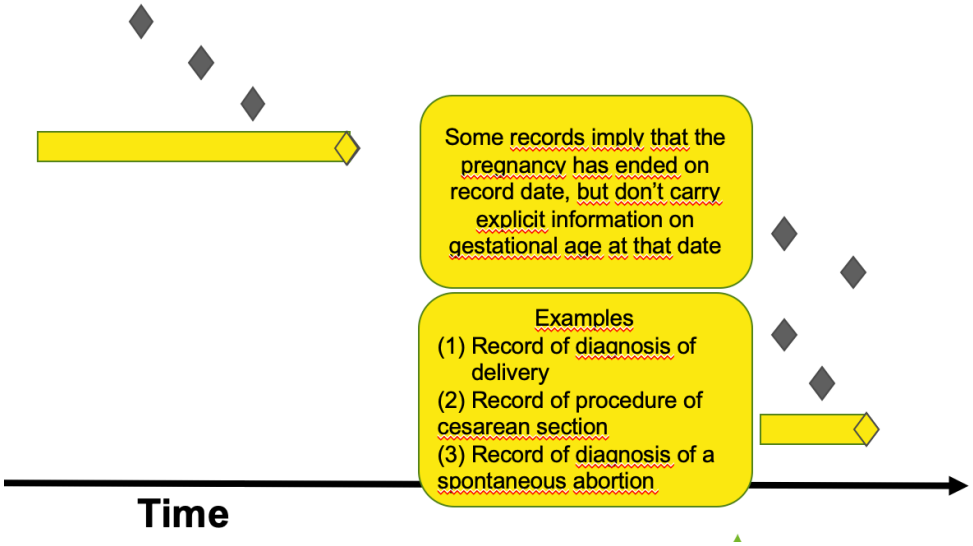
Below is a graphical representation of the quality colors assignment to the different type of records retrieved from the data sources (**Figure 2.1**). The more refined ranking of quality assigned to each record is presented in **Table 2.2**.

Figure 2.1. Quality colors according to the type of pregnancy record retrieved. (Please refer to Table 2.2 for a more refined ranking of quality) In the figure, the diamond represents the date in which the record is recorded, circle represents a recorded start date of pregnancy and length of the bar represents the record wise estimation of the duration of the pregnancy

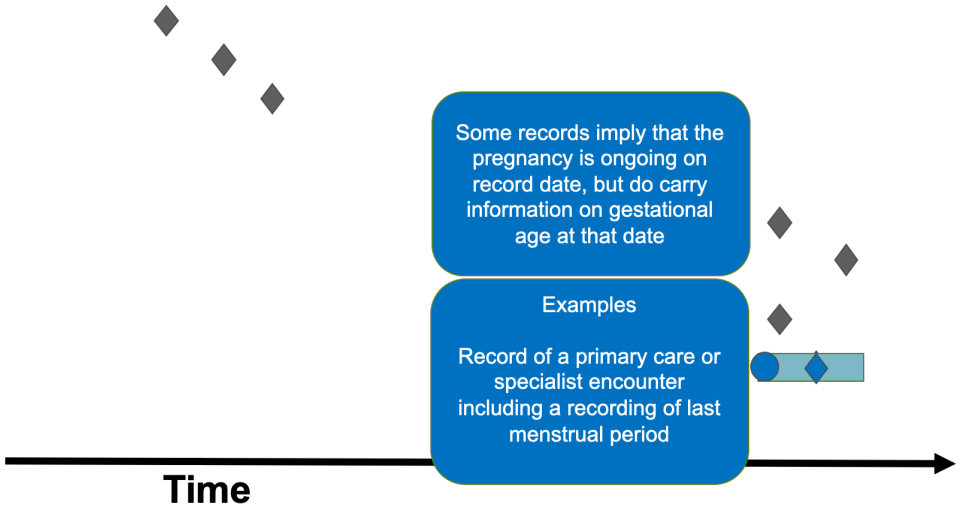
Panel A. Green quality



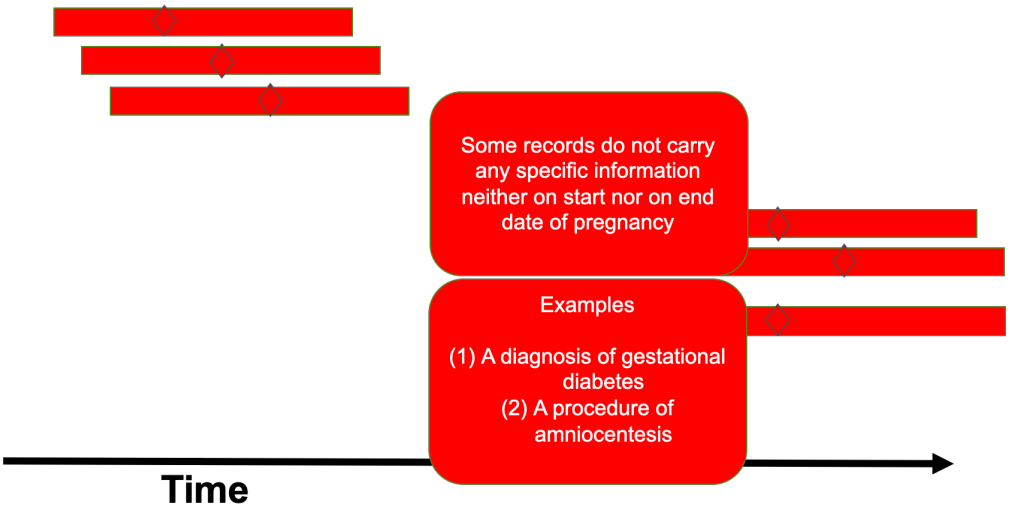
Panel B. Yellow quality



Panel C. Blue quality



Panel D. Red quality



Beyond the color code, other combinations of information included in the records can be identified and contribute to a more refined ranking of quality. One of this information is the type of setting originating the record, for instance, primary and non-primary care. In the majority of data sources, records from primary care setting are assigned a lower quality as compared to non-primary care, as primary care is not the designated setting for pregnancy care registration: for this reason, the assigned ranking differs among them.

The ranking of quality assigned to all the retrieved records ranges from 1 (i.e., the highest quality, having quality color Green from stream EUROCAT) to 99 (Table 2.2).

Table 2.2. Ranking of quality based on provenance of data and type of information

Quality Ranking	Quality Colour	Stream	Specification
1	Green	EUROCAT	Both, pregnancy start date and pregnancy end date are recorded
2	Green	PROMPT	Both, pregnancy start date and pregnancy end date are recorded
3	Green	ITEMSETS	Both, pregnancy start date and pregnancy end date are recorded
4	Green	CONCEPTSETS diagnosis codes	Both, pregnancy start date and pregnancy end date are recorded
5	Yellow	EUROCAT	pregnancy completed and pregnancy start date not available and imputed
6	Yellow	PROMPT	pregnancy completed and pregnancy start date not available and imputed
7	Yellow	ITEMSETS	pregnancy completed and pregnancy start date not available and imputed
8	Yellow	CONCEPTSETS diagnosis codes	pre-term and at term delivery with live birth, pregnancy start date not available and imputed (primary care excluded)
9	Yellow	CONCEPTSETS diagnosis codes	Delivery with live birth, pregnancy start date not available and imputed, (primary care excluded)
10	Yellow	CONCEPTSETS procedure codes	Delivery with live birth, pregnancy start date not available and imputed, (primary care excluded)
11	Yellow	CONCEPTSETS diagnosis codes	stillbirth, pregnancy start date not available and imputed, (primary care excluded)
12	Yellow	CONCEPTSETS diagnosis codes	At term, post-term, pre-term birth outcome unspecified, pregnancy start date not available and imputed, (primary care excluded)
13	Yellow	CONCEPTSETS diagnosis codes	Childbirth with birth outcome unspecified, pregnancy start date not available and imputed, (primary care excluded)
14	Yellow	CONCEPTSETS procedure codes	Delivery, pregnancy start date not available and imputed, (primary care excluded)
15	Yellow	CONCEPTSETS diagnosis codes	Elective termination narrow, pregnancy start date not available and imputed, (primary care excluded)
16	Yellow	CONCEPTSETS	Elective termination and medicated voluntary termination of pregnancy,

		procedure codes	pregnancy start date not available and imputed, (primary care excluded)
17	Yellow	CONCEPTSETS diagnosis codes	Spontaneous abortion narrow, pregnancy start date not available and imputed, (primary care excluded)
18	Yellow	CONCEPTSETS procedure codes	spontaneous abortion, pregnancy start date not available and imputed, (primary care excluded)
19	Yellow	CONCEPTSETS diagnosis codes	Ectopic pregnancy, pregnancy start date not available and imputed, (primary care excluded)
20	Yellow	CONCEPTSETS procedure codes	Ectopic pregnancy, pregnancy start date not available and imputed, (primary care excluded)
21	Yellow	CONCEPTSETS diagnosis codes	Stillbirth possible, elective termination possible, spontaneous abortion possible, pregnancy start date not available and imputed, (primary care excluded)
22	Yellow	COCEPTSETS procedure codes	Procedures of end of pregnancy with unfavorable unspecified outcome, pregnancy start date not available and imputed, (primary care excluded)
23	Yellow	CONCEPTSETS diagnosis codes	Delivery with birth outcome unknown, pregnancy start date not available and imputed, (primary care excluded)
24	Yellow	CONCEPTSETS procedure codes	Procedures of end of pregnancy with outcome unknown, pregnancy start date not available and imputed, (primary care excluded)
25	Yellow	CONCEPTSETS diagnosis codes	Birth possible, pregnancy start date not available and imputed, (primary care excluded)
30	Yellow	CONCEPTSETS diagnosis codes	primary care records, pregnancy start date not available and imputed, end date estimated with record date
40	Blue	all Streams	ongoing pregnancy and pregnancy start date recorded
50	Red	all Streams	ongoing pregnancy having pregnancy start date not available and imputed
99	-	ITEMSET	Record with information about pregnancies, but not necessarily implying pregnancies

Record sorting

Within each person the records are sorted first per ranking of quality (**Table 2.2**) and then based on record_date, from most recent to oldest. The data source may require a different set of hierarchy rules (see **Table 2.4** for data sources' specific rules).

Reconciliation

Then, to reconcile pregnancies, the first record, as resulted after sorting, will be compared in turn with all the subsequent records, one at a time. If the time period of the pregnancy of the next record is compatible with the time period of pregnancy defined by the first record, the reconciliation takes place, that is, the variables of the two records (start date, end date, and type of end) are reconciled. Otherwise, the record is labeled as belonging to a second pregnancy episode. See section below **2.3.6 Create Pregnancies (Reconciliation).**

Once all records have been either reconciled or moved to the pregnancy group two, the procedure starts again on the second group, and so on iteratively until all records have been reconciled.

The reconciliation defines three variables for each pregnancy episode: start date of pregnancy, end date of pregnancy and type of pregnancy end. The recordwise information on start, end and type of end is reconciled in a hierarchical manner: information carried by records with higher quality is prioritized over records of lower quality.

Predictive model

Additionally, in data sources that have information on the start of pregnancy (e.g. birth registry), a predictive model is applied to predict the start date of pregnancy. First, record wise imputation is made, then a new start date of pregnancy is imputed using a weighted average of the prediction across records of the pregnancy. See more details in the sections **2.3.7 Predictive model.**

Final refinement

At this stage, if some pregnancies are too long or overlap, a final refinement is enacted, see section **2.3.8 Final refinement.**

Additional step for pregnancy with type of end UNK

As a next step, pregnancies with type of end assigned as UNK undergo an additional revision. Specifically, if the date of end of pregnancy falls outside the observation period of the pregnant person, the type of pregnancy end is updated and defined as LOSTFU. If the end date of pregnancy is after the date on which the data are extracted, the type of pregnancy end is updated and defined as ONGOING.

Output of the algorithm

The output of the algorithm has one record per pregnancy. Each pregnancy is stored with its main variables (start, end and type of end) as well as secondary variables (e.g., description of the records composing the pregnancy). The data model of the final output is presented in **Table 2.3**. Additionally, at the end of the algorithm a sample of 30 pregnancies is extracted from the output for data sources' experts review (**2.3.9 Verification of a sample of pregnancies**).

Table 2.3. Data model of the final output

Variable name	Description	Type	Vocabulary
pregnancy_id	unique identifier of a pregnancy	string	-
person_id	unique identifier of the pregnant person	string	-
age_at_start_of_pregnancy	age at start of pregnancy	int	-
pregnancy_start_date	best estimate of the date of pregnancy start	date	-
pregnancy_end_date	best estimate of the date of pregnancy end	date	-
meaning_start_date	method by which pregnancy_start_date was obtained	string	from_itemsets_ITEMSET_*NAME* from_conceptset_Gestation_*WEEK* imputed_from_*OTHER_STREAM_NAME* updated_from_blue_record
meaning_end_date	method by which pregnancy_end_date was obtained	string	from_conceptset_*CONCEPT_NAME* *REGISTRY_NAME*
type_of_pregnancy_end	Type of pregnancy end	string	LB = livebirth SB = stillbirth SA = spontaneous abortion T = termination ECT-MOL = ectopic or molar pregnancy UKN = unknown UNF = other non-live birth ONGOING = pregnancy was ongoing at the time of CDM instance creation
date_of_principal_record	date when the record of highest quality of the pregnancy was recorded	date	-
meaning_of_principal_record	meaning of the principal record	string	Among others: birth_registry_mother hospitalisation_primary spontaneous_abortion_registry induced_termination_registry emergency_room_diagnosis hospitalisation_secondary
date_of_oldest_record	date of oldest record	date	
date_of_most_recent_record	date of most recent record	date	
imputed_start_of_pregnancy	whether the start of pregnancy was imputed	int	1 = imputed 0 = not imputed
imputed_end_of_pregnancy	whether the end of pregnancy was imputed	int	1 = imputed 0 = not imputed
highest_quality	quality of the highest quality record	string	green yellow

			blue red
number_of_records_in_the_group	number of records in the group	int	-
number_green	number of records in the group of green quality	int	-
number_yellow	number of records in the group of yellow quality	int	-
number_blue	number of records in the group of blue quality	int	-
number_red	number of records in the group of red quality	int	-
PROMPT	whether the pregnancy was included by the PROMPT stream	string	yes' = yes 'no' = no
CONCEPTSET	whether the pregnancy was included by the CONCEPTSET stream	string	yes' = yes 'no' = no
EUROCAT	whether the pregnancy was included by the EUROCAT stream	string	yes' = yes 'no' = no
ITEMSET	whether the pregnancy was included by the ITEMSET stream	string	yes' = yes 'no' = no
algorithm_for_reconciliation	string that explain the reconciliation	string	-
description	string that reports the name of the concept (or the meaning) of all the records that compose the pregnancy	string	-
GGDE	whether the pregnancy is composed by two green records that are discordant on the end of pregnancy	int	1 = GGDE 0 = not GGDE
GGDS	whether the pregnancy is composed by two green records that are discordant on the start of pregnancy	int	1 = GGDS 0 = not GGDS
INSUF_QUALITY	whether the pregnancy is composed by only blue or red records	int	1 = INSUF_QUALITY 0 = not INSUF_QUALITY
gestage_greater_44	whether the gestational age is greater than 44 weeks	int	1 = greater than 44w 0 = not greater than 44w
sex_at_instance_creation	most recent measurement of the sex of the pregnant person	string	M = "male" F = "female"
n_child	number of children linked to the pregnancy, created using PERSON_RELATIONSHIP table	int	-
child_in multiple_pregnancies	Variable used to check whether a child linked to the pregnancy is also linked to another pregnancy	int	1 = at least 1 child linked to multiple preg 0 = no child linked to multiple preg

In some data sources, information that links the identifiers of a person with the identifiers of his/her birth mother is available (**2.3.1 Data sources' specific parameters**). Such data sources store this information in a specific table of the ConcePTION CDM (PERSON_RELATIONSHIP). In the first steps of the PA, this information is retrieved among other records (see the prompt section below) and used to identify pregnancies. Moreover, as an output, the data set storing identifiers of pregnancies alongside identifiers of children is also created. See the section **2.4 Mother-children linkage** below.

2.3 Ingredients of the algorithm

2.3.1 Data sources' specific parameters

The PA takes advantage from the ConcePTION Common Data Model (CDM) which preserves the granularity of each data sources data (*Thurin et al., 2021*). The CDM ensures that the origin of each record can be retrieved during data processing, since each record contains a 'origin' variable where the name of the origin table (in the original language of the data source) is stored, and can be referred to the data model of the origin table, and to the rules for its ETL in the ConcePTION CDM, as stored in the ConcePTION Catalogue (https://vac4eu.molgeniscloud.org/conception/catalogue/#/).

Moreover, each record composing a pregnancy contains a variable named 'meaning' that stores a summary description of the provenance of the data used to generate the pregnancy record (e.g., 'hospitalization primary diagnosis', or 'birth registry'). This allows to carry over to the pregnancy list produced at the end of the algorithm the information on data diversity that originated the pregnancy, therefore allowing the upfront selection of the characteristics of the pregnancies needed for each research questions.

The PA can be tailored according to the characteristics of the data source and needs of the data partner by setting specific parameters at different steps of the script, as described below:

datasource_that_does_not_modify_PROMPT	In the reconciliation process, the pregnancy start date is defined using the contribution from all available records. The data sources listed here use only the information provided by prompts, when available
datasource_with_conceptsets	the data sources in this have diagnosis or procedure codes that can be used to retrieve pregnancies
datasource_with_itemsets_stream_from_medical_obs	the data sources present in this list have records of MEDICAL_OBSERVATIONS that can be used to detect pregnacies (itemsets)
datasource_with_person_rel_table	the data sources in this have the PERSON_RELATIONSHIP table that can be used to define pregnancies by mother-child relationship
datasource_with_procedures	the data sources in this list have records of PROCEDURES that can be used to detect pregnancies (conceptset)
datasource_with_promt	the data sources in this list have records of SURVEY_ID that can be used to detect pregnancies (prompt)
datasource_with_promt_child	the data sources in this list have records in SURVEY_ID/SURVEY_OBSERVATION that are

	related to the child (instead of the mother)
datasource_with_related_id_correspondig_to_child	the data sources in this list have records in SURVEY_ID/SURVEY_OBSERVATION that are related to the child, and in PERSON_RELATIONSHIP the person_id is related to the mother, and the related_id corresponds to the child
datasource_with_visit_occurrence_prompt	the data sources in this list have records of VISIT_OCCURENCE that can be used to detect pregnancies (prompt)
datasources_EUROCAT	the data sources in this list have the EUROCAT table
datasources_prescription	the data sources in this list use prescription
datasources_that_do_not_use_prediction_on_red	the data sources in this list do not use predictive model to impute the pregnancies start date for yellow and red records
datasources_that_end_red_pregnancies	the data sources in this list consider the date of the most recent record as the pregnancy end date for red pregnancies
datasources_with_specific_algorithms	the data sources in this list have a specific algorithm to impute pregnancy information
datasources_with_subpopulations	the data sources in this list have subpopulations
Maxgap	indicates the period after (or before) a pregnancy in which pregnancy are implausible, it is set at 28 days
maxgap_specific_meanings	indicates the period after (or before) a pregnancy in which pregnancy are implausible, for a specific list of record meaning
list_of_meanings_with_specific_maxgap	specific list of record meaning for which the period after (or before) a pregnancy in which other pregnancy are implausible is different from “maxgap”
gap_allowed_red_record	indicates the maximum time that can elapse between pregnancy records of the same pregnancy that do not contain start or end information
max_gestage_yellow_no_LB	maximum gestational age not-LB pregnancies

Table 2.4. DAPs Data sources' specific parameters

Parameters	Data Source																			
	UOSL	VID	SNDS	BIFAP	CASERTA	GePaRD	EpiChron	HSD	SAIL Databank	PHARMO	CPRD	SIDIAP	DANREG	KI	ARS	FERR	EFEMERIS	POMME	THL	RDRU FISABIO
datasource_that_does_not_modify_PROMPT	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
datasource_with_conceptsets	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
datasource_with_itemsets_stream_from_medical_obs	0	1	0	1	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0
datasource_with_person_rel_table	1	1	0	0	0	0	0	0	1	0	0	0	1	0	1	1	1	1	1	1
datasource_with_procedures	0	1	1	1	1	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0
datasource_with_prompt	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1
datasource_with_prompt_child	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	1	1
datasource_with_related_id_correspondig_to_child	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
datasource_with_visit_occurrence_prompt	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
datasources_EUROCAT	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
datasources_prescriptions	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0
datasources_that_do_not_use_prediction_on_red	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	1	1	1
datasources_that_end_red_pregnancies	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
datasources_with_specific_algorithms	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
datasources_with_subpopulations	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Maxgap	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28	28
maxgap_specific_meanings	168																			
list_of_meanings_with_specific_maxgap	primary_care_diagnosis - primary_care																			
gap_allowed_red_record	56	56	56	56	56	56	56	180	56	56	56	56	56	56	56	56	56	56	56	56
max_gestage_yellow_no_LB	84																			

2.3.2 CONCEPTSET

The stream CONCEPTSET takes advantage from the concept sets - which are a subset of code lists - to retrieve records that carry a diagnostic code or a procedure code implying that the person is experiencing an ongoing pregnancy or an end of pregnancy, such as a diagnosis of preeclampsia or of spontaneous onset of labor, or a procedure of amniocentesis or of a cesarean section. In this stream we query the CDM tables EVENTS, MEDICAL_OBSERVATIONS, SURVEY_OBSERVATIONS for diagnostic codes, and PROCEDURES for procedure codes.

Procedure codes records typically lack the gestational age at record date and can be either yellow (e.g., procedure of cesarean section) or red quality (e.g., procedure of amniocentesis). In contrast, diagnosis codes records implying a delivery can also have information on the gestational age (green quality). Diagnosis codes records can also refer to a delivery or a spontaneous abortion without any other information (yellow quality), or to gestational diabetes (red quality).

Diagnostic codes
The list of diagnostic codes used to retrieve records implying that a person is experiencing an ongoing pregnancy, or an end of pregnancy was initially sourced from literature (*Matcho et al. 2018*). The medical concepts retrieved from the literature were mapped to the coding systems of the data sources (SNOMEDCT_US, SCTSPA, ICD9CM, ICD10, ICD10CM, READ, ICPC2P, ICPC, MTHICD9) using the tool Codemapper (*Becker et al, 2017*) which is based on the Unified Medical Language System (UMLS Terminology Service), and were further refined by data partners. Finally, the diagnosis codes populated a set of code lists that are named according to the corresponding medical concepts. The full set of lists of reviewed diagnostic codes is publicly available (*Girardi et al, 2024*).

Notably, the personal identifier of the records carrying such diagnosis codes are normally interpreted as the identifiers of the pregnant persons. However, we also included an additional set of concepts including diagnostic codes whose personal identifier is a newborn. In the PA, such codes are associated to the date of pregnancy end, rather than the record date, and to the identifier of the person indicated as the gestational mother of the child in the PERSON_RELATIONSHIPS table of the ConcePTION CDM (see more details below in the specific section on **2.4 Mother-children linkage**).

All the diagnostic codes included in the PA underwent an initial review from the data partners, according to the local expertise, followed by a review conducted by the leading data partner responsible for developing the PA. The first review aimed to enhance the original code lists to ensure comprehensiveness, whereas the second review focused on the assignment of the appropriate concept set to each included diagnostic code. This last involved tagging each code within its original code list. The criteria applied for tag assignment was agreed with medical experts from the participating data partners (*Girardi et al, 2024*).

The concept set is a subset of code lists. As an example, codes belonging to the lists “Elective Termination” are tagged as *narrow* when they clearly refer to a diagnosis of elective termination occurred at record date (e.g. Legal termination of pregnancy), populating the concept set “Elective Termination_narrow”, or tagged as *possible* when the association with elective termination outcome

is not straightforward (e.g. Failed medical abortion), populating the concept set "Elective Termination_possible". The two concept sets differ in terms of type of pregnancy end assigned: elective termination and unfavorable, respectively.

Each of the concept set is assigned a type of pregnancy end (see Table 2.1). The mapping from concept sets to codelist and type of end is presented in **Table 2.5.** In Table 2A (in the Annex to this Chapter) a more refined mapping is displayed where each concept set is mapped to the rules to assign start date, end date, corresponding meanings, and quality color and ranking.

Table 2.5. Mapping from concept set to type of pregnancy end_diagnostic codes

Concept set	Code list name	Tag	Type of end	implies pregnancy at the record date	implies end of pregnancy at record date
Gestation_less24_UNK	P_GestationlessthanU_PrA	n.a.	UKN	yes	no
Gestation_24_UNK	P_24weeksUNK_PrA	n.a.	UKN	yes	no
Gestation_25_26_UNK	P_Gestation2526weeksUNK_PrA	n.a.	UKN	yes	no
Gestation_27_28_UNK	P_Gestation2728weeksUNK_PrA	n.a.	UKN	yes	no
Gestation_29_30_UNK	P_Gestation2930weeksUNK_PrA	n.a.	UKN	yes	no
Gestation_31_32_UNK	P_Gestation3132weeksUNK_PrA	n.a.	UKN	yes	no
Gestation_33_34_UNK	P_Gestation3334weeksUNK_PrA	n.a.	UKN	yes	no
Gestation_35_36_UNK	P_Gestation3536weeksUNK_PrA	n.a.	UKN	yes	no
Gestation_more37_UNK	P_Gestation37weeksUNK_PrA	n.a.	UKN	yes	no
Gestation_less24_LB	P_GestationlessthanL_PrA	n.a.	LB	yes	yes
Gestation_24_LB	P_24weeksLB_PrA	n.a.	LB	yes	yes
Gestation_25_26_LB	P_Gestation2526weeksLB_PrA	n.a.	LB	yes	yes
Gestation_27_28_LB	P_Gestation2728weeksLB_PrA	n.a.	LB	yes	yes
Gestation_29_30_LB	P_Gestation2930weeksLB_PrA	n.a.	LB	yes	yes
Gestation_31_32_LB	P_Gestation3132weeksLB_PrA	n.a.	LB	yes	yes
Gestation_33_34_LB	P_Gestation3334weeksLB_PrA	n.a.	LB	yes	yes
Gestation_35_36_LB	P_Gestation3536weeksLB_PrA	n.a.	LB	yes	yes
Gestation_more37_LB	P_Gestation37weeksLB_PrA	n.a.	LB	yes	yes
Ongoingpregnancy	P_OngoingPregnancy[1-7]_PrA	n.a.	UKN	yes	no
Ongoingpregnancy	P_StartofPregnancy_PrA	n.a.	UKN	yes	no
GESTDIAB	P_GESTIAB_PrA	narrow	UKN	yes	no
GESTDIAB	P_GESTIAB_PrA	possible	UKN	yes	no
FGR	P_FGR_PrA	narrow	UKN	yes	no
FGR	P_FGR_PrA	possible	UKN	yes	no
PREECLAMP	P_PREECLAMP_PrA	narrow	UKN	yes	no

PREECLAMP	P_PREECLAMP_PrA	possible	UKN	yes	no
PREG_BLEEDING	P_BLEEDING_PrA	narrow	UKN	yes	no
PREG_BLEEDING	P_BLEEDING_PrA	possible	UKN	yes	no
Birth_possible	P_BirthPossible[1-3]_PrA	n.a.	UKN	yes	no
BirthNarrow_LB	P_BirthNarrowLB_PrA	n.a.	LB	yes	yes
BirthNarrow_BUNSP	P_BirthNarrowBUNSP_PrA	n.a.	BUNSP	yes	yes
EndUnspecified	P_EndUnspecified_PrA	n.a.	UNSP	yes	yes
Atterm_LB	P_AtTermLB_PrA	n.a.	LB	yes	yes
Atterm_BUNSP	P_AtTermBUNSP_PrA	n.a.	BUNSP	yes	yes
Postterm_BUNSP	P_PostTermBUNSP_PrA	n.a.	BUNSP	yes	yes
Preterm_LB	P_PretermLB_PrA	n.a.	LB	yes	yes
Preterm_BUNSP	P_PretermBUNSP_PrA	n.a.	BUNSP	yes	yes
Stillbirth_narrow	P_Stillbirth_PrA	narrow	SB	yes	yes
Stillbirth_possible	P_Stillbirth_PrA	possible	UNF	yes	yes
Interruption_narrow	P_ELECTTERM_PrA	narrow	T	yes	yes
Interruption_possible	P_ELECTTERM_PrA	possible	UNF	yes	no
Spontaneousabortion_narrow	P_SpontaneousAbortion_PrA	narrow	SA	yes	yes
Spontaneousabortion_possible	P_SpontaneousAbortion_PrA	possible	UNF	yes	no
Ectopicpregnancy	P_EctopicPregnancy_PrA	narrow	ECT-MOL	yes	yes
Molarpregnancy	P_MolarPregnancy_PrA	n.a.	ECT-MOL	yes	yes
Gestation_less24_CHILD*	P_Gestationlessthan24weeksC_PrA	n.a.	LB	na	na
Gestation_24_CHILD*	P_24weeksCHILD_PrA	n.a.	LB	na	na
Gestation_25_26_CHILD*	P_Gestation2526weeksCHILD_PrA	n.a.	LB	na	na
Gestation_27_28_CHILD*	P_Gestation2728weeksCHILD_PrA	n.a.	LB	na	na
Gestation_29_30_CHILD*	P_Gestation2930weeksCHILD_PrA	n.a.	LB	na	na
Gestation_31_32_CHILD*	P_Gestation3132weeksCHILD_PrA	n.a.	LB	na	na
Gestation_33_34_CHILD*	P_Gestation3334weeksCHILD_PrA	n.a.	LB	na	na
Gestation_35_36_CHILD*	P_Gestation3536weeksCHILD_PrA	n.a.	LB	na	na
Gestation_more37_CHILD*	P_Gestation37weeksCHILD_PrA	n.a.	LB	na	na

* The codes included in this code list have to be associated to the date of birth (i.e. the date on which the pregnancy ended), rather than the record date, and to the related id linked to the child in PERSON_RELATIONSHIPS

The PA leverages both the record date of the diagnosis code and the medical information carried (i.e. the type of assigned concept set) to define the three variables associated to a pregnancy episode.

The PA underwent several implementations since its first release, including parameter settings for

data sources, debugging, and implemented code lists after the revision's rounds described at the beginning of this section. The latest available version of the script is 5.2.7 (ConcePTIONAlgorithmPregnancies). The upcoming version 5.3 of the PA addressed a limitation of the code lists associated to the type of pregnancy end "Live Birth" (LB) and aimed to overcome one of the limitations of the previous version of the PA. As a matter of fact, in the script version 5.2.7, diagnosis codes associated to LB type of end referred to maternal conditions that occur at the time of delivery and imply that the pregnancy has already ended or will end shortly. However, not all these codes carry the information of a live born baby, and this is the reason why we recommend considering the output of the PA v5.2.7 at face value regarding the LB type of pregnancy end, whereas valid for the start and end date values.

We succeeded in re-classifying diagnosis codes included in the code lists "BirthNarrow", "AtTerm", "PreTerm", "Postterm". Specifically, codes indicating that the pregnancy ended with a live birth are now included in the new code lists "_LB"; codes suggesting a delivery with unspecified outcome for the baby (e.g. Cesarean section) are now included in the new code lists "_BUNSP", and codes with unspecified type of end, and for which gestational age at the end of pregnancy is not specified, are now included in the new code list "EndUnspecified".

The PA version 5.3 has two new types of end of pregnancy. The type of end LB will exclusively pertain codes from the code lists "_LB", while the new type of end BUNSP is assigned to codes from "_BUNSP" and new type of end UNSP to those from "EndUnspecified" (**Table 2.1**).

Procedure codes

The list of procedure codes has been refined by data partners according to their expertise and finally categorized according to the defined concept sets (**Table 2.6**). Procedure codes included in the PA can be consulted in the GitHub page.

Table 2.6. Mapping from concept set to type of pregnancy end_procedure codes

Concept set	Specifications	Type of end
procedure_end_UNK	procedures carrying info on a delivery (e.g., cesarean section)	UNK
procedures_termination	procedures carrying info on an elective termination (e.g., Aspiration curettage of uterus for termination of pregnancy)	T
procedures_spontaneous_abortion	procedures carrying info on spontaneous abortion (currently empty)	SA
procedures_ectopic	procedures carrying info on ectopic pregnancy (e.g., Salpingectomy with removal of tubal pregnancy)	ECT-MOL
procedures_ongoing	procedures carrying info on ongoing pregnancy	UNK
fetal_nuchal_transucency*	procedures of fetal nuchal	UNK
Chorionic_villus_sampling*	procedures of chorionic villus sampling	UNK
Amniocentesis*	procedures of amniocentesis	UNK
Others*	other procedures during pregnancy	UNK

*Information provided by each data partner according to data source specific coding system

2.3.3 ITEMSET

ITEMSET: this is the method to retrieve records that are collected from surveys, or from other complex data banks, such as birth registries, and contain information about pregnancies. An example of an ITEMSET is a record related to surveys conducted at the end of pregnancy, such as those related to birth registries, that contain information about gestational age or type of pregnancy end. Another example of an ITEMSET is a record collected by the general practitioner during a primary care visit where the date of the last menstruation is reported and can be used to define the start of pregnancy.

The information of the ITEMSET category is stored by the DAPs in the SURVEY_OBSERVATIONS and MEDICAL_OBSERVATIONS tables of the ConcePTION CDM, using the data source-specific names of related variable.

Not all ITEMSET implies that the person is pregnant at the time of recording. Some data sources, have information available about the date of the last menstrual period recorded by the primary care physician, which however does not imply that the subject is pregnant, as it can be recorded during a routine visit.

ITEMSET related to surveys conducted at the end of pregnancy are retrieved from the SURVEY_OBSERVATIONS table and are grouped into the following categories:

- GESTAGE_FROM_LMP_WEEKS: gestational age defined by the date of the last menstrual period, in weeks
- GESTAGE_FROM_LMP_DAYS: gestational age defined by the date of the last menstrual period, in days
- GESTAGE_FROM_USOUNDS_WEEKS: gestational age obtained through ultrasound, in weeks
- GESTAGE_FROM_USOUNDS_DAYS: gestational age obtained through ultrasound, in days
- GESTAGE_FROM_DAPs_CRITERIA_WEEKS: gestational age obtained through different methods, in weeks
- GESTAGE_FROM_DAPs_CRITERIA_DAYS: gestational age obtained through different methods, in days
- DATESTARTPREGNANCY: start date of pregnancy
- DATEENDPREGNANCY: end date of pregnancy
- END_LIVEBIRTH: end date of pregnancy for a live birth
- END_STILLBIRTH: end date of pregnancy for a stillbirth
- END_INTERRUPTION: end date of pregnancy for a termination
- END_ABORTION: end date of pregnancy for a spontaneous abortion
- TYPE: type of pregnancy end

Other ITEMSET are instead retrieved from the MEDICAL_OBSERVATION table and are grouped into the following categories:

- LastMenstrualPeriod: date of the last menstrual period
- LastMenstrualPeriodImplyingPregnancy: date of the last menstrual period of a pregnant

person
- GestationalAge: gestational age
- PregnancyTest: date when a pregnancy test was conducted with a positive result

In the upcoming version 5.3 of the algorithm, ITEMSET records produced by birth registries or surveys generated at the end of pregnancy are retrieved from the following data sources: "ARS", "PHARMO", "UOSL", "VID", "CPRD", "GePaRD", "EpiChron", "SIDIAP", "SAIL Databank", "EFEMERIS", "POMME", "DANREG", "KI", "THL", "FERR", "RDRU_FISABIO", "CASERTA".

The data sources "BIFAP", "VID", "PHARMO", "EpiChron", "HSD" have ITEMSET from the MEDICAL_OBSERVATION table.

2.3.4 PROMPTSET

We define as PROMPTSET those records whose existence itself implies that a person is pregnant at time of registration, for example, a record of a birth registry.

In the ConcePTION CDM, the column “meaning” stores information about provenance of the original record was prompted into existence. A record of a birth registry will be loaded into the SURVEY_ID table with meaning “birth_registry”, and this implies that the woman is pregnant at the time of registration, irrespectively of the content of the record itself.

According to the meaning, the date on which the record is recorded may correspond to the end of pregnancy or to a date when the pregnancy is ongoing.

Often PROMPT records are retrieved from VISIT_OCCURRENCE and SURVEY_ID. This means that there is additional information from the origin records that are stored in other tables of the CDM that can be linked to, through visit_occurrence_id and survey_id respectively, and the information to assign start, end, type of end and quality of the prompt record may be retrieved via this linkage. If so, this is governed by ITEMSET record (see next section).

Finally, in some data sources, there are tables containing information about individuals' relationships. In these cases, the PA actively seeks out mother-child relationships stored in the PERSON_RELATIONSHIP table of the CDM, using the meanings “birth_mother” or “gestational_mother”. If located, the child's ID identified in the table is then used to access the date of birth from the PERSONS table in the CDM, and this is used as pregnancy end date.

In the current version of the algorithm, PROMPT records produced by birth registries or surveys generated at the end of pregnancy are retrieved from the following data sources: "ARS", "PHARMO", "UOSL", "VID", "CPRD", "GePaRD", "EpiChron", "SIDIAP", "SAIL Databank", "EFEMERIS", "POMME", "DANREG", "KI", "THL", "FERR", "RDRU_FISABIO", "CASERTA". In addition to records related to the birth registry, "ARS" has also records generated during pregnancies in a community primary care center. PROMPT generated by mother-child relationships are included in the datasources: "EFEMERIS", "POMME", "THL", "ARS", "FERR", "UOSL", "SAIL Databank", "RDRU_FISABIO", "DANREG", "VID".

2.3.5 EUROCAT

EUROCAT corresponds to European network of population-based registries for the epidemiological surveillance of congenital anomalies. The data sources "SAIL Databank" and "VID" include this databank.

Information about pregnancies present in EUROCAT are incorporated in the pregnancy algorithm.

2.3.6 Create pregnancies (Reconciliation)

Within each person the records are sorted first per order_quality (data sources may require different hierarchy rules) and then in reverse order of record_date, from the most recent to the oldest record (Figure 2.2). Then, to reconcile pregnancies, the first record will be compared with all the subsequent records, one at a time: if the time period of the next record is plausible with the time period of pregnancy defined by the first record, the reconciliation takes place, otherwise the record is declared not belonging to the pregnancy group and is moved to a second pregnancy group.

Once all records have been either reconciled or moved to the pregnancy group two, the procedure starts again on the second group, and so on iteratively until all records have been reconciled (see Box 2.1).

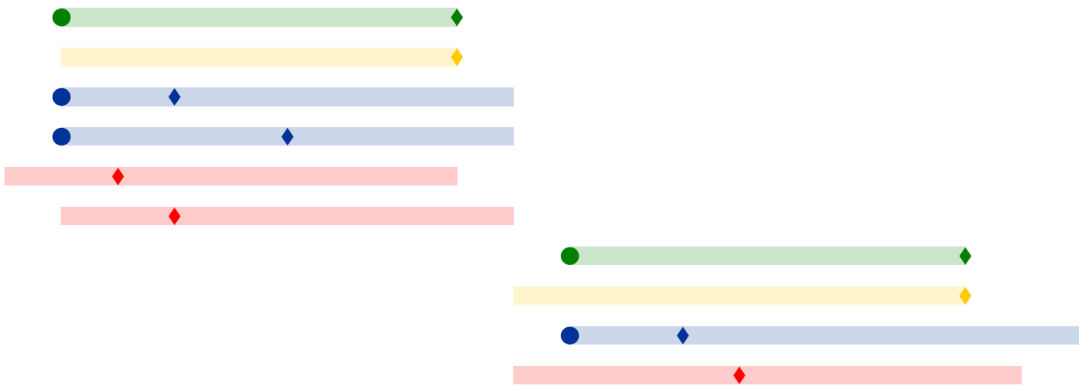


Figure 2.2. Graphical representation of two groups of records. In the graphical representation, a diamond represents the date of the record, a circle represents a record of a date of start of pregnancy, and the bar represents the interval between start and end. When the diamond is not at the end of the bar, it means that the record was recorded before the end of the pregnancy, which implies that the end of the pregnancy was imputed. When there is no circle, the start of pregnancy is imputed. The color of the interval represents the quality of the records, as indicated in step A) above.

Each time a comparison is made, a string describing the reconciliation result will be added to the variable "algorithm_for_reconciliation". In the meantime, it checks whether the pregnancy information needs to be updated. The first review concerns the end of pregnancy type. If the type of the first record is different from the comparison record, and the type of the comparison record is not “UNK”, the string "DiffType" will be pasted to the variable “algorithm_for_reconciliation”, followed by the color of the first record and the color of the comparison record (e.g. “TypeDiff:green/yellow”). Therefore meaning, start dates and end dates of pregnancy will be reviewed. If DAP did not require different hierarchical rules, nine possible comparisons exist:

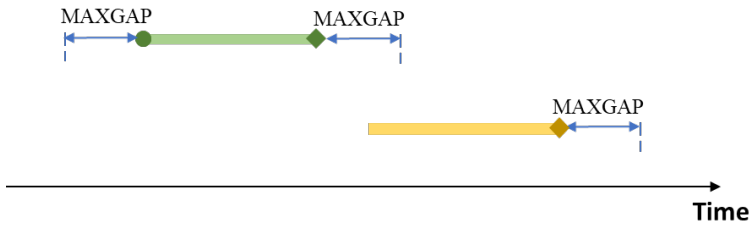
1. Green / Green
2. Green / Yellow
3. Green / Blue

4. Green / Red
5. Yellow / Yellow
6. Yellow / Blue
7. Yellow / Red
8. Blue / Blue
9. Blue / Red

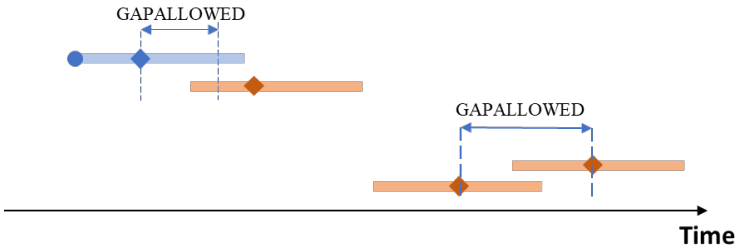
Rules for reconciliation are described in the following Box 2.1.

Parameters:

- **MAXGAP**: indicates the period after (or before) a pregnancy in which pregnancy are implausible, it is set at 28 days;



- **GAPALLOWED**: indicates the maximum time that can elapse between pregnancy records of the same pregnancy that do not contain start or end information, set according to DAPs.



Regardless of the type, two records are assigned to different pregnancies if at least one of those condition is satisfied:

1) Abs(Record date – record date next record) > 280

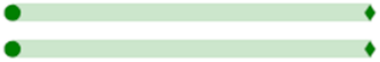
2) end date < start date next record

3) start date > end date next record

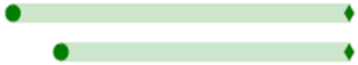
Reconciling Green with Green:

Thus, among records belonging to the same pregnancy:

a) if start dates and end dates are concordant, algorithm_for_reconciliation = "GG:concordant_"



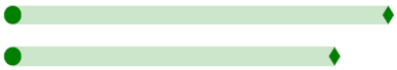
b) if they have different start dates:



i. if the two dates are less than 7 days apart, algorithm_for_reconciliation = "GG:SlightlyDiscordantStart_"

ii. if the difference between the two dates is larger, algorithm_for_reconciliation = "GG: DiscordantStart_"

c) if they have different end dates



i. if the two dates are less than 7 days apart, algorithm_for_reconciliation = "GG:SlightlyDiscordantEnd_"

ii. if the difference between the two dates is larger, algorithm_for_reconciliation = "GG: DiscordantEnd_"

If two different pregnancies overlap:

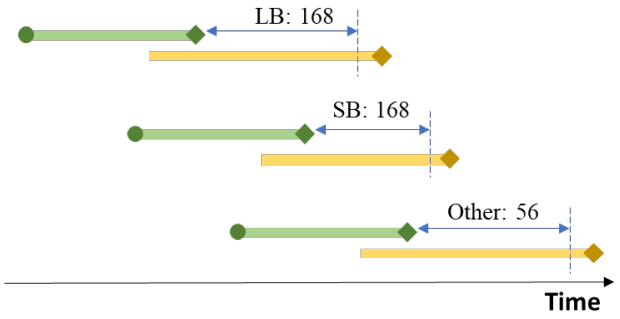
- When a non-LB record is compared with a LB record, LB pregnancies is selected
- Otherwise most recent pregnancy is selected

Overlapping pregnancies are flagged as green discordant (GGD = 1)

Reconciling Green with Yellow:

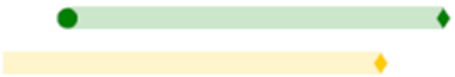
When a green record is compared with a yellow record, the records are assigned to different pregnancies if:

- If LB: End date + 168 < End date next record
- If SB: End date + 168 < End date next record
- If other: End date + 56 < End date next record



if the end dates are concordant, algorithm_for_reconciliation = "GY:concordant_"

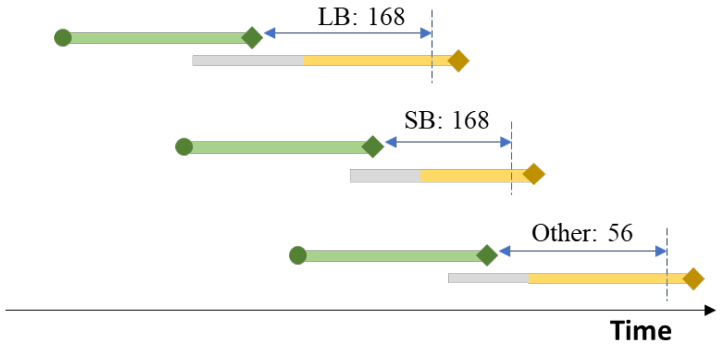
if the inconsistency is only on dates and they are of less than 7 days algorithm_for_reconciliation = "GY:SlightlyDiscordantEnd_"



if the inconsistency is only on dates and they are more than 7 days, algorithm_for_reconciliation = "GY:DiscordantEnd_"

If pregnancies overlap, start of yellow pregnancies is set to:

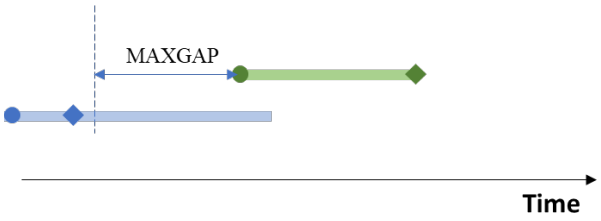
- If LB: End date - 154
- If SB: End date - 154
- If other: End date - 42



Reconciling Green with Blue:

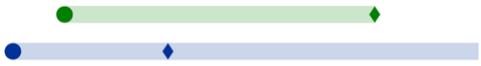
When a green record is compared with a blue record, the records are assigned to different pregnancies if:

start date - MAXGAP > record date next record



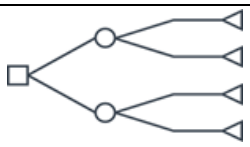
if the start dates are concordant, algorithm_for_reconciliation = "GB:concordant"

if the start dates are disconcordant,

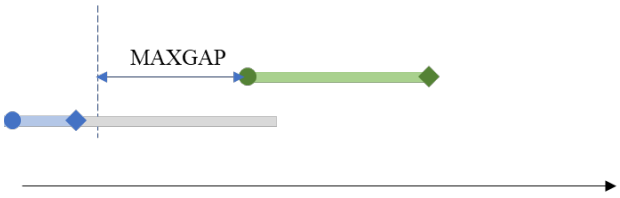


- as a default: update the value of pregnancy_start_date, the meaning_start_date, and set algorithm_for_reconciliation = "GB:StartUpdated"; the rationale is that the start of pregnancy recorded during pregnancy is of higher quality;

- upon indication of the DAP, this rule may vary according to specific characteristics of the quality blue record



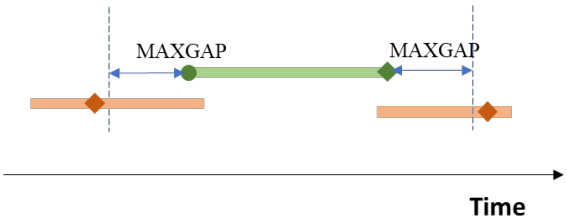
pregnancies overlap, end of blue pregnancy is set at record date



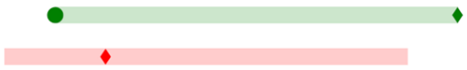
Reconciling Green with red:

When a green record is compared with a red record, the records are assigned to different pregnancies if:

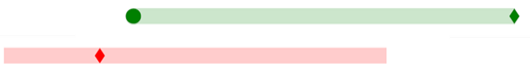
- End date green + MAXGAP < record date next record
- start date green - MAXGAP > record date next record



if 'record date' of the red record is between the start and end of the green record = “GR:NoInconsistecies”

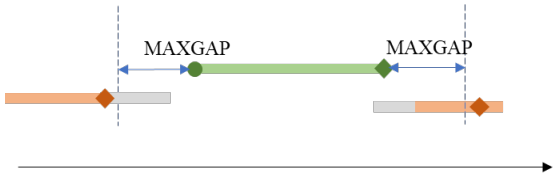


if 'record date' of the red record is not between the start and end of the green record = “GR:Inconsistecies”



pregnancies overlap:

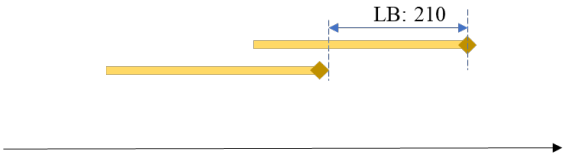
- If red pregnancy overlap on the left, end of red pregnancy is set at most recent record date
- If red pregnancy overlap on the right, start of red pregnancy is set at oldest record date – (MAXGAP/2)



Reconciling yellow with yellow:

When a yellow record is compared with a yellow record, the records are assigned to different pregnancies if:

- If LB: End date - 168 > End date next record
- If SB: End date - 168 > End date next record
- If other: End date - 56 > End date next record



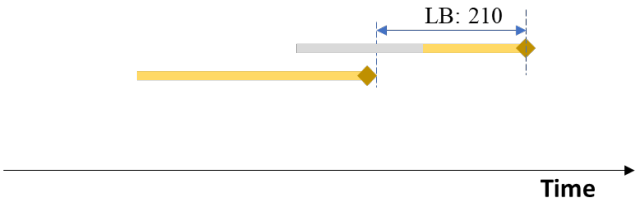
if they have same end date, algorithm_for_reconciliation = "YY:concordant"

if the inconsistency is on dates and they are of less than 7 days apart, algorithm_for_reconciliation = "YY:SlightlyDiscordantEnd_"

if the inconsistency is on dates and they are more than 7 days apart, algorithm_for_reconciliation = "YY:DiscordantEnd_"

pregnancies overlap, start of yellow pregnancies is set to:

- If LB: End date - 154
- If SB: End date - 154
- If other: End date - 42



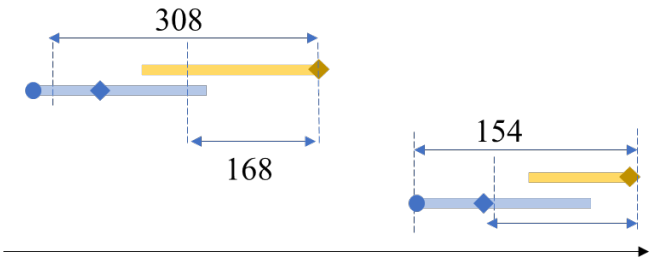
Reconciling yellow with blue:

When a LB/SB yellow record is compared with a blue record, the records are assigned to different pregnancies if:

- End date - 308 > start date next record & end – 168 < record date next record

When a non-LB/SB record is compared with a blue record, the records are assigned to different pregnancies if:

- End date - 154 > start date next record & end – (GAPALLOWED) < record date next record



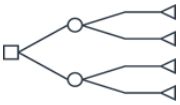
When a yellow record is reconciled with a blue record, the "meaning_start_date" and "imputed_start" variables are updated, and:

if the start dates are concordant, algorithm_for_reconciliation = “YB:concordant”

if the start dates are disconcordant,

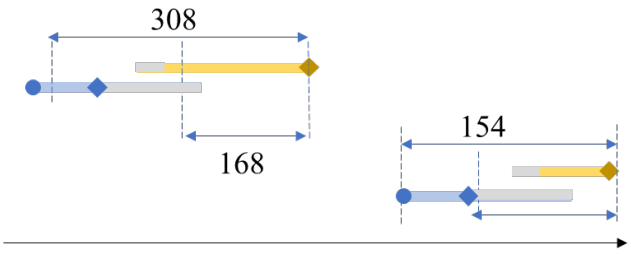
- as a default: update the value of pregnancy_start_date, the meaning_start_date and set algorithm_for_reconciliation = “YB:StartUpdated_”; the rationale is that the start of pregnancy recorded during pregnancy is of higher quality;

- upon indication of the DAP, this rule may vary according to specific characteristics of the quality blue record



If pregnancies overlap:

- start of yellow pregnancy is set at 154/ GAPALLOWED
- End of blue pregnancies is set a record date



Reconciling yellow with red:

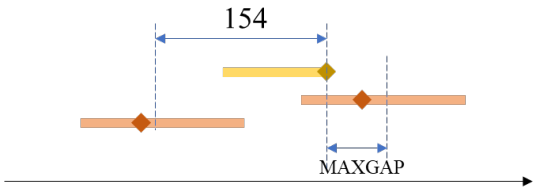
The records are assigned to different pregnancies if:

- end date + MAXGAP < record date next record

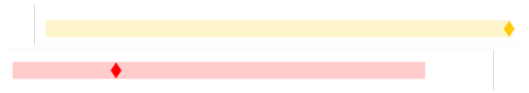
When a yellow non-LB/SB record is compared with a red record, the records are assigned to different

pregnancies if:

- end date – (154) > record date next record



if 'record date' of the red record is between the start and end of the yellow record = "YR:NoInconsistecies_"



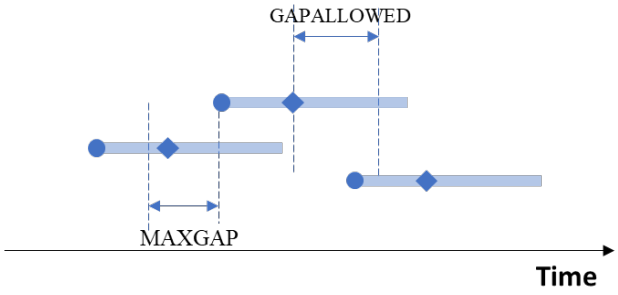
if 'record date' of the red record is not between the start and end of the green record = "YR:Inconsistecies_"

If pregnancies overlap: End of red pregnancy is set at most recent record date / start of red pregnancy is set at oldest record date – 14 days

Reconciling Blue with Blue:

wo records are classified as belonging to different pregnancies if:

- Start – MAXGAP > record date next record
- record date + GAPALLOWED < start date next record

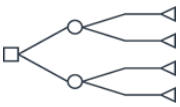


if the start dates are concordant, algorithm_for_reconciliation = "BB:concordant_"

if the start dates are disconcordant,

- as a default: update the value of pregnancy_start_date, the meaning_start_date and set algorithm_for_reconciliation = "BB:StartUpdated_"; the rationale is that the start of pregnancy recorded earlier is of higher quality;

- upon indication of the DAP, this rule may vary according to specific characteristics of the quality blue record

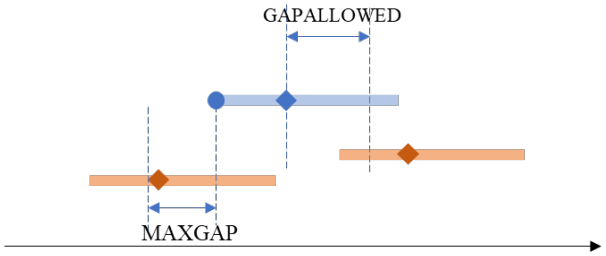


If pregnancies overlap: End of pregnancy is set at most recent record date

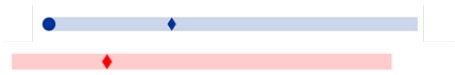
Reconciling Blue with Red:

Two records are classified as belonging to different pregnancies if:

- Start – MAXGAP > record date next record
- record date + GAPALLOWED < record date next record



if 'record date' of the red record is between the start and end of the yellow record = "BR:NoInconsistecies_"



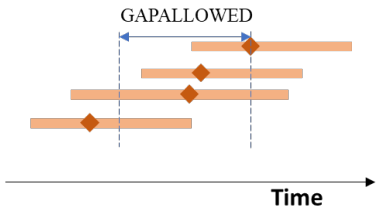
if 'record date' of the red record is not between the start and end of the green record = "BR:Inconsistecies_"

pregnancies overlap: End of pregnancy is set at most recent record date / start of pregnancy is set at end - (MAXGAP – 14)

econciling Red with Red:

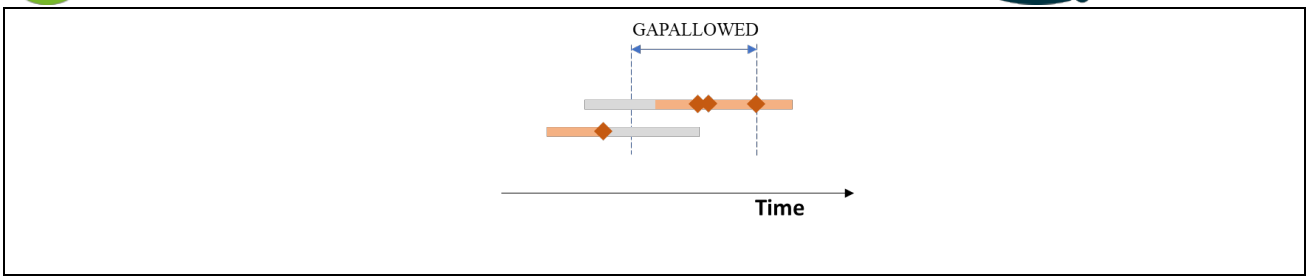
Two records are classified as belonging to different pregnancies if:

- Abs(record date – record date next record) > GAPALLOWED



If pregnancies overlap:

- end of pregnancy is set at most recent record date
- start of red pregnancy is set at oldest record date – (GAPALLOWED/2)



Box 2.1. Algorithm to reconcile records of the same group. In the graphical representation, a diamond represents the date of the record, a circle represents a record of a date of start of pregnancy, and the bar represents the interval between start and end. When the diamond is not at the end of the bar, it means that the record was recorded before the end of the pregnancy, which implies that the end of the pregnancy was imputed. When there is no circle, the start of pregnancy is imputed. The color of the interval represents the quality of the records, as indicated in step A) above.

2.3.7 Predictive Model

Not all records contain information on the start date of pregnancy. Therefore, some pregnancies may have their start date imputed based on selected values a priori (e.g. record date - 280 days for birth_narrow diagnosis codes, record date – 55 days for codes belonging to ongoing concepts, etc.). The predictive model aims to improve imputations for pregnancy start by leveraging groups of records that contain both records with information on the start date and those without. The selected model for prediction is a random forest. A subset of pregnancies is chosen that contains at least one record with information about the start (e.g., pregnancies that have at least one record in the birth registry indicating gestational age). This subset is used to train two different random forest, one for red record prediction and one for yellow record prediction. The covariates for the random forest are as follows:

- Record type: This corresponds to the record description at the finest possible aggregation level. If the record is a diagnosis code, then the "record type" value will match the code itself. When a diagnosis code is not available, the "record type" value will correspond to the meaning of the record.
- Origin: Table from which the record originated.
- Mother's age at the beginning of pregnancy.
- Year in which the record was registered:
  - record_year < 2000 → record_year := 1
  - 2000 ≤ record_year < 2005 → record_year := 2
  - 2005 ≤ record_year < 2010 → record_year := 3
  - 2010 ≤ record_year < 2015 → record_year := 4
  - 2015 ≤ record_year < 2020 → record_year := 5
  - record_year ≥ 2020 → record_year := 6
- Distance from the oldest record belonging to the same pregnancy.

The random forest is implemented in R using the 'ranger' package. The function's parameters will be selected through cross-validation performed individually for each DAP. The tested parameters are as follows:

- Number of trees: 100, 500
- Number of selected variables (mtry): 2, 3, 4
- Always split variable: none, record type

The final model is executed with the parameter combination that minimizes the mean squared error. Once the model is trained, it will predict the start dates of pregnancies for all records not belonging to pregnancies with "highest quality" as yellow or red.
Subsequently, the pregnancy's start date is defined as a weighted average of the start dates of all records composing the pregnancy, where the weights correspond to the inverse of the variance of gestational age at the record date.

2.3.8 Final refinement

We define range of gestational age for each type of pregnancies end:

- LB: between 147 and 310 day
- T: between 14 and 154 day
- SA: between 14 and 154 day
- UNF: between 14 and 310 day
- SB: between 154 and 310 day
- ECT: between 14 and 154 day
- LOSTFU: between 14 and 310 day
- ONGOING: between 14 and 310 day
- UNK: between 14 and 310 day

For each type of end we calculated the mean gestational age in the pregnancies with start and end not imputed. If the predicted gestation age falls outside this interval, we use this value to re-impute the gestation age.

Finally, the gestation age of overlapping pregnancies is modified, according to section “reconciliation”.

2.3.9 Verification of a sample of pregnancies

At the end of the script, a sample of 30 pregnancies is extracted and made available for manual review by data source’s experts. This procedure aims at assessing whether the local experts’ choice of start date, end date, type of end, and division of pregnancy episodes would have been different. Notably, the extracted pregnancies are among those with the most challenging reconciliation process. The result of the manual assessment is then reviewed together with the leading data partner responsible for the development of the PA, and if corrections or data source’s specific variations are needed, they will be included in the script.

2.4 Mother-children linkage

As described in the section Structure of the Algorithm, the PA processes the information available in the data source on mother-child linkage (which may be originated from birth registry, primary care medical records, etc.) and stores it readily as a product that allows linking the mother to the specific pregnancy where the child was born.

2.5 Future developments

A new variable will be implemented in the PA to identify if a pregnancy is multiple and provide this

information in the final output of the script. This variable will be based on the list of the diagnosis codes for singleton and multiple births (Annex Table 2B). Moreover, if this occurs, the PA will enable us to create an additional data set based on the fetus rather than mother. This additional data set will allow more than one outcome for a same pregnancy when it is recognized as multiple.

2.6 Discussion

In the present deliverable we extensively describe the ConcePTION PA by detailing each step of the algorithm. The novel aspects introduced with the ConcePTION PA are illustrated below.

In previous experience from multi-database European studies, algorithms exploiting diverse data banks were implemented separately by each research partner (*Charlton RA et al, 2014, Cohen et al. 2020*) or records were processed independently from their origin (*Matcho et al 2018*). The ConcePTION PA has been successfully implemented in multiple European Projects (*Abtahi et al, 2023, Duran et al, 2023, Hurley et al, 2024, Covid Vaccine Monitoring Project)* with a different strategy: all data sources participating to the project have been mapped to the ConcePTION Common Data Model that ensures that origin of each record can be retrieved throughout the entire process. It was designed to incorporate all possible scenarios in a transparent and common framework (Lot4 oral retinoids, CVM) with a different strategy: all data sources participating to the project have been mapped to the ConcePTION Common Data Model that ensures that origin of each record can be retrieved throughout the entire process. It was designed to incorporate all possible scenarios in a transparent and common framework.

Another novel aspect of the ConcePTION PA is the inclusion of the predictive model based on a random forest model, to refine the imputation on the pregnancy start date. In previous research (*Zhu et al. 2023*) that aimed at developing and validating an algorithm for the estimation of the gestational age in claims data, the algorithm based on a random forest model resulted to be the best performing one.

Moreover, the ConcePTION PA provides more complete information on the periods when a person in the data source is pregnant, as it allows a wider range of types of end of pregnancy compared to previous algorithms. For instance, pregnancies for which type of end could not be established are retained and displayed in the final cohort of pregnancies. Only one previous study based on U.S. claims data included the unknown outcome of pregnancy, based on the fact that pregnant individuals may lose or switch insurance before the pregnancy has ended (*Sarayani et al. 2020*).

Finally, the ConcePTION PA allows identification of pregnancies that are still ongoing at the time of data extraction, which made the PA particularly suitable for monitoring concomitant drug exposure as it allows to timely detect potential side effects of drugs.

Perinatal pharmacoepidemiology studies can benefit from using the algorithm presented herein to harmonize the methods used, maximizing their reproducibility and transparency, as well as facilitating the comparison of results across studies.

To this purpose, we are providing remarks and recommendations on the use of the ConcePTION PA. Firstly, it is important to note that all individuals included in the cohort generated by the algorithm are confirmed to be pregnant during the recorded dates that compose the pregnancy episode, i.e. the period between the start and end date of pregnancy, irrespective of the assigned quality colour.

Therefore, it would not be appropriate to exclude individuals based on the quality colour assigned by the algorithm, as this colour pertains to the origin and definition of the start, end and type of pregnancy end, rather than the existence of the pregnancy itself. Additionally, pregnancies classified as having an unknown type of end require consideration in relation to the data source used. If the data source has no information from the birth registry nor from the hospital admission registry, or if individuals in that country have the option to choose home delivery, then there is no reason not to consider these pregnancies comparable to pregnancies ending in livebirth. Conversely, if data form the birth registry and hospital admission registry are available and the great majority of individuals gave birth at a hospital, then pregnancies classified as UNK likely have a high prevalence of premature outcomes. In this scenario, excluding individuals with pregnancies with type of end "unknown" from the study cohort could introduce selection bias based on the outcome.

In **Table 2.7** we are providing recommendations for using the ConcePTION PA, specifically in cases where pregnancies with type of end "unknown" and "ongoing" are used for cohort selection in a drug utilization study, cohort selection in a drug safety study, matching criteria, covariate, outcome definition.

Table 2.7 Recommendation for the use of the ConcePTION PA

	Type of end UNKNOWN		**Type of end ONGOING**	
Intended use	**Data source in which this is a proxy for a prematurely terminated pregnancy**	**Other data sources**	**Data source in which this is a proxy for a prematurely terminated pregnancy**	**Other data sources**
Matching criteria	Match using person-time	Match using per-sontime	Match using person-time	Match using per-sontime
Covariate	Classify pregnancy status based on time of covariate measurement	Classify pregnancy status based on time of covariate measurement	Classify pregnancy status based on time of covariate measurement	Classify pregnancy status based on time of covariate measurement
Outcome definition	Can be assumed to be premature end	No assumptions can be made	Not occurred during study period	Not occurred during study period
Cohort selection criteria – drug utilization study (including studies on risk minimization measurement)	Pregnancies **must be included,** at least during 'pregnant person-time', because they are different from the other pregnancies	Pregnancies **may be discarded** at the price of reducing power	If part of the study period, pregnancies **must be included**	If part of the study period, pregnancies **may be included**
Cohort selection criteria – drug safety study	Pregnancies **cannot be discarded**: selection based on outcome. Uncertainty on exposure time must be handled	Pregnancies **must be discarded**, although this reduces power, and still representativeness is at risk	**Study period** when pregnancies have to had a chance to end **must be discarded**	**Study period** allowing pregnancies to have a chance to end **must be discarded**

Another important aspect of the ConcePTION pregnancy algorithm is the inclusion of a verification file for each data source in the output of the script. This file is generated by extracting a pool of pregnancies from the final output, allowing for an *a priori* check of the reliability of the obtained results. This verification tool may also help prevent misclassification of pregnancy episodes, such as the inclusion of diagnosistic codes that do not necessarily imply that the person is pregnant at the record date. This situation was observed, for example, in the CPRD GOLD data while evaluating the

impact of the pregnancy prevention program in post-authorization safety studies (*Lee C et al. 2023*). In this instance, the inflated number of pregnancy episodes during the post-implementation period resulted from a heightened use of records for advice consultations, with no other codes indicating evidence of an ongoing pregnancy for those individuals.

Some limitations must be acknowledged. Up to and including script version 5.2.7, the outcome of the LB type from the pregnancy algorithm should be considered at face value, although the start and end dates were valid. Specifically, potential misclassification of LB type can occur due to the assignment of LB when, for instance, there is a diagnosis code of cesarean delivery and no other codes or records of pregnancy end. This may be incorrect since this diagnosis code indicates the end of pregnancy rather than the outcome of the pregnancy. Also, misclassification of LB/SB can occur in case of discordant records for a single pregnancy, as the algorithm chooses the best outcome, which is live birth. This limitation was overcome in script version 5.3.

Also, from the verification procedure, it appeared that discordant yellow-yellow and green-green pregnancies present errors. However, it should be noted that these are small numbers when compared to the total number of pregnancies. It is possible to decide to exclude them from the final output or proceed with manual verification. In general, the algorithm allows manual review of all generated pregnancies, and it will be added as a recommendation to manually review the yellow-yellow discordant cases.

Acknowledgements. We would like to thank all the researchers who contributed to the creation and development of the diagnosis codes lists used in the ConcePTION pregnancy algorithm: Judit Riera, Carlos Duràn, Afonso Ana Sofia Afonso, Caitlin Dodd, Vera Ehrenstein, Patricia Garcia, Giulia Hyeraci, Eimily Hurley, Maryline Le Noan-Laine, Mar Martin-Pérez, Vera Mitter, Sima Mohammadi, Hedvig Nordeng, Romin Pajohueshnia, Tania Shink, Miriam Sturkenboom. We also extend our gratitude to all the data partners who participated in the project for their efforts in providing local expertise, running the script, and validating the results.

2.7 References

(Abtahi et al. 2023) Abtahi S, Pajouheshnia R, Durán CE, Riera-Arnau J, Gamba M, Alsina E, et al. Impact of 2018 EU Risk Minimisation Measures and Revised Pregnancy Prevention Programme on Utilisation and Prescribing Trends of Medicinal Products Containing Valproate: An Interrupted Time Series Study. Drug Saf. 2023 Jul;46(7):689-702.

(Bertoia et al, 2022) Bertoia ML, Phiri K, Clifford CR, et al. Identification of pregnancies and infants within a US commercial healthcare administrative claims database. Pharmacoepidemiol Drug Saf. 2022 Aug;31(8):863-874.

(Becker et al, 2017) Becker BFH, Avillach P, Romio S, van Mulligen EM, Weibel D, Sturkenboom MCJM, et al. CodeMapper: semiautomatic coding of case definitions. A contribution from the ADVANCE project. Pharmacoepidemiol Drug Saf. 2017 Aug;26(8):998–1005.

(Campbell et al, 2022) Campbell J, Krishnan BST, Williams R, McDonald HI, and Minassian C. 2022. 'Investigating the Optimal Handling of Uncertain Pregnancy Episodes in the CPRD GOLD Pregnancy Register: A Methodological Study Using UK Primary Care Data'. BMJ Open 12 (2): e055773.

(Charlton et al, 2014) Charlton RA, Neville AJ, Jordan S, et al. 'Healthcare Databases in Europe for Studying Medicine Use and Safety during Pregnancy'. Pharmacoepidemiology and Drug Safety 23 (6): 586–94.

(Chomstek et al, 2023) Chomistek AK, Phiri K, Doherty MC, et al. Development and Validation of ICD-10-CM-based Algorithms for Date of Last Menstrual Period, Pregnancy Outcomes, and Infant Outcomes. Drug Saf. 2023 Feb;46(2):209-222.

(Cohen et al, 2020) Cohen JM, Cesta CE, Furu K, et al. Prevalence trends and individual patterns of antiepileptic drug use in pregnancy 2006-2016: A study in the five Nordic countries, United States, and Australia. Pharmacoepidemiol Drug Saf. 2020 Aug;29(8):913-922.

(Covid Vaccine Monitoring Project) EMA COVID-19 Vaccine Monitor (CVM): Rapid Safety Assessment of SARS-CoV-2 Vaccines in EU Member States Using Electronic Health Care Datasources. GitHub repository: https://github.com/VAC4EU/CVM/wiki/Covid-Vaccine-Monitoring-final-scripts-repository

(Donati et al, 2011) Donati S, Senatore S, Ronconi A, Regional maternal mortality working group. Maternal mortality in Italy: a record-linkage study. BJOG. 2011 Jun;118(7):872–9.

(Donati et al, 2018) Donati S, Maraschini A, Lega I, D'Aloja P, Buoncristiano M, Manno V, et al. Maternal mortality in Italy: Results and perspectives of record-linkage analysis. Acta Obstet Gynecol Scand. 2018 Nov;97(11):1317–24.

(Duran et al, 2023) Durán CE, Riera-Arnau J, Abtahi S, Pajouheshnia R, Hoxhaj V, Gamba M, et al. Impact of the 2018 revised Pregnancy Prevention Programme by the European Medicines Agency on the use of oral retinoids in females of childbearing age in Denmark, Italy, Netherlands, and Spain: an interrupted time series analysis. Front. Pharmacol. 2023 14:1207976.

(Girardi et al, 2024) Girardi A, Riera-Arnau J, Duràn CE, Sturkenboom M, Afonso A, Dodd C, et al. Pregnancy-related diagnosis codelist developed for the ConcePTION pregnancy algorithm. 2024. 10.5281/zenodo.13341149

(Hurley et al, 2024) Hurley E, Geisler BP, Lupattelli A, Poblador-Plou B, Lassalle R, Jové J, et al. COVID-19 and pregnancy: A European study on pre- and post-infection medication use. Eur J Clin Pharmacol. 2024 May;80(5):707–16.

(Lee et al, 2024) Lee C, Rizzi S, Bierrenbach AL, et al. Identification of misclassified pregnancy episodes in women of childbearing potential exposed to drugs with known teratogenic potential in the CPRD GOLD Pregnancy Register. Pharmacoepidemiol Drug Saf. 2024 Feb;33(2):e5761.

(Margulis et al, 2023) Margulis AV, Huybrechts K. Identification of pregnancies in healthcare data: A changing landscape. Pharmacoepidemiol Drug Saf. 2023 Jan;32(1):84-86. doi: 10.1002/pds.5526. Epub 2022 Sep 3.

(Matcho et al 2018) Matcho A, Ryan P, Fife D, Gifkins D, Knoll C, Friedman A. 2018. 'Inferring Pregnancy Episodes and Outcomes within a Network of Observational Databases'. PLoS ONE 13 (2).

(Nordeng et al, 2024) Nordeng H, Lupattelli A, Engjom, HM et al. Detecting and dating early non-live pregnancy outcomes: generation of a novel pregnancy algorithm from Norwegian linked health registries. Authorea. April 18, 2024.

(Sanchez-Ortiz et al, 2020) Ortiz SS, García AL, Astasio P, Huerta C, Soriano LC. An algorithm to identify pregnancies in BIFAP Primary Care database in Spain: Results from a cohort of 155 419 pregnancies. Pharmacoepidemiology and Drug Safety. 2020;29(1):57–68.

(Sarayani et al, 2020) Sarayani, Amir, Xi Wang, Thuy Nhu Thai, Yasser Albogami, Nakyung Jeon, and Almut G. Winterstein. 2020. 'Impact of the Transition from ICD-9-CM to ICD-10-CM on the Identification of Pregnancy Episodes in US Health Insurance Claims Data'. Clinical Epidemiology 12: 1129–38.

(Shink et al, 2020) Schink T, Wentzell N, Dathe K, Onken M, Haug U. 'Estimating the Beginning of Pregnancy in German Claims Data: Development of an Algorithm With a Focus on the Expected Delivery Date'. Frontiers in Public Health 2020 8: 350.

(Thurin et al, 2021) Thurin NH, Pajouheshnia R, Roberto G, et al. 'From Inception to ConcePTION: Genesis of a Network to Support Better Monitoring and Communication of Medication Safety During Pregnancy and Breastfeeding'. Clinical Pharmacology & Therapeutics 2021 Jan;111(1):321-331.

(Zhu et al, 2023) Zhu Y, Thai TN, Hernandez-Diaz S, et al. Development and Validation of Algorithms to Estimate Live Birth Gestational Age in Medicaid Analytic eXtract Data. Epidemiology. 2023 Jan 1;34(1):69-79.

Annex to Chapter 2

Table 2A. Mapping between concept sets and pregnancy variables.

		type of end	Start date	Ongoing Date	End date	meaning start	meaning ongoing	meaning end	imputed start	imputed end	color quality	order quality
Start recorded, end UNK	Gestation_less24_UNK	UNK	Record date - 154	record date	start date + 280	from_CONCEPTSET	Record_date_CONCEPTSET	imputed_from_CONCEPTSET	0	1		40
	Gestation_24_UNK	UNK	Record date - 168	record date	start date + 280	from_CONCEPTSET	Record_date_CONCEPTSET	imputed_from_CONCEPTSET	0	1		40
	Gestation_25_26_UNK	UNK	Record date - 178	record date	start date + 280	from_CONCEPTSET	Record_date_CONCEPTSET	imputed_from_CONCEPTSET	0	1		40
	Gestation_27_28_UNK	UNK	Record date - 192	record date	start date + 280	from_CONCEPTSET	Record_date_CONCEPTSET	imputed_from_CONCEPTSET	0	1		40
	Gestation_29_30_UNK	UNK	Record date - 206	record date	start date + 280	from_CONCEPTSET	Record_date_CONCEPTSET	imputed_from_CONCEPTSET	0	1		40
	Gestation_31_32_UNK	UNK	Record date - 220	record date	start date + 280	from_CONCEPTSET	Record_date_CONCEPTSET	imputed_from_CONCEPTSET	0	1		40
	Gestation_33_34_UNK	UNK	Record date - 234	record date	start date + 280	from_CONCEPTSET	Record_date_CONCEPTSET	imputed_from_CONCEPTSET	0	1		40
	Gestation_35_36_UNK	UNK	Record date - 248	record date	start date + 280	from_CONCEPTSET	Record_date_CONCEPTSET	imputed_from_CONCEPTSET	0	1		40
	Gestation_more37_UNK	UNK	Record date - 266	record date	start date + 280	from_CONCEPTSET	Record_date_CONCEPTSET	imputed_from_CONCEPTSET	0	1		40

Start recorded, end LB	Gestation_less24_LB	LB	Record date - 154	NA	record date	from_CONCEPTSET	Record_date_CONCEPTSET	from_CONCEPTSET	0	0		4
	Gestation_24_LB	LB	Record date - 168	NA	record date	from_CONCEPTSET	Record_date_CONCEPTSET	from_CONCEPTSET	0	0		4
	Gestation_25_26_LB	LB	Record date - 178	NA	record date	from_CONCEPTSET	Record_date_CONCEPTSET	from_CONCEPTSET	0	0		4
	Gestation_27_28_LB	LB	Record date - 192	NA	record date	from_CONCEPTSET	Record_date_CONCEPTSET	from_CONCEPTSET	0	0		4
	Gestation_29_30_LB	LB	Record date - 206	NA	record date	from_CONCEPTSET	Record_date_CONCEPTSET	from_CONCEPTSET	0	0		4
	Gestation_31_32_LB	LB	Record date - 220	NA	record date	from_CONCEPTSET	Record_date_CONCEPTSET	from_CONCEPTSET	0	0		4
	Gestation_33_34_LB	LB	Record date - 234	NA	record date	from_CONCEPTSET	Record_date_CONCEPTSET	from_CONCEPTSET	0	0		4
	Gestation_35_36_LB	LB	Record date - 248	NA	record date	from_CONCEPTSET	Record_date_CONCEPTSET	from_CONCEPTSET	0	0		4
	Gestation_more37_LB	LB	Record date - 266	NA	record date	from_CONCEPTSET	Record_date_CONCEPTSET	from_CONCEPTSET	0	0		4
Ongoing	Ongoingpregnancy	UNK	Record date - 55	record date	start date + 280	imputed_from_CONCEPTSET	Record_date_CONCEPTSET	imputed_from_CONCEPTSET	1	1		50
	GESTDIAB	UNK	Record date - 55	record date	start date + 280	imputed_from_CONCEPTSET	Record_date_CONCEPTSET	imputed_from_CONCEPTSET	1	1		50
	FGR	UNK	Record date - 55	record date	start date + 280	imputed_from_CONCEPTSET	Record_date_CONCEPTSET	imputed_from_CONCEPTSET	1	1		50

	PREECLAMP	UNK	Record date - 55	record date	start date + 280	imputed_from_CONCEPTSET	Record_date_CONCEPTSET	imputed_from_CONCEPTSET	1	1		50
	PREG_BLEEDING	UNK	Record date - 55	record date	start date + 280	imputed_from_CONCEPTSET	Record_date_CONCEPTSET	imputed_from_CONCEPTSET	1	1		50
	procedures_ongoing	UNK	Record date - 55	record date	start date + 280	imputed_from_CONCEPTSET	Record_date_CONCEPTSET	imputed_from_CONCEPTSET	1	1		50
	fetal_nuchal_translucency	UNK	Record date - 55	record date	start date + 280	imputed_from_CONCEPTSET	Record_date_CONCEPTSET	imputed_from_CONCEPTSET	1	1		50
	amniocentesis	UNK	Record date - 55	record date	start date + 280	imputed_from_CONCEPTSET	Record_date_CONCEPTSET	imputed_from_CONCEPTSET	1	1		50
	Chorionic_Villus_Sampling	UNK	Record date - 55	record date	start date + 280	imputed_from_CONCEPTSET	Record_date_CONCEPTSET	imputed_from_CONCEPTSET	1	1		50
	others	UNK	Record date - 55	record date	start date + 280	imputed_from_CONCEPTSET	Record_date_CONCEPTSET	imputed_from_CONCEPTSET	1	1		50
End LB	AtTermLB	LB	Record date -280	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		8
	BirthNarrowLB	LB	Record date -280	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		9
	PretermLB	LB	Record date -250	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		8
	procedures_livebirth	LB	Record date -280	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		10
End BUNSP	BirthNarrowBUNSP	UNSP	Record date -280	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		13

	PretermBUNSP	UNSP	Record date -250	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		12
	AtTermBUNSP	UNSP	Record date -280	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		12
	PostTermBUNSP	UNSP	Record date -300	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		12
	procedures_delivery	UNSP	Record date -280	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		14
End UNSP UNK	EndUnspecified	UNK	Record date -280	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		23
	Birth_possible	UNK	Record date -280	NA	record date	imputed_from_CONCEPTSET	NA	imputed_from_CONCEPTSET	1	1		25
	procedures_end_UNK	UNK	Record date -280	NA	record date	imputed_from_CONCEPTSET	NA	imputed_from_CONCEPTSET	1	0		24
End T SA SB ECT	Stillbirth_narrow	SB	Record date - 280	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		11
	Interruption_narrow	T	Record date - 70	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		15
	Spontaneousabortion_narrow	SA	Record date - 70	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		17
	Ectopicpregnancy	ECT-MOL	Record date - 70	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		19
	Molarpregnancy	ECT-MOL	Record date - 70	NA	Record date	imputed_from_CONCEPTSET	NA	rom_CONCEPTSET	1	0		19

	procedures_termination	T	Record date - 70	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		16
	procedures_spontaneous_abortion	SA	Record date - 70	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		18
	procedures_ectopic	ECT	Record date - 70	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		20
End UNF	Interruption_possible	UNF	Record date - 70	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		21
	Stillbirth_possible	UNF	Record date - 280	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		21
	Spontaneousabortion_possible	UNF	Record date - 70	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		21
	procedures_end_UNK	UNF	Record date - 71	NA	record date	imputed_from_CONCEPTSET	NA	from_CONCEPTSET	1	0		24
End LB	Gestation_less24_CHILD	LB	Record date - 154	NA	record date	from_CONCEPTSET	Record_date_CONCEPTSET	from_CONCEPTSET	0	0		4
	Gestation_24_CHILD	LB	Record date - 168	NA	record date	from_CONCEPTSET	Record_date_CONCEPTSET	from_CONCEPTSET	0	0		4
	Gestation_25_26_CHILD	LB	Record date - 178	NA	record date	from_CONCEPTSET	Record_date_CONCEPTSET	from_CONCEPTSET	0	0		4
	Gestation_27_28_CHILD	LB	Record date - 192	NA	record date	from_CONCEPTSET	Record_date_CONCEPTSET	from_CONCEPTSET	0	0		4
	Gestation_29_30_CHILD	LB	Record date - 206	NA	record date	from_CONCEPTSET	Record_date_CONCEPTSET	from_CONCEPTSET	0	0		4

Gestation_31_32_CHILD	LB	Record date - 220	NA	record date	from_CONCEPTSET	Record_date_CONCEPTSET	from_CONCEPTSET	0	0		4
Gestation_33_34_CHILD	LB	Record date - 234	NA	record date	from_CONCEPTSET	Record_date_CONCEPTSET	from_CONCEPTSET	0	0		4
Gestation_35_36_CHILD	LB	Record date - 248	NA	record date	from_CONCEPTSET	Record_date_CONCEPTSET	from_CONCEPTSET	0	0		4
Gestation_more37_CHILD	LB	Record date - 266	NA	record date	from_CONCEPTSET	Record_date_CONCEPTSET	from_CONCEPTSET	0	0		4

Table 2B. Diagnosis codes for singleton and multiple pregnancy.

event_definition	coding_system	code	code_name	tags
Singleton_mother	ICD10	O80	Single spontaneous delivery	Narrow
Singleton_mother	ICD10	O80.0	Spontaneous vertex delivery	Narrow
Singleton_mother	ICD10	O80.1	Spontaneous breech delivery	Narrow
Singleton_mother	ICD10	O80.8	Other single spontaneous delivery	Narrow
Singleton_mother	ICD10	O80.9	Single spontaneous delivery, unspecified Spontaneous delivery NOS	Narrow
Singleton_mother	ICD10	O81	Single delivery by forceps and vacuum extractor	Narrow
Singleton_mother	ICD10	O81.0	Low forceps delivery	Narrow
Singleton_mother	ICD10	O81.1	Mid-cavity forceps delivery	Narrow
Singleton_mother	ICD10	O81.2	Mid-cavity forceps with rotation	Narrow
Singleton_mother	ICD10	O81.3	Other and unspecified forceps delivery	Narrow
Singleton_mother	ICD10	O81.4	Vacuum extractor delivery	Narrow
Singleton_mother	ICD10	O81.5	Delivery by combination of forceps and vacuum extractor	Narrow
Singleton_mother	ICD10	O82	Single delivery by caesarean section	Narrow
Singleton_mother	ICD10	O82.0	Delivery by elective caesarean section Repeat caesarean section NOS	Narrow
Singleton_mother	ICD10	O82.1	Delivery by emergency caesarean section	Narrow
Singleton_mother	ICD10	O82.2	Delivery by caesarean hysterectomy	Narrow
Singleton_mother	ICD10	O82.8	Other single delivery by caesarean section	Narrow
Singleton_mother	ICD10	O82.9	Delivery by caesarean section, unspecified	Narrow
Singleton_mother	ICD10	O83	Other assisted single delivery	Narrow
Singleton_mother	ICD10	O83.0	Breech extraction	Narrow
Singleton_mother	ICD10	O83.1	Other assisted breech delivery, Breech delivery NOS	Narrow
Singleton_mother	ICD10	O83.2	Other manipulation-assisted delivery, Version with extraction	Narrow
Singleton_mother	ICD10	O83.3	Delivery of viable fetus in abdominal pregnancy	Narrow
Singleton_mother	ICD10	O83.4	Destructive operation for delivery	Narrow
Singleton_mother	ICD10	O83.8	Other specified assisted single delivery	Narrow
Singleton_mother	ICD10	O83.9	Assisted single delivery, unspecified	Narrow
Singleton_mother	ICD10	Z37.0	Single live birth	Narrow

Singleton_mother	ICD10	Z37.1	Single stillbirth	Narrow
Singleton_child	ICD10	Z38.0	Singleton, born in hospital	Narrow
Singleton_child	ICD10	Z38.1	Singleton, born outside hospital	Narrow
Singleton_child	ICD10	Z38.2	Singleton, unspecified as to place of birth Liveborn infant NOS	Narrow
Multiple_mother	ICD10	O30	Multiple gestation	Narrow
Multiple_mother	ICD10	O30.0	Twin pregnancy	Narrow
Multiple_mother	ICD10	O30.1	Triplet pregnancy	Narrow
Multiple_mother	ICD10	O30.2	Quadruplet pregnancy	Narrow
Multiple_mother	ICD10	O30.8	Other multiple gestation	Narrow
Multiple_mother	ICD10	O30.9	Multiple gestation, unspecified	Narrow
Multiple_mother	ICD10	O31	Complications specific to multiple gestation	Narrow
Multiple_mother	ICD10	O31.0	Papyraceous fetus	Narrow
Multiple_mother	ICD10	O31.1	Continuing pregnancy after abortion of one fetus or more	Narrow
Multiple_mother	ICD10	O31.2	Continuing pregnancy after intrauterine death of one fetus or more	Narrow
Multiple_mother	ICD10	O31.8	Other complications specific to multiple gestation	Narrow
Multiple_mother	ICD10	O32.5	Maternal care for multiple gestation with malpresentation of one fetus or more	Narrow
Multiple_mother	ICD10	O63.2	Delayed delivery of second twin, triplet, etc.	Narrow
Multiple_mother	ICD10	O66.1	Obstructed labour due to locked twins	Narrow
Multiple_mother	ICD10	O84	Multiple delivery ***Note***: Use additional code (O80-O83), if desired, to indicate the method of delivery of each fetus or infant.	Narrow
Multiple_mother	ICD10	O84.0	Multiple delivery, all spontaneous	Narrow
Multiple_mother	ICD10	O84.1	Multiple delivery, all by forceps and vacuum extractor	Narrow
Multiple_mother	ICD10	O84.2	Multiple delivery, all by caesarean section	Narrow
Multiple_mother	ICD10	O84.8	Other multiple delivery	Narrow
Multiple_mother	ICD10	O84.9	Multiple delivery, unspecified	Narrow
Multiple_mother	ICD10	Z37.2	Twins, both liveborn	Narrow
Multiple_mother	ICD10	Z37.3	Twins, one liveborn and one stillborn	Narrow
Multiple_mother	ICD10	Z37.4	Twins, both stillborn	Narrow
Multiple_mother	ICD10	Z37.5	Other multiple births, all liveborn	Narrow
Multiple_mother	ICD10	Z37.6	Other multiple births, some liveborn	Narrow
Multiple_mother	ICD10	Z37.7	Other multiple births, all stillborn	Narrow
Multiple_child	ICD10	Z38.3	Twin, born in hospital	Narrow

Multiple_child	ICD10	Z38.4	Twin, born outside hospital	Narrow
Multiple_child	ICD10	Z38.5	Twin, unspecified as to place of birth	Narrow
Multiple_child	ICD10	Z38.6	Other multiple, born in hospital	Narrow
Multiple_child	ICD10	Z38.7	Other multiple, born outside hospital	Narrow
Multiple_child	ICD10	Z38.8	Other multiple, unspecified as to place of birth	Narrow
Single_child	ICD9	V30	Single liveborn	Narrow
Multiple_mother	ICD9	651	Multiple gestation	Narrow
Multiple_mother	ICD9	651.0	Twins, unspecified	Narrow
Multiple_mother	ICD9	651.1	Triplet, unspecified	Narrow
Multiple_mother	ICD9	660.51	Locked twins, delivered	Narrow
Multiple_mother	ICD9	678.1	Fetal conjoined twins	Narrow
Multiple_child	ICD9	V31	Twin birth mate liveborn	Narrow
Multiple_child	ICD9	V32	Twin birth mate stillborn	Narrow
Multiple_child	ICD9	V33	Twin birth unspecified whether mate liveborn or stillborn	Narrow
Multiple_child	ICD9	V34	Other multiple birth (three or more) mates all liveborn	Narrow
Multiple_child	ICD9	V35	Other multiple birth (three or more) mates all stillborn	Narrow
Multiple_child	ICD9	V36	Other multiple birth (three or more) mates liveborn and stillborn	Narrow
Multiple_child	ICD9	V37	Other multiple birth (three or more) unspecified whether mates liveborn or stillborn	Narrow
Multiple_child	ICD9	652.6	Multiple gestation with malpresentation of one fetus or more	Narrow
Multiple_child	ICD9	652.61	Multiple gestation with malpresentation of one fetus or more, delivered	Narrow
Singleton_child	ICD9CM	V27.0	Outcome of delivery, single liveborn	Narrow
Singleton_child	ICD9CM	V27.1	Outcome of delivery, single stillborn	Narrow
Singleton_child	ICD9CM	V30	Single liveborn	Narrow
Singleton_child	ICD9CM	V30.0	Singleton, born in hospital	Narrow
Singleton_child	ICD9CM	V30.00	Single liveborn, born in hospital, delivered without mention of cesarean section	Narrow
Singleton_child	ICD9CM	V30.01	Single liveborn, born in hospital, delivered by cesarean section	Narrow
Singleton_child	ICD9CM	V30.1	Single liveborn, born before admission to hospital	Narrow
Singleton_child	ICD9CM	V30.2	Singleton, born outside hospital and not hospitalised	Narrow
Multiple_mother	ICD9CM	651	Multiple gestation	Narrow
Multiple_mother	ICD9CM	651.0	Twin pregnancy	Narrow
Multiple_mother	ICD9CM	651.00	Twin pregnancy, unspecified as to episode of care or not applicable	Narrow
Multiple_mother	ICD9CM	651.01	Twin pregnancy, delivered, with or without mention of antepartum condition	Narrow

Multiple_mother	ICD9CM	651.03	Twin pregnancy, antepartum condition or complicatio	Narrow
Multiple_mother	ICD9CM	651.1	Triplet pregnancy	Narrow
Multiple_mother	ICD9CM	651.10	Triplet pregnancy, unspecified as to episode of care or not applicable	Narrow
Multiple_mother	ICD9CM	651.11	Triplet pregnancy, delivered, with or without mention of antepartum condition	Narrow
Multiple_mother	ICD9CM	651.2	Quadruplet pregnancy	Narrow
Multiple_mother	ICD9CM	651.20	Quadruplet pregnancy, unspecified as to episode of care or not applicable	Narrow
Multiple_mother	ICD9CM	651.21	Quadruplet pregnancy, delivered, with or without mention of antepartum condition	Narrow
Multiple_mother	ICD9CM	651.23	Quadruplet pregnancy, antepartum condition or complication	Narrow
Multiple_mother	ICD9CM	651.3	Twin pregnancy with fetal loss and retention of one fetus	Narrow
Multiple_mother	ICD9CM	651.30	Twin pregnancy with fetal loss and retention of one fetus, unspecified as to episode of care or not applicable	Narrow
Multiple_mother	ICD9CM	651.31	Twin pregnancy with fetal loss and retention of one fetus, delivered, with or without mention of antepartum condition	Narrow
Multiple_mother	ICD9CM	651.33	Twin pregnancy with fetal loss and retention of one fetus, antepartum condition or complication	Narrow
Multiple_mother	ICD9CM	651.4	Triplet pregnancy with fetal loss and retention of one or more fetus(es)	Narrow
Multiple_mother	ICD9CM	651.40	Triplet pregnancy with fetal loss and retention of one or more fetus(es), unspecified as to episode of care or not applicable	Narrow
Multiple_mother	ICD9CM	651.41	Triplet pregnancy with fetal loss and retention of one or more fetus(es), delivered, with or without mention of antepartum condition	Narrow
Multiple_mother	ICD9CM	651.43	Triplet pregnancy with fetal loss and retention of one or more fetus(es), antepartum condition or complication	Narrow
Multiple_mother	ICD9CM	651.5	Quadruplet pregnancy with fetal loss and retention of one or more fetus	Narrow
Multiple_mother	ICD9CM	651.50	Quadruplet pregnancy with fetal loss and retention of one or more fetus(es), unspecified as to episode of care or not applicable	Narrow
Multiple_mother	ICD9CM	651.51	Quadruplet pregnancy with fetal loss and retention of one or more fetus(es), delivered, with or without mention of antepartum condition	Narrow
Multiple_mother	ICD9CM	651.53	Quadruplet pregnancy with fetal loss and retention of one or more fetus(es), antepartum condition or complication	Narrow
Multiple_mother	ICD9CM	651.6	Other multiple pregnancy with fetal loss and retention of one or more fetus(es)	Narrow
Multiple_mother	ICD9CM	651.60	Other multiple pregnancy with fetal loss and retention of one or more fetus(es), unspecified as to episode of care or not applicable	Narrow
Multiple_mother	ICD9CM	651.61	Other multiple pregnancy with fetal loss and retention of one or more fetus(es), delivered, with or without mention of antepartum condition	Narrow
Multiple_mother	ICD9CM	651.63	Other multiple pregnancy with fetal loss and retention of one or more fetus(es),	Narrow

			antepartum condition or complication	
Multiple_mother	ICD9CM	651.7	Multiple gestation following (elective) fetal reduction	Narrow
Multiple_mother	ICD9CM	651.70	Multiple gestation following (elective) fetal reduction, unspecified as to episode of care or not applicable	Narrow
Multiple_mother	ICD9CM	651.71	Multiple gestation following (elective) fetal reduction,delivered, with or without mention of antepartum condition	Narrow
Multiple_mother	ICD9CM	651.73	Multiple gestation following (elective) fetal reduction, antepartum condition or complication	Narrow
Multiple_mother	ICD9CM	651.8	Other specified multiple gestation	Narrow
Multiple_mother	ICD9CM	651.80	Other specified multiple gestation, unspecified as to episode of care or not applicable	Narrow
Multiple_mother	ICD9CM	651.81	Other specified multiple gestation, delivered, with or without mention of antepartum condition	Narrow
Multiple_mother	ICD9CM	651.83	Other specified multiple gestation, antepartum condition or complication	Narrow
Multiple_mother	ICD9CM	651.9	Unspecified multiple gestation	Narrow
Multiple_mother	ICD9CM	651.90	Unspecified multiple gestation, unspecified as to episode of care or not applicable	Narrow
Multiple_mother	ICD9CM	651.91	Unspecified multiple gestation, delivered, with or without mention of antepartum condition	Narrow
Multiple_mother	ICD9CM	651.93	Unspecified multiple gestation, antepartum condition or complication	Narrow
Multiple_mother	ICD9CM	652.6	Multiple gestation with malpresentation of one fetus or more	Narrow
Multiple_mother	ICD9CM	652.60	Multiple gestation with malpresentation of one fetus or more, unspecified as to episode of care or not applicable	Narrow
Multiple_mother	ICD9CM	652.61	Multiple gestation with malpresentation of one fetus or more, delivered	Narrow
Multiple_mother	ICD9CM	652.63	Multiple gestation with malpresentation of one fetus or more, antepartum	Narrow
Multiple_mother	ICD9CM	660.5	Locked twins	Narrow
Multiple_mother	ICD9CM	660.50	Locked twins, unspecified as to episode of care or not applicable	Narrow
Multiple_mother	ICD9CM	660.51	Locked twins, delivered, with or without mention of antepartum condition	Narrow
Multiple_mother	ICD9CM	660.53	Locked twins, antepartum condition or complication	Narrow
Multiple_mother	ICD9CM	662.3	Delayed delivery of second twin triplet	Narrow
Multiple_mother	ICD9CM	662.30	Delayed delivery of second twin, triplet, etc., unspecified as to episode of care or not applicable	Narrow
Multiple_mother	ICD9CM	662.31	Delayed delivery of second twin, triplet, etc., delivered, with or without mention of antepartum condition	Narrow
Multiple_mother	ICD9CM	662.33	Delayed delivery of second twin, triplet, etc., antepartum condition or complication	Narrow
Multiple_mother	ICD9CM	678.1	Fetal conjoined twins	Narrow
Multiple_mother	ICD9CM	678.10	Fetal conjoined twins, unspecified as to episode of care or not applicable	Narrow

Multiple_mother	ICD9CM	678.11	Fetal conjoined twins, delivered, with or without mention of antepartum condition	Narrow
Multiple_mother	ICD9CM	678.13	Fetal conjoined twins, antepartum condition or complication	Narrow
Multiple_child	ICD9CM	V27.2	Outcome of delivery, twins, both liveborn	Narrow
Multiple_child	ICD9CM	V27.3	Outcome of delivery, twins, one liveborn and one stillborn	Narrow
Multiple_child	ICD9CM	V27.4	Outcome of delivery, twins, both stillborn	Narrow
Multiple_child	ICD9CM	V27.5	Outcome of delivery, other multiple birth, all liveborn	Narrow
Multiple_child	ICD9CM	V27.6	Outcome of delivery, other multiple birth, some liveborn	Narrow
Multiple_child	ICD9CM	V27.7	Outcome of delivery, other multiple birth, all stillborn	Narrow
Multiple_child	ICD9CM	V31	Twin birth mate liveborn	Narrow
Multiple_child	ICD9CM	V31.0	Twin birth mate liveborn born in hospital	Narrow
Multiple_child	ICD9CM	V31.00	Twin birth, mate liveborn, born in hospital, delivered without mention of cesarean section	Narrow
Multiple_child	ICD9CM	V31.01	Twin birth, mate liveborn, born in hospital, delivered by cesarean section	Narrow
Multiple_child	ICD9CM	V31.1	Twin birth, mate liveborn, born before admission to hospital	Narrow
Multiple_child	ICD9CM	V31.2	Twin birth, mate liveborn, born outside hospital and not hospitalized	Narrow
Multiple_child	ICD9CM	V32	Twin birth mate stillborn	Narrow
Multiple_child	ICD9CM	V32.0	Twin birth mate stillborn born in hospital	Narrow
Multiple_child	ICD9CM	V32.00	Twin birth, mate stillborn, born in hospital, delivered without mention of cesarean section	Narrow
Multiple_child	ICD9CM	V32.01	Twin birth, mate stillborn, born in hospital, delivered by cesarean section	Narrow
Multiple_child	ICD9CM	V32.1	Twin birth, mate stillborn, born before admission to hospital	Narrow
Multiple_child	ICD9CM	V32.2	Twin birth, mate stillborn, born outside hospital and not hospitalized	Narrow
Multiple_child	ICD9CM	V33	Twin birth unspecified whether mate liveborn or stillborn	Narrow
Multiple_child	ICD9CM	V33.0	Twin birth unspecified whether mate liveborn or stillborn born in hospital	Narrow
Multiple_child	ICD9CM	V33.00	Twin birth, unspecified whether mate liveborn or stillborn, born in hospital, delivered without mention of cesarean section	Narrow
Multiple_child	ICD9CM	V33.01	Twin birth, unspecified whether mate liveborn or stillborn, born in hospital, delivered by cesarean section	Narrow
Multiple_child	ICD9CM	V33.1	Twin birth, unspecified whether mate liveborn or stillborn, born before admission to hospital	Narrow
Multiple_child	ICD9CM	V33.2	Twin birth, unspecified whether mate liveborn or stillborn, born outside hospital and not hospitalized	Narrow
Multiple_child	ICD9CM	V34	Other multiple birth (three or more) mates all liveborn	Narrow
Multiple_child	ICD9CM	V34.0	Other multiple birth (three or more) mates all liveborn born in hospital	Narrow
Multiple_child	ICD9CM	V34.00	Other multiple birth (three or more), mates all liveborn, born in hospital, delivered without	Narrow

			mention of cesarean section	
Multiple_child	ICD9CM	V34.01	Other multiple birth (three or more), mates all liveborn, born in hospital, delivered by cesarean section	Narrow
Multiple_child	ICD9CM	V34.1	Other multiple birth (three or more), mates all liveborn, born before admission to hospital	Narrow
Multiple_child	ICD9CM	V34.2	Other multiple birth (three or more), mates all liveborn, born outside hospital and not hospitalized	Narrow
Multiple_child	ICD9CM	V35	Other multiple birth (three or more) mates all stillborn	Narrow
Multiple_child	ICD9CM	V35.0	Other multiple birth (three or more), mates all still born, born in hospital	Narrow
Multiple_child	ICD9CM	V35.00	Other multiple birth (three or more), mates all still born, born in hospital, delivered without mention of cesarean section	Narrow
Multiple_child	ICD9CM	V35.01	Other multiple birth (three or more), mates all still born, born in hospital, delivered by cesarean section	Narrow
Multiple_child	ICD9CM	V35.1	Other multiple birth (three or more), mates all stillborn, born before admission to hospital	Narrow
Multiple_child	ICD9CM	V35.2	Other multiple birth (three or more), mates all stillborn, born outside of hospital and not hospitalized	Narrow
Multiple_child	ICD9CM	V36	Other multiple birth (three or more) mates liveborn and stillborn	Narrow
Multiple_child	ICD9CM	V36.0	Other multiple birth (three or more) mates liveborn and stillborn born in hospital	Narrow
Multiple_child	ICD9CM	V36.00	Other multiple birth (three or more), mates liveborn and stillborn, born in hospital, delivered without mention of cesarean section	Narrow
Multiple_child	ICD9CM	V36.01	Other multiple birth (three or more), mates liveborn and stillborn, born in hospital, delivered without mention of cesarean section	Narrow
Multiple_child	ICD9CM	V36.1	Other multiple birth (three or more), mates liveborn and stillborn, born before admission to hospital	Narrow
Multiple_child	ICD9CM	V36.2	Other multiple birth (three or more), mates liveborn and stillborn, born outside hospital and not hospitalized	Narrow
Multiple_child	ICD9CM	V37	Other multiple birth (three or more) unspecified whether mates liveborn or stillborn	Narrow
Multiple_child	ICD9CM	V37.0	Other multiple birth (three or more) unspecified whether mates liveborn or stillborn born in hospital	Narrow
Multiple_child	ICD9CM	V37.00	Other multiple birth (three or more), unspecified whether mates liveborn or stillborn, born in hospital, delivered without mention of cesarean section	Narrow
Multiple_child	ICD9CM	V37.01	Other multiple birth (three or more), unspecified whether mates liveborn or stillborn, born in hospital, delivered by cesarean section	Narrow
Multiple_child	ICD9CM	V37.1	Other multiple birth (three or more), unspecified whether mates liveborn or stillborn, born before admission to hospital	Narrow
Multiple_child	ICD9CM	V37.2	Other multiple birth (three or more), unspecified whether mates liveborn or stillborn, born outside of hospital	Narrow

Multiple_child	ICD9CM	V31	Twin birth mate liveborn	Narrow
Multiple_child	ICD9CM	V31.0	Twin birth mate liveborn born in hospital	Narrow
Multiple_child	ICD9CM	V31.00	Twin birth, mate liveborn, born in hospital, delivered without mention of cesarean section	Narrow
Multiple_child	ICD9CM	V31.01	Twin birth, mate liveborn, born in hospital, delivered by cesarean section	Narrow
Multiple_child	ICD9CM	V31.1	Twin birth, mate liveborn, born before admission to hospital	Narrow
Multiple_child	ICD9CM	V31.2	Twin birth, mate liveborn, born outside hospital and not hospitalized	Narrow
Multiple_child	ICD9CM	V32	Twin birth mate stillborn	Narrow
Multiple_child	ICD9CM	V32.0	Twin birth mate stillborn born in hospital	Narrow
Multiple_child	ICD9CM	V32.00	Twin birth, mate stillborn, born in hospital, delivered without mention of cesarean section	Narrow
Multiple_child	ICD9CM	V32.01	Twin birth, mate stillborn, born in hospital, delivered by cesarean section	Narrow
Multiple_child	ICD9CM	V32.1	Twin birth, mate stillborn, born before admission to hospital	Narrow
Multiple_child	ICD9CM	V32.2	Twin birth, mate stillborn, born outside hospital and not hospitalized	Narrow
Multiple_child	ICD9CM	V33	Twin birth unspecified whether mate liveborn or stillborn	Narrow
Multiple_child	ICD9CM	V33.0	Twin birth unspecified whether mate liveborn or stillborn born in hospital	Narrow
Multiple_child	ICD9CM	V33.00	Twin birth, unspecified whether mate liveborn or stillborn, born in hospital, delivered without mention of cesarean section	Narrow
Multiple_child	ICD9CM	V33.01	Twin birth, unspecified whether mate liveborn or stillborn, born in hospital, delivered by cesarean section	Narrow
Multiple_child	ICD9CM	V33.1	Twin birth, unspecified whether mate liveborn or stillborn, born before admission to hospital	Narrow
Multiple_child	ICD9CM	V33.2	Twin birth, unspecified whether mate liveborn or stillborn, born outside hospital and not hospitalized	Narrow
Multiple_child	ICD9CM	V34	Other multiple birth (three or more) mates all liveborn	Narrow
Multiple_child	ICD9CM	V34.0	Other multiple birth (three or more) mates all liveborn born in hospital	Narrow
Multiple_child	ICD9CM	V34.00	Other multiple birth (three or more), mates all liveborn, born in hospital, delivered without mention of cesarean section	Narrow
Multiple_child	ICD9CM	V34.01	Other multiple birth (three or more), mates all liveborn, born in hospital, delivered by cesarean section	Narrow
Multiple_child	ICD9CM	V34.1	Other multiple birth (three or more), mates all liveborn, born before admission to hospital	Narrow
Multiple_child	ICD9CM	V34.2	Other multiple birth (three or more), mates all liveborn, born outside hospital and not hospitalized	Narrow
Multiple_child	ICD9CM	V35	Other multiple birth (three or more) mates all stillborn	Narrow
Multiple_child	ICD9CM	V35.0	Other multiple birth (three or more), mates all still born, born in hospital	Narrow
Multiple_child	ICD9CM	V35.00	Other multiple birth (three or more), mates all still born, born in hospital, delivered without	Narrow

			mention of cesarean section	
Multiple_child	ICD9CM	V35.01	Other multiple birth (three or more), mates all still born, born in hospital, delivered by cesarean section	Narrow
Multiple_child	ICD9CM	V35.1	Other multiple birth (three or more), mates all stillborn, born before admission to hospital	Narrow
Multiple_child	ICD9CM	V35.2	Other multiple birth (three or more), mates all stillborn, born outside of hospital and not hospitalized	Narrow
Multiple_child	ICD9CM	V36	Other multiple birth (three or more) mates liveborn and stillborn	Narrow
Multiple_child	ICD9CM	V36.0	Other multiple birth (three or more) mates liveborn and stillborn born in hospital	Narrow
Multiple_child	ICD9CM	V36.00	Other multiple birth (three or more), mates liveborn and stillborn, born in hospital, delivered without mention of cesarean section	Narrow
Multiple_child	ICD9CM	V36.01	Other multiple birth (three or more), mates liveborn and stillborn, born in hospital, delivered without mention of cesarean section	Narrow
Multiple_child	ICD9CM	V36.1	Other multiple birth (three or more), mates liveborn and stillborn, born before admission to hospital	Narrow
Multiple_child	ICD9CM	V36.2	Other multiple birth (three or more), mates liveborn and stillborn, born outside hospital and not hospitalized	Narrow
Multiple_child	ICD9CM	V37	Other multiple birth (three or more) unspecified whether mates liveborn or stillborn	Narrow
Multiple_child	ICD9CM	V37.0	Other multiple birth (three or more) unspecified whether mates liveborn or stillborn born in hospital	Narrow
Multiple_child	ICD9CM	V37.00	Other multiple birth (three or more), unspecified whether mates liveborn or stillborn, born in hospital, delivered without mention of cesarean section	Narrow
Multiple_child	ICD9CM	V37.01	Other multiple birth (three or more), unspecified whether mates liveborn or stillborn, born in hospital, delivered by cesarean section	Narrow
Multiple_child	ICD9CM	V37.1	Other multiple birth (three or more), unspecified whether mates liveborn or stillborn, born before admission to hospital	Narrow
Multiple_child	ICD9CM	V37.2	Other multiple birth (three or more), unspecified whether mates liveborn or stillborn, born outside of hospital	Narrow
Singleton_mother	RCD2 AURUM	Lyu57	Assisted single delivery, unspecified	Narrow
Singleton_child	RCD2 AURUM	^ESCTSI800434	Single liveborn born in hospital by vaginal delivery	Narrow
Multiple_mother	RCD2 AURUM	Lyu56	Other multiple delivery	Narrow
Multiple_mother	RCD2 AURUM	Lyu58	Multiple birth delivery	Narrow
Multiple_mother	RCD2 AURUM	^ESCTTO812505	Total number of registerable births at delivery	Narrow
Multiple_mother	RCD2 AURUM	^ESCTMU510562	Multiple delivery	Narrow

Multiple_mother	RCD2 AURUM	L213	Multiple delivery	Narrow
Multiple_mother	RCD2 AURUM	L2130	Multiple delivery, all spontaneous	Narrow
Multiple_mother	RCD2 AURUM	L2131	Multiple delivery, all by forceps and vacuum extractor	Narrow
Multiple_mother	RCD2 AURUM	L2132	Multiple delivery, all by caesarean section	Narrow
Multiple_mother	RCD2 AURUM	L323	Delayed delivery of second twin, triplet etc	Narrow
Multiple_mother	RCD2 AURUM	L3230	Delayed delivery second twin unspecified	Narrow
Multiple_mother	RCD2 AURUM	L3231	Delayed delivery second twin - delivered	Narrow
Multiple_mother	RCD2 AURUM	L3232	Delayed delivery second twin with antenatal problem	Narrow
Multiple_mother	RCD2 AURUM	L323z	Delayed delivery second twin etc NOS	Narrow
Multiple_child	RCD2 AURUM	ESCTOR 2	Order of birth at delivery	Narrow
Multiple_child	RCD2 AURUM	633D	Order of birth at delivery	Narrow
Singleton_mother	RCD2 GOLD	Lyu5000	Other single spontaneous delivery	Narrow
Singleton_mother	RCD2 GOLD	Lyu5200	Other single delivery by caesarean section	Narrow
Singleton_mother	RCD2 GOLD	Lyu5500	Other specified assisted single delivery	Narrow
Singleton_mother	RCD2 GOLD	Lyu5700	Assisted single delivery, unspecified	Narrow
Multiple_mother	RCD2 GOLD	Lyu5600	Other multiple delivery	Narrow
Multiple_mother	RCD2 GOLD	Lyu5800	Multiple delivery, unspecified	Narrow
Multiple_mother	RCD2 GOLD	Z254E11	Multiple birth delivery	Narrow
Multiple_mother	RCD2 GOLD	L213.00	Multiple delivery	Narrow
Multiple_mother	RCD2 GOLD	L213200	Multiple delivery, all by caesarean section	Narrow
Multiple_mother	RCD2 GOLD	L213100	Multiple delivery, all by forceps and vacuum extractor	Narrow
Multiple_mother	RCD2 GOLD	L213000	Multiple delivery, all spontaneous	Narrow
Multiple_child	RCD2 GOLD	L323000	Delayed delivery second twin unspecified	Narrow
Multiple_child	RCD2 GOLD	L323z00	Delayed delivery second twin etc NOS	Narrow
Multiple_child	RCD2 GOLD	L323.00	Delayed delivery of second twin, triplet etc	Narrow
Multiple_child	RCD2 GOLD	L323100	Delayed delivery second twin - delivered	Narrow
Multiple_child	RCD2 GOLD	L323200	Delayed delivery second twin with antenatal problem	Narrow
Multiple_child	RCD2 GOLD	633D.00	Order of birth at delivery	Narrow
.	ICPC2	W90	Uncomplicate labour/delivery live	possible
.	ICPC2	W91	Uncomplicate labour/delivery still	possible
.	ICPC2	W92	Complicate labour/ delivery livebirth	possible

.	ICPC2	W93	Complicate labour/delivery stillbirth	possible

Chapter 3. Days of Treatment (CreateDoT)

3.1 Introduction

When a pharmacoepidemiology study is conducted using electronic healthcare data sources, person-level exposure to medications is assessed based on the electronic records collecting information on either prescription, dispensing or administration of one or more medicinal products of interest (*Overbeek et al, 2017, Rasmussen et al, 2022, Roberto et al, 2020*). Notably, the information collected in a drug utilization record is usually data source-specific. In fact, while the date of dispensing, prescription or administration is usually available in electronic healthcare data sources, other items concerning the drug utilization event might be incomplete or even not captured in the specific electronic healthcare data source of interest (*ConcePTION Catalogue*). In particular, the prescribed/administered daily dose and/or the treatment duration are often missing or even not recorded in large electronic healthcare data sources. Nevertheless, they can be imputed or calculated based on assumptions that are made upfront by the investigator (*Overbeek et al, 2017, Rasmussen et al, 2022, Meaidi et al, 2021*).

Although it is a well-established assumption that the date of the first drug utilization record of interest is the start of drug use, missing daily dose and/or treatment duration need to be calculated or imputed based on specific assumptions that take into account the information items available in the datasource-specific drug utilization record of interest (e.g. days supplied, prescribed daily dose, medicinal product strenght, number of dosage units), as well as the characteristics of the specific pharmacoepidemiologic study (e.g. indication for drug use in the study population) (*Rasmussen et al, 2022*).

Notably, even when the prescribed/administered daily dose is known and data from one unique healthcare data source are considered, several approaches can be adopted for calculating the duration of treatment to be associated to the same drug utilization records of interest (*Meaidi et al, 2021, Gardarsdottir et al, 2010*). Consequently, in multi-database studies the standardization of such calculations for assigning treatment duration to each drug utilization record of interest becomes fundamental to facilitate both the documentation of study methods and the comparison of results from the different data sources contributing to the study.

3.2 Purpose

The ‘Create Days of Treatment’ function (CreateDoT), which is available on GitHub at this link (*CreateDoT Wiki, 2024*), is meant to be used for standardizing the calculation of days of exposure associated to any type of electronic drug utilization record (i.e. prescription, dispensing or administration of a medicinal product) across different observational healthcare data sources. CreateDoT aims at defining a set of standardized calculations, that take as input a dataset of longitudinal drug utilization records along with other information such as the medicinal product package characteristics and the daily dose prescribed by the physician or assumed by the investigator.

3.3 General consideration and caveat

We use the term «calculation approach» for indicating the algorithm used to calculate the number of days of treatment corresponding to a single drug utilization record. The output of CreateDoT is a new variable containing the number of days of treatment per each single drug utilization record. CreateDoT may also be specified to provide, as output, a calculation of the daily dose, measured as amount of active substance, based on input parameters specified by the user to calculate the treatment duration.

After CreateDoT has created the number of days of treatment for each drug utilization record of interest, it is the responsibility of the user to make further assumptions on whether such days of treatment happen immediately and regularly, or not. The function in itself does not compute a variable indicating the end of continuous treatment episodes that consist in more than one consecutive drug utilization record. For the latter, further assumption are needed to define how to manage overlaps and gaps between the estimated duration of two or more consecutive records of utilization of the medicinal product(s) of interest (*Gardarsdottir et al, 2010*).Therefore, the output of this function needs to be further processed to assess either the duration of continuous treatment episodes, the average daily dose consumed during a continuous treatment episode, adherence or persistence to treatment, trajectories of use, etc. The choices underlying such additional steps are not explored in this document. However, the output of CreateDoT can be used both as an analysis variable itself, or, for instance, as an input to build episodes of continuous treatment using AdhereR.

General examples of the application of CreateDoT using three distinct calculation approaches are reported below (For more detailed examples and function parameters concerning all the five calculation approaches proposed in this document, see section 3.5 Description of calculation approaches, as well as the function wiki (*CreateDoT Wiki, 2024*).

1) Application of the calculation approach "Active substance amount per day"

If a drug dispensing record contains information that 3 packages of medicinal product were dispensed, containing a total of 100mg of active ingredient per box, and the investigator assumed that a daily dose of 5mg/day (i.e. input parameter to be specified by the investigator) was taken, because this is the Defined Daily Dose of the World Health Organization for this ingredient, the function will calculate the DoT of the record as 3x100/5 = 60 days.

2) Application of the calculation approach "Unit of presentations per day"

If a drug prescription record contains information that a patient was dispensed one medicinal product package of 90 tablets, containing 3mg of active substance each, and the prescribed daily dose was 2 tablets per day, the function will calculate the DoT of the record as 1x90/2 = 45 days, with a corresponding Daily Dose (DD) of 6 mg of active substance per day.

3) Application of the calculation approach "Fixed record duration"

If a drug dispensing record contains information that a woman of childbearing age received a medicinal product with teratogenic risk for which a dispensing corresponding to a treatment no longer than 30 days is recommended, the algorithm "Fixed record duration" might be chosen, and the function will associate to the drug utilization record a DoT of 30 days independently from the dispensed number of packages, dosage units or strength of the medicinal product. For instance, if a

*drug dispensing record contains information that 1 package of NEOTIGASON*30 cps 10 mg and 1 package of NEOTIGASON*20 cps 25 mg (acitretin) was dispensed and the investigator assumed a fixed duration of the drug utilization record of 30 days, the function will assign a Dot of 30 days to the drug utilization record and will calculate the corresponding DD of active substance as ((30x100)+(20x25))/30 = 26,7 mg/day.*

The user must specify at least one additional parameter (from the statistical analysis plan), corresponding to the specific calculation approach to be applied and the relevant daily dose or fixed number of days of duration that will be used to estimate the days of treatment.

The calculation approaches may be applied either sequentially (e.g. first apply one, and where input values are missing in the drug utilization record, then apply a second/third/n-th calculation approach) or iteratively to calculate days of treatments for the same set of drug utilization records based on different calculation approaches and/or assumptions.

3.4 Glossary

The definitions listed here are inspired by the document Standard Terms of the European Directorate for the Quality of Medicines (EDQM) available at this link. We also indicate the variables capturing the definitions in the ConcePTION Common Data Model (CDM) and the corresponding items from the ISO standands for the Identification of Medicinal Products (IDMP) (*ISO 11615:2017, ISO 11239:2023, ISO IDMP standards*).

This section described below is part of the wiki of the CreateDoT function (*CreateDoT Wiki, 2024*).

3.4.1 Medicinal product

It is a product of medicinal nature authorized for marketing (see a more sophisticated definition here). It is uniquely identified at package level by a national/international medicine identifier code. Some examples of medicinal product packages marketed in Italy are:

- 'ATENOLOLO AHCL 50CPR 50MG', which includes 50 tablets, each containing 50 mg of atenol, a cardiovascular drug used for treating hypertension and other cardiovascular conditions
- 'DROPTIMOL COLL FL 5ML 2,5MG/ML', which includes one multi-dose container of 5ml of eye drops, where an anti-glaucoma drug, timolol, has a concentration of 2.5 mg/ml
- 'LUMIGAN COLL30FL0,4ML 0,3MG/ML', which includes 30 single-dose containers (units of presentation) of 0.4 ml of anti-glaucoma eye drops each. The concentration of the active substance, bimatoprost, is 0.3 mg/ml
- 'CLOBESOL*CREMA 30G 0,05%', which includes a multi-dose tube, containing 30 g of cream to be applied on the skin. The 0.05% of the total amount of cream contained in the multi-dose tube is a corticosteroid, clobetasol.
- 'LOSARTAN ID ALM*28CPR 100+25MG', which includes 28 oral tablets of an antihypertensive medication. Each tablet contains 100 g of losartan (substance 1) and 25 mg of idrochlorotiazide (substance 2)

A medicinal product package may contain, for instance, a defined number of tablets or capsules, or it may correspond to a solution containing a defined concentration of one or more active substances, for instance a bottle of syrup. A medicinal product package may contain one or more vials of an injectable solution, one multi-dose container of eye drops or even a defined number of single-dose eye drop containers...

In the ConcePTION CDM, the code of the medicinal product is stored in the variable medicinal_product_id which is the linkage key between the MEDICINES and the PRODUCTS tables. In PRODUCTS, the description of the medicinal product in the local language can also be stored, in the variable medicinal_product_name.

The ISO standard corresponding to the ''medicinal_product_id'' of the ConcePTION CDM is the ''Medicinal Product Package Identifier'' (ISO 11615:2017).

3.4.2 Drug utilization record

We use this term to indicate a record of a prescription, dispensing or administration event of one or more packages of one or more medicinal products in electronic healthcare databases.

In the ConcePTION CDM, a drug utilization record corresponds to a record of the MEDICINES tables.

3.4.3 Unit of presentation

Units of presentations allows to identify the countable entity in which the strength(s) of the medicinal product is presented and described: for example tablets, syringes. “Unit of presentation” is a controlled vocabulary of the EDQM. For example, the strength of a modified-release tablet, such as “10 mg per tablet”, is expressed in terms of strength (‘10 mg’) per each unit of presentation (‘per tablet’). Similarly, where the quantity of product in a pre-filled syringe needs to be expressed, for example “10 ml per syringe”, a unit of presentation (‘syringe’) is also used. While a unit of presentation will often share the same name as another concept, such as pharmaceutical dose form (e.g., tablets: see *pharmaceutical dose form* below) or container (e.g., bottles), it is important that a separate list of terms is maintained for units of presentation, pharmaceutical dose forms, and containers. There are other ways of expressing strength/quantity (e.g., as a concentration, using standard units of measurement, such as “0.5 mg/ml” or “5 mg per 100 ml”), see below.

In the PRODUCTS table of the ConcePTION CDM, this is described by the variable unit_of_presentation_type: its vocabulary listed here corresponds to the “unit of presentation” controlled vocabulary of the EDQM. The number of units of presentations included in the medicinal product is stored in the variable unit_of_presentation_num.

The definitions of ''unit of presentation'' used in the ConcePTION CDM is in line with the corresponding ISO standard ''Unit of presentation'' (ISO 11239:2023).

3.4.4 Pharmaceutical dose form and pharmaceutical product

The pharmaceutical dose form is the physical manifestation of a product that contains the active ingredient(s) and/or inactive ingredient(s) that are intended to be delivered to the patient (e.g. cream, capsule, tablet, powder, oral solution). A slightly more sophisticated concept is the pharmaceutical

product (*ISO 11615:2017, ISO 11239:2023, ISO IDMP standards*) which represents the qualitative and quantitative composition of a medicinal product in the pharmaceutical dose form authorized for administration by a medicines regulatory agency and as represented with any corresponding regulated product information, we will not use this concept in a direct manner.

Depending on the specific type of pharmaceutical dose form, the medicine can be administered as it is (e.g., 1 tablet) or as a defined part (e.g., 1 spoon of syrup, 5 ml of cream). In some cases, the pharmaceutical dose form has the same name as the unit of presentation (e.g., tablet, capsule). It must be noted that the amount of active substance usually represents only a portion of the total amount of the pharmaceutical product (see below): 'dicloreum cream 1%' (medicinal product) is a tube containing 100 mg of cream (amount of pharmaceutical product) including 1 mg of diclofenac (amount of active substance).

In the PRODUCTS table of the ConcePTION CDM, the categorical variable representing the pharmaceutical dose form is the variable administration_dose_form, which correspond to the EDQM standard term and ISO standard "basic dose form" (Gardarsdottir et al, 2010, ISO 11615:2017, ISO 11239:2023). The corresponding controlled vocabulary from the EDQM is listed here. The amount of pharmaceutical product correspond to the size of the container of the medicinal product of interest and it is stored in the variable concentration_total_content of the ConcePTION CDM, whose unit of measurement is stored in the variable concentration_total_content_unit. For example: a tube containing 60 mg of cream (concentration_total_content = 60, concentration_total_content_unit = mg), or a bottle containing 100ml of oral solution (concentration_total_content = 100, concentration_total_content_unit = ml). If the medicinal product contains multiple units of presentation, this variable refers to each unit of presentation: e.g., if the medicinal product contains 30 single-dose containers of 1.4 ml, concentration_total_content = 1.4 and concentration_total_content_unit = ml.

3.4.5 Amount of active substance and concentration of active substance

The quantity of active substance in one unit of presentation may be stored in the data in two different manners

- as a direct measurement (e.g. tablet containing 10mg ramipril, suppository containing 500mg of acetaminophen);
- as a fraction/ratio of the pharmaceutical product
    - either the mass/volume of active substance as a fraction of the mass/volume of the pharmaceutical product in a unit of presentation (e.g., 30g of cream containg 0,05% of active principle, that is, 1.5mg of active principle), or
    - as a concentration, i.e. the ratio between the mass of active substance and the volume of pharmaceutical poduct in a unit of presentation (e.g., a contained of 5ml of eye drops, with concentration of 2.5mg/ml).

In the PRODUCTS table of the ConcePTION CDM, the amount of active substance can be described by two sets of variables, according to whether it is available as a direct measurement or as a concentration

- *a direct measurement is stored in the variable subst1_amount_per_form, whose unit of measurement is stored in subst1_amount_unit. If the medicinal product contains multiple active principles, their amount is stored also in subst2_amount_per_form/subst2_amount_unit and if necessary subst3_amount_per_form/subst3_amount_unit.*
- *a concentration is stored in subst1_concentration, whose unit is stored in subst1_concentration_unit (e.g, mg/ml). If the medicinal product contains multiple active principles, their concentration is stored also in subst2_concentration/subst2_concentration_unit and if necessary subst3_concentration/subst3_concentration_unit.*

3.5 Description of calculation approaches

The calculation approaches described below can be distinguished in two main families:

1) Daily Dose-based calculation approaches, where the investigator must specify the daily dose (DD) and its unit of measure as the input parameter of the function, and
2) Fixed duration-based calculation approaches, where the investigator must define a fixed number of days to be associated to either each unit of medicinal product package or each drug utilization record (i.e. regardless of the corresponding number of medicinal product packages).

3.5.1 DD-based calculation approaches

When using this family of calculation approaches, the investigator must choose a specific daily dose (DD). There are three possible DD-based calculation approaches which correspond to the three units of measure the investigator can choose to quantify the DD. A DD can be expressed as:

- a defined number of units of presentation (e.g., 1 tablet/day of atorvastatin), or
- a defined amount of active substance (e.g., 150 mg/day of metoprolol), or
- a defined amount of pharmaceutical product form from a multi-dose container (e.g., 10 ml/day of cough syrup).

Notably, the investigator may choose as the DD the actual **prescribed daily dose**, whenever available in the MEDICINES table. Another common option is to use the **Defined Daily Dose assigned by the WHO**. Other assumptions on the DD are possible and can be based on many different elements: the target population of users, the expected indication of use, and/or the information from the Summary of Product Characteristics (SmPCs).

The three calculation approaches are described below in detail, including examples.

3.5.1.2 Units of presentations per day

This calculation approach can be used when the chosen DD correspond to one or more units of presentation of the medicinal product package of interest (e.g., one or more tablets). It allows calculating the days of treatment of one single drug utilization record based on the total number of

units of presentation contained in the medicinal product package(s) that the record contains, divided by the number of units of presentation to be taken daily according to the assumed DD (e.g., 2 tablets per day).

The calculation approach also calculates the CALCULATED_DD_active_subst, i.e., the amount of active substance corresponding to the amount of pharmaceutical product chosen as DD.

The calculation approach can be used only if the DD chosen by the investigator is homogeneous with respect to the unit of presentation that describes the medicinal product of interest (e.g., both DD and unit of presentation are described as «tablets»). In case this rule does not hold, one of the two other DD-based calculation approaches might be used.

Using the calculation approach « Unit of presentation per day » the DoT is calculated as:

DoT=(number of packages of medicinal product)×(number of units of presentation contained per medicinal product package)/DD

The CALCULATED_DD_active_subst is obtained as follow:

CALCULATED DD active subst=(number of packages of medicinal product)×(total active substance amount per medicinal product)/DoT

The calculation of “total active substance amount per medicinal product” is described in Section 3.6.

Examples of the application of the calculation approach Units of presentations per day

- The medicinal product labelled "ELIQUIS*20CPR RIV 2,5MG" in Italy contains 20 tablets (unit of presentation), each containing 2.5 mg of apixaban, a direct anticoagulant. Other medicinal products containing apixaban may have different strengths (2.5mg or 5mg). The recommended posology is 2 tablets per day for apixaban, independently from the specific strength. Therefore, a DD of 2 tablets/day may be chosen by the investigator. If so, the DoT for one package of this medicinal product is equal to 10 days, i.e., (1 package of medicinal product) x (20 tablets) / 2 tablets/day. The CALCULATED_DD_active_subst will be 5 mg/day, i.e., (1 package of medicinal product) x 50mg / 10 days

- The medicinal product labelled 'LUMIGAN*COLL30FL0,4ML 0,3MG/ML' in Italy contains anti-glaucoma eye drops, in multiple single-dose containers. It includes 30 containers (units of presentation) of 1.4 ml, each including a concentration of active substance of 0.3mg/ml of bimatoprost. One single-dose container per day is recommended, therefore, a DD of 1 single-dose container may be chosen by the investigator: DD = 1 container/day. If so, the DoT for 2 packages of medicinal product will be equal to 60 days, i.e., (2 packages of medicinal product) x (30 containers) / 1 container/day. The CALCULATED DD will be 1.2 mg/day, i.e., 1 package of medicinal product x 36 mg / 30 days

The R code developed to implement the calculation approach “Unit of presentation per day” according to the variables of the ConcePTION CDM is reported below. Additional details are available in the GitHub repository of the function (*CreateDoT Wiki, 2024*).

```
output <- CreateDOT(dataframe = input,
                    calculation_approach = "Units of presentations per day",
                    output_var = "CALCULATED_DoT",
                    disp_num_medicinal_product = "disp_num_medicinal_product",
                    unit_of_presentation_num = "unit_of_presentation_num",
                    subst1_amount_per_form = "subst1_amount_per_form",
                    subst1_amount_per_form_unit = "subst1_amount_unit",
                    subst1_concentration = "subst1_concentration",
                    subst1_concentration_unit = "subst1_concentration_unit",
                    concentration_total_content= "concentration_total_content",
                    dd="dd",
                    dd_unit = "unit_dd",
                    unit_of_presentation="unit_of_presentation_type",
                    output_dd1="CALCULATE_DD_subst1",
                    output_dd1_unit="CALCULATE_DD_subst1_unit"
                    )
```

3.5.1.2 Active substance amount per day

This calculation approach can be used when the chosen DD correspond to an amount of active substance contained in the medicinal product of interest.

It allows calculating the days of treatment of a drug utilization record based on the total amount of active substance contained in the corresponding medicinal product package(s) divided by the active substance amount to be taken daily according to the chosen DD (e.g., 10 mg per day).

It can be used only if the unit of measure of the chosen DD (e.g., 10 mg) is equal to the unit of measure that describes the total active substance amount in the medicinal product of interest (e.g., 100 mg). If the unit of measure of the DD is a multiple or submultiple of the unit of measure of the total active substance amount (e.g., mg and g) then conversion should be done before starting the calculation. In case the units of measure are not homogeneous (e.g., mg and ml), one of the two other DD-based calculation approaches might be used.

The DoT is calculated as:

DoT=(number of packages of medicinal products)×(total active substance amount per package of medicinal product)/DD

The calculation of "total active substance amount per medicinal product" is described in the Appendix available online in the function wiki (*CreateDoT Wiki, Appendix. 2024*) and in Section 3.6.

Examples of calculation approach Active substance amount per day

In both examples, the relevant DDDs assigned by the WHO, which correspond to a specific amount of active substance, are used as the assumed DD.

- The medicinal product labelled 'ATENOLOLO AHCL*50CPR 50MG' in Italy contains 50 tablets, each containing 50 mg of atenolol, a cardiovascular drug used for treating hypertension and other cardiovascular conditions. According to the WHO, the DDD of atenol is 75mg/day.

The total active substance amount contained in one medicinal product package is calculated by multiplying the active substance amount contained in one tablet by the total number of tablets contained in one package of medicinal product, i.e. 50 x 50mg = 2500mg. Then, if the investigator chooses as DD the DDD assigned by the WHO, the DoT for one unit of this medicinal product package will be equal to: (1 package of medicinal product) x (2500 mg) / 75 mg/day = 33.33 days.

- The medicinal product labelled 'DEPAKIN*OS FL 40ML 200MG/ML' in Italy contains one multi-dose container of 40ml of oral solution (i.e., a liquid), where an antiepileptic drug, valproate, has a concentration of 200mg/ml. According to the WHO, the DDD of valproate is 1.5g/day = 1500mg/day. The total active substance amount contained in the medicinal product package is calculated multiplying the concentration of the active substance amount in the oral solution by the total amount of oral solution contained in one package of medicinal product, i.e. 200 mg/ml x 40 ml = 8000 mg. Then, if the investigator chooses as DD the DDD assigned by the WHO, the DoT for 2 units of this medicinal product package will be equal to: (2 packages of medicinal product) x (8000 mg) / (1500 mg/day) = 10.66 days.

  The R code developed to implement the calculation approach “Activee substance amount per day” according to the variables of the ConcePTION CDM is reported below. Additional details are available in the GitHub repository of the function (*CreateDoT Wiki, 2024*).”.

```
output <- CreateDOT(dataframe = input,
                    calculation_approach = "Active substance amount per day",
                    output_var = "CALCULATED_DoT",
                    disp_num_medicinal_product = "disp_number_medicinal_product",
                    unit_of_presentation_num = "unit_of_presentation_num",
                    subst1_amount_per_form = "subst1_amount_per_form",
                    subst1_amount_per_form_unit = "subst1_amount_unit",
                    subst1_concentration= "subst1_concentration",
                    concentration_total_content= "concentration_total_content",
                    subst1_concentration_unit= "subst1_concentration_unit",
                    concentration_total_content_unit= "concentration_total_content_unit",
                    dd = "dd",
                    dd_unit = "unit_dd"
                    )
```

3.5.1.3 Amount of pharmaceutical product per day

This calculation approach allows calculating DoT of a drug utilization record dividing the total amount of pharmaceutical product contained in the relevant medicinal product package(s) (e.g., 100 mg of cream) by the amount of pharmaceutical product to be taken daily according to the chosen DD (e.g. 5 mg of cream/day).

The calculation approach also allows calculating the CALCULATED_DD_active_subst, i.e. the amount of active substance corresponding to the amount of pharmaceutical product chosen as DD. It can be applied only if the unit of measure of the chosen DD (e.g., 2 mg of cream) is equal to the unit of measure that describes the total amount of pharmaceutical product contained in the medicinal product package of interest (e.g., 100mg). If the unit of measure of the chosen DD is a multiple or submultiple of the unit of measure of the total amount of pharmaceutical product (e.g., mg and g)

conversion should be done before starting the calculation. In case the units o measure are not homogeneous, one of the two other DD-based calculation approaches might be used. The difference with the previous calculation approach is that here the DD is not a portion of the total amount of active substance contained in the medicinal product (e.g., the active substance included in the cream), but rather it corresponds to a portion of the total amount of pharmaceutical product (the cream itself).

The DoT is calculated as:

DoT=(number of packages of medicinal products)×(total amount of pharmaceutical product contained in the medicinal product package)/DD

The CALCULATED_DD_active_subst is obtained as follow:

CALCULATED_DD_active_subst=(number of packages of medicinal product)×(total active substance amount per medicinal product)/DoT

The calculation of "total active substance amount per medicinal product" is described in the Appendix available online (*CreateDoT Wiki Appendix, 2024*) and in Section 3.6.

Examples of calculation approach Amount of pharmaceutical product per day

- The medicinal product labelled 'CLOBESOL*CREMA 30G 0,05%' in Italy is a multi-dose tube, containing 30 g of cream to be applied on the skin. The active substance, a corticosteroid, correspond to 0,05% of the total pharmaceutical dose form amount (i.e., 30g x 0.05% = 15mg). If the investigator assumes that every day the patient uses 2 grams of cream (i.e., DD= 2g/day), DoT for one unit of the medicinal product package containing Clobesol will be equal to: (1 package of medicinal product) x 30g / (2g/day) = 15 days. The CALCULATED_DD_active_subst will be 1 mg/day, i.e. (1 package of medicinal product) x 15mg / 15 days

- The medicinal product labelled 'DROPTIMOL*COLL FL 5ML 2,5MG/ML' in Italy contains one multi-dose container of 5ml of eye drops where an antiglaucoma drug, timolol, has a concentration of 2.5mg/ml. The total active substance amount in one package of medicinal product is 5ml x 2.5mg/ml = 12.5mg. According to the recommendation of the WHO for establishing a DDD, two eye drops (one in each eye) correspond to 0.1 ml. If the investigator assumes that the medicinal product is administered twice daily then the assumed DD will be 0.2 ml/day. Therefore, DoT for 2 medicinal product package units of Droptimol will be equal to: (2 packages of medicinal product) x 5ml / (0.2 ml/day) = 50 days. The CALCULATED_DD_active_subst will be 0.5mg/day, i.e., (2 packages of medicinal product) x 12.5mg / 50 days.

  The R code developed to implement the calculation approach "Amount of pharmaceutical product per day" according to the variables of the ConcePTION CDM is reported below. Additional details are available in the GitHub repository of the function (*CreateDoT Wiki, 2024*).CreateDOT".

```
output <- CreateDOT(dataframe = input,
                    calculation_approach = "Amount of pharmaceutical product per day",
                    output_var = "DOT_recipe_3",
                    output_dd1="CALCULATE_DD_subst1",
                    output_dd1_unit="CALCULATE_DD_subst1_unit",
                    disp_num_medicinal_product = "disp_number_medicinal_product",
                    concentration_total_content= "concentration_total_content",
                    concentration_total_content_unit= "concentration_total_content_unit",
                    subst1_concentration= "subst1_concentration",
                    subst1_concentration_unit= "subst1_concentration_unit",
                    dd = "dd",
                    dd_unit = "unit_dd"
                    )
```

3.5.2 Fixed duration-based calculation approaches

The calculation of DoT relies on a fixed duration that the investigator may choose to assign either to each drug utilization record or to each package of medicinal product contained in the drug utilization record of interest, irrespective from the number of dosage units, amount of active substance or amount of pharmaceutical product contained in the medicinal product package unit(s) of interest.

The fixed duration-based calculation approaches calculation approaches also generate the CALCULATED_DD_active_subst, i.e. the amount of active substance that was taken daily by the patients according to the a priori assumed duration of the relevant drug utilization record of interest.

3.5.2.1 Fixed record duration

This calculation approach can be applied when the investigator chooses to assign a fixed duration to each drug utilization record. The DoT of each drug utilization record of interest will be equal to the duration assumed *a priori* (e.g. 30 days), irrespective from everything else (the number of units of presentations, the strength, the amount of pharmaceutical product contained in each medicinal product package, and even the number of packages of medicinal product in the drug utilization record). This calculation approach may be chosen whenever the daily dose varies significantly across patients and indications of use, while a fixed duration of each drug utilization record can be assumed based on the expected prescribing behaviours and/or recommendations.

DoT=(fixed duration of the drug utilization record)

The CALCULATED_DD_active_subst is obtained as follows:

CALCULATED DD active subst=(number of packages of medicinal product)×(total active substance amount per medicinal product)DoT

The calculation of “total active substance amount per medicinal product” is described online (*CreateDoT Wiki Appendix, 2024*) and in Section 3.6

Examples of calculation approach base on ‘’Fixed record duration’’

Consider two medicinal product packages, labelled respectively 'NEOTIGASON 30CPS 10MG' and 'NEOTIGASON 20CPS 25MG'. The former contains 30 tablets of 10mg, the latter 20 tablets of 25mg of the same antipsoriatic drug, acitretin. The total active substance amount contained in one package

of medicinal product is 30 x 10mg = 300mg in the first example, and 20 x 25mg = 500mg in the second example. In a study were DoT has to be estimated in women of childbearing age it would be reasonable to assume a fixed duration, since this product has a teratogenic potential and its use in this population must be strictly controlled: the recommendation is that prescription of acitretin-based therapy should not exceed 30 days of treatment. Therefore, DoT = 30 days, in both examples. The CALCULATED_DD_active_subst for one drug utilization record containing 1 unit of the first medicinal product package will be 10 mg, i.e. (1 package of medicinal product) x 300mg / (30 days). The CALCULATED_DD_active_subst for one drug utilization record containing 3 units of the second medicinal product will be 25mg/day, i.e. (3 packages of medicinal product) x 500mg/ 30 days.

The R code developed to implement the calculation approach “Fixed record duration” according to the variables of the ConcePTION CDM is reported below. Additional details are available in the GitHub repository of the function (*CreateDoT Wiki, 2024*).CreateDOT”.

```
output <- CreateDOT(dataframe = input,
                    calculation_approach = "Fixed record duration",
                    output_var = "CALCULATED_DOT",
                    output_dd1 = "CALCULATE_DD_subst1",
                    output_dd1_unit="CALCULATE_DD_subst1_unit",
                    fixed_duration_days="fixed_duration_days",
                    disp_num_medicinal_product = "disp_number_medicinal_product",
                    unit_of_presentation_num = "unit_of_presentation_num",
                    subst1_amount_per_form = "subst1_amount_per_form",
                    subst1_amount_per_form_unit = "subst1_amount_unit"

                    )
```

3.5.2.2 Fixed medicinal product package duration

This calculation approach can be applied when the investigator chooses to assign a fixed duration (e.g., 30 days) to each package of medicinal product contained in the relevant drug utilization record of interest, irrespective from the number of unit of presentations, the strength or the amount of pharmaceutical product contained in each package of medicinal product.

This calculation approach may be chosen when the number of dosage units, the amount of active substance and the amount of pharmaceutical product per day vary significantly across patients and indications of drug use while a fixed duration of each unit of the medicinal product of interest can be assumed based on recommended dosing schedule.

DoT will be equal to the chosen value for the fixed duration of the medicinal product, multiplied by the number of units of the medicinal product of interest contained in the drug utilization record.

DoT=number of packages of medicinal products)×(fixed duration of the medicinal product)

The CALCULATED_DD_active_subst is obtained as follow:

CALCULATED DD active subst=(number of packages of medicinal product)×(total active substance amount per medicinal product)DoT

The calculation of "total active substance amount per medicinal product" is described in the Appendix.

Examples of calculation approach based on "Fixed medicinal product duration"

- The medicinal product package labelled 'LOSARTAN ID ALM*28CPR 100+25MG' in Italy contains 28 oral tablets of an antihypertensive medication. Each tablet contains 100g of losartan (substance 1) and 25 mg of idrochlorotiazide (substance 2), to be administered once daily. The total amount of substance 1 (i.e., losartan) is 28 x 100mg = 2800 mg, and the total amount of substance 2 (i.e., hydrochlorothiazide) is 28 x 25mg = 700mg. The investigator may choose to approximate the duration of one package of medicinal product 30 day to account, for instance, for imperfect treatment adherence. Therefore, DoT for a drug utilization record containing 3 units of this medicinal product will be equal to 90 days, i.e. (3 packages of medicinal product) X 30 days. The CALCULATED DD for substance 1 (i.e., losartan) will be 93.33 mg/day, i.e. (3 packages of medicinal product) x 2800mg/90 days. The CALCULATED_DD_active_subst for substance 2 (i.e., hydrochlorothiazide) will be 23.33mg/day, i.e. (3 packages of medicinal product) x 700mg/90 days.

  The R code developed to implement the calculation approach "Fixed medicinal product package duration" according to the variables of the ConcePTION CDM is reported below. Additional details are available in the GitHub repositoryof the function (*CreateDoT Wiki, 2024*).

```
output <- CreateDOT(dataframe = input,
                    calculation_approach = "Fixed medical product package duration",
                    output_var = "CALCULATED_DOT",
                    output_dd1 = "CALCULATE_DD_subst1",
                    output_dd1_unit="CALCULATE_DD_subst1_unit",
                    output_dd2 = "CALCULATE_DD_subst2",
                    output_dd2_unit="CALCULATE_DD_subst2_unit",
                    fixed_duration_days="fixed_duration_days",
                    disp_num_medicinal_product = "disp_number_medicinal_product",
                    unit_of_presentation_num = "unit_of_presentation_num",
                    subst1_amount_per_form = "subst1_amount_per_form",
                    subst1_amount_per_form_unit = "subst1_amount_unit",
                    subst2_amount_per_form = "subst2_amount_per_form",
                    subst2_amount_per_form_unit = "subst2_amount_unit"

)
```

3.6 Computing the total active substance amount per medicinal product

To calculate the "total active substance amount per medicinal product", two different sets of parameters can be used depending on the available descriptors of the medicinal product package of interest:

- if the "active substance amount per unit of presentation" is available, the total amount of active substance contained in the medicinal product of interest is obtained multiplying the active substance amount contained in one unit of presentation by the total number of unit of presentations contained in one unit of the medicinal product of interest; *In the ConcePTION CDM, the total substance amount is obtained as follows:* (subst_amount1_per_form) X (unit_of_presentation_num)

- if the "concentration of the active substance" contained in the pharmaceutical form of the medicinal product is available, the total amount of active substance in the medicinal product package of interest is obtained multiplying the concentration of the active substance contained in the pharmaceutical dose form of the medicinal product by the total amount of pharmaceutical product contained in the medicinal product. If the medicinal product contains more than one unit of presentation (e.g., 30 dispensers of drops) the total amount of pharmaceutical product contained in the medicinal product can be obtain multiplying the amount of pharmaceutical product contained in one unit of presentation by the total number of units of presentations. *In the ConcePTION CDM, the total substance amount per medicinal product is obtained as follows* (Subst1_concentration) X (concentration_total_content) X (unit_of_presentation_num)

3.7 Structure of input data

Input data consists of a data.table (R format) data set containing entries of prescriptions, dispensings or administration at the individual record level for a selection of the variables (parameters) listed in the table below. For example, the data.table table may be derived from the combination of information from multiple tables (for example, formatted according to the ConcepTION Common Data Model, or according to a different common data model). The function assumes that every drug utilization record represents one prescription, dispensing or administration event. If not already the case, prior to applying the function, the input data must be pre-processed so that there is one row per drug utilization record.

Examples for the application of the CreateDoT function to different types of medicinal products and calculation approaches are available at this link (*CreateDoT Wiki, 2024*).

3.8 Utilization of the CreateDoT function in other projects

The createDoT function is going to be used within the project "*A framework for the post-authorisation safety monitoring and evaluation of vaccines in the EU"* (Reopening of competition no. 18 under framework contract following procurement procedure EMA/2020/46/TDA (Lot 5) and also in the EMMA project (Exposure to Medications Measured with ATC/DDD Classification System) among recommendations for the design and creation of an on-line freely available application for the certified calculation of number of Defined Daily Doses per medicinal product package.

3.9 References

(ConcePTION Catalogue) ConcePTION data catalogue. Available on:

https://vac4eu.molgeniscloud.org/conception/catalogue/#/ Accessed August 2024.

(CreateDoT Wiki, 2024) https://github.com/IMI-ConcePTION/CreateDoT/wiki. Accessed August 2024.

(CreateDoT Wiki, Appendix.2024) https://github.com/IMI-ConcePTION/CreateDoT/wiki/Appendix:-computing-the-total-active-substance-amount-per-medicinal-product. Accessed August 2024.

(Gardarsdottir et al, 2010) Gardarsdottir H, Souverein PC, Egberts TCG, Heerdink ER. Construction of drug treatment episodes from drug-dispensing histories is influenced by the gap length. J Clin Epidemiol. Apr 2010;63(4):422–7

(ISO 11615:2917) ISO 11615:2017(en), Health informatics — Identification of medicinal products — Data elements and structures for the unique identification and exchange of regulated medicinal product information [Internet]. Available on: https://www.iso.org/obp/ui/en/#iso:std:iso:11615:ed-2:v1:en

(ISO 11239:2023) ISO 11239:2023(en), Health informatics — Identification of medicinal products — Data elements and structures for the unique identification and exchange of regulated information on pharmaceutical dose forms, units of presentation, routes of administration and packaging [Internet]. Available on: https://www.iso.org/obp/ui/en/#iso:std:iso:11239:ed-2:v1:en:term:3.1.4

(ISO IDMP standards) Data on medicines (ISO IDMP standards): Overview | European Medicines Agency [Internet]. Available on: https://www.ema.europa.eu/en/human-regulatory-overview/research-development/data-medicines-iso-idmp-standards-overview#ema-inpage-item-18327

(Meaidi et al, 2021) Meaidi M, Støvring H, Rostgaard K, Torp-Pedersen C, Kragholm KH, Andersen M, et al. Pharmacoepidemiological methods for computing the duration of pharmacological prescriptions using secondary data sources. Eur J Clin Pharmacol. Dec 2021;77(12):1805–14

(Overbeek et al, 2017) Overbeek JA, Heintjes EM, Prieto-Alhambra D, Blin P, Lassalle R, Hall GC, et al. Type 2 Diabetes Mellitus Treatment Patterns Across Europe: A Population-based Multi-database Study. Clin Ther. aprile 2017;39(4):759–70

(Rasmussen et al, 2022) Rasmussen L, Wettermark B, Steinke D, Pottegård A. Core concepts in pharmacoepidemiology: Measures of drug utilization based on individual-level drug dispensing data. Pharmacoepidemiol Drug Saf. Oct 2022;31(10):1015–26

(Roberto et al, 2020) Roberto G, Spini A, Bartolini C, Moscatelli V, Barchielli A, Paoletti D, et al. Real word evidence on rituximab utilization: Combining administrative and hospital-pharmacy data. PLoS ONE. 12 marzo 2020;15(3):e0229973

Chapter 4. Breastfeeding

4.1 Introduction

This document complements the protocol (*Jordan et al, 2024*) and Statistical Analysis Plan (SAP) (*Jordan et al., 2022*) of ConcePTION Work Package 1 Demonstration Study 2 (DP2). The background pharmacoepidemiology and glossary of terms is available elsewhere (*Jordan et al, 2022a, Jordan et al, 2023*)

Infant feeding is a public health issue. However, whole-population data collection of breastfeeding data is not universal, particularly where public health data collection is not embedded in practice: there are no legal obligations to register infant feeding strategies, and there are no 'breastfeeding registers' to complement birth registers. Much routine health and social care data are based on administration, transactions and reimbursements, for example for prescription medicines or hospitalization. However, breastfeeding is rarely and unpredictably associated with reimbursable healthcare claims. Therefore, for breastfeeding data, ConcePTION relies on public health databases, where data are systematically reported, which are maintained for the public good from altruistic motives. Where reporting is incomplete, selection bias will be explored and reported as a limitation for each study (*Jordan et al 2013*). Although more databanks are beginning to collect breastfeeding data (*Jordan et al. 2022a, Jordan et al. 2023*), only three have substantial records available for the timeframe of this study.

In the present deliverable, we describe the information available on breastfeeding, and how this is mapped to the ConcePTION common data model. In the ConcePTION Project, use of this information is piloted in DP2. However, breastfeeding data, as described here, will be available for use in future studies.

DP2 will focus on breastfeeding at 4-8 weeks of age. Selection of this timepoint was predicated on concerns that: a) 'breastfeeding at birth' may be only transient, rendering data unduly vulnerable to social desirability response bias (*McAndrews et al., 2012*), particularly when the hospital's WHO 'Baby Friendly Initiative' status is at stake; and b) breastfeeding status in later infancy is influenced by mothers' return to work, and other social, non-biological factors (*Whitney et al., 2023*).

To facilitate the aims of task 1.3.7, as set out in the protocol (*Jordan et al, 2024*), we shall:

- Develop definitions and validate proposed algorithms to identify outcomes, exposures and confounders of interest;
- Produce background and disease-specific prevalence rates for breastfeeding at 4-8 weeks and at birth;
- Develop the criteria for determining which DAPS have data suitable for analysis of breast-feeding;
- Provide recommendations on any specific analyses relating to breastfeeding as an outcome, predictor, confounding or mediator variable.

Given the multifaceted components of breastfeeding success (*Jordan et al., 2022a*), it will be necessary to explore several covariates before any associations can be quantified.

4.2 Data sources

The association between breastfeeding and neurodevelopment (*Kramer et al 2008, Ghozy et al 2020*) obliges investigators to include breastfeeding as a covariate in any analysis of neurodevelopmental outcomes; accordingly, these parameters must be explored together. The data sources available for this protocol as of January 2021 are summarised in **Table 4.1**, reproduced from Jordan et al 2022, under cc licence. Information was collated in January 2021.

Data sources were identified by contacting representatives of all countries in Europe and searching the literature to compile the FAIR Data Catalogue for the ConcePTION project, as described. To identify data sources containing both these variables plus prescription records during pregnancy, the breastfeeding and neurodevelopmental data source lists were cross-referenced and data were discussed with the data access providers.

Data from three Data Access Partners (DAPs) have been further analysed for inclusion in pharmaco-epidemiological studies: Wales, France (Haute-Garone) and Italy (Tuscany). Data managers from University of Dundee, Scotland withdrew from the ConcePTION consortium. Data from Finland starts from 2019 and is therefore largely outside the timeframe of ConcePTION.

- In Wales, birth registrations and data relating to child health are collected in the National Community Child Health Database (NCCHD) and Maternal indicators (MIDS). Breastfeeding data are recorded by midwives at birth and collected by health visitors at 10 days, 4-8 weeks and 6 months. Health visitor contact is mandatory.
- In Haute-Garonne, health certificates are administered to children at 8 days, 9 months and 24 months from the paediatrician or the GP. Information on duration of breastfeeding is sought.
- In Tuscany, information is collected in hospital during hospitalization for childbirth and by primary care professionals working in community services. These services are open to all, but attendance is voluntary: accordingly, there is a risk that data might be vulnerable to volunteer and collider biases (*Jordan et al 2022*). They will therefore not used in DP2.

4.3 Study Population of DP2

All infants surviving to time point of data collection - 4-8 weeks, and excluding those:

- not included in the relevant national/ regional birth register as a live birth (e.g. moved into the country after birth, adopted infants);
- where mother-infant linkage is not available;
- not linked to maternal prescription data throughout pregnancy and up to 8 weeks after childbirth.

Table 4.1. European Population-based Data Sources with Data on Breastfeeding plus Medicines use during pregnancy plus Neurodevelopment

Country	Data sources (breastfeeding data sources italicised)	Neurodevelopmental measurement available	Breastfeeding information categories as they appear in the data source	Pregnancies per year (1,000s)	Birth years with breastfeeding plus neurodevelopment data
Finland	Care Register for Health Care, Primary Health Care, Drugs and Pregnancy Database, *Finnish Medical Birth Register*, CA registry	ICD codes recorded in outpatient or GP care	Assessed and recorded by midwives at discharge or 7 days *postpartum*. Categories: exclusive breastfeeding, partial breastfeeding, 'artificial milk' only.	50	2017 onwards
France (Haute-Garonne)	EFEMERIS* *database (pregnant women and their children up to 24 months)*	Certificates completed at 9 and 24 months by a general practitioner or a pediatrician – include 14 items designed to detect children at risk of psychomotor development abnormalities	Self-report, recorded on health certificates completed during mandatory medical examinations at 8 days, 9 months and 24 months. Categories: 'any' breastfeeding (Yes/No), duration of breastfeeding (in weeks), and duration of exclusive breastfeeding (weeks)	10	mid 2004 onwards
	POMME *databases (breastfeeding data up to 24 months)*	As above plus medicines and health care reimbursements	As above	18.5	mid 2010 to mid 2011 + mid 2015 to mid 2016
Italy – Tuscany	Mental health services, *birth registry*, medicines dispensed in community pharmacies, and hospital pharmacies for outpatient use	Outpatient and mental health service ICD codes	Hospital records documenting how the new-born was fed during the hospital stay. Categories: Only breast milk, breast milk with the addition of water or liquids other than milk, breast milk and infant formula, infant formula only.	30	2010-
UK-Scotland	*Child Health Systems, Programme – Pre-School, Child Health Systems Programme – School,* Support Needs	Children registered on the Support Needs System, Child health developmental	Health visitors' records of self-report at 10 days, 6 weeks and 13 months.	53	2013-

	System, Maternity hospital discharge records (including delivery records), Prescribing Information System	examinations	Categories: breast milk only, fed formula milk only, or fed both breast and formula milk		
UK – Wales	In-patient and out-patient records, Primary Care GP data†, *National Community Child Health Database,* National Pupil Database Wales, congenital anomaly registry	ICD/Read codes, child health developmental examinations, special education needs, and educational attainment from 7 to 16 years	Health visitors' records of self-report: At birth and 6-8 weeks. At 6 and 12 months Categories: 'any' breastfeeding (yes/no)	33	2005-2015-

Notes to table:
ICD: international classification of disease, as issued by the World Health Organisation (WHO).
We use 'neurodevelopment' as an umbrella term for cognitive, social, motor, and behavioural development. How these data can be usefully combined and standardised is being investigated.
Exclusive breastfeeding is as defined by the WHO (2008): Infant receives only breast milk from his/her mother or a wet nurse, or expressed breast milk via tube, cup or syringe, and no other liquids or solids, with the exception of drops or syrups consisting of vitamins, mineral supplements or medicine (WHO 2008). Where breastfeeding is self-reported at certain time-points, the duration of 'breastmilk only' or 'any breastfeeding' is taken as 'from birth'. We acknowledge this may introduce imprecision.
**EFEMERIS covers the 80% of the population covered by the state-controlled French Health Insurance in Haute-Garonne (Lacroix et al 2009).*
†In Wales, ~80% primary care providers voluntarily supply medicines data to the databank. Any selection bias is due to healthcare providers, not subjects. All pregnancies identified can be followed for life, unless the individual leaves the country.
Papers relating medicines use to breastfeeding are available for France (Lacroix et al 2009) and Wales (Jordan et al 2019, Davies et al 2020).

4.4 Data retrieval

Breastfeeding at birth is documented routinely by midwives in primary, secondary or tertiary care, as a component of the birth records.

For data at 4-8 weeks, investigators explored the possibility of using primary care data as a strategy to obtain breastfeeding information. Primary care codes (Read v2) including the term 'breastfe*' were retrieved and reviewed. In the main, these noted problems or support with breastfeeding, rather than practice. Relating these to infant feeding practice or age was not straightforward. Accordingly, this strategy was not pursued, and is not recommended.

Relatively few dyads are admitted to hospital around 4-8 weeks, and rarely for breastfeeding problems. Therefore, we do not recommend using the ICD10 code P92.5 (neonatal difficulty in feeding at breast) or codes relating to mastitis to obtain population wide data on breastfeeding. We did not identify other ICD10 codes relating to breastfeeding. Accordingly, we discounted *post-partum* hospital admissions data as a source of breastfeeding information.

Categorical variables coding infant feeding status were identified in the data sources contributing to ConcePTION. Such variables are loaded to the ConcePTION CDM table named SURVEY_OBSERVATIONS. In DP2, script was programmed in R, to retrieve such data from the CDM, as described in **Table 4.2**

4.4.1 Breastfeeding variables

These will be analysed separately, and prioritised:

1. Breastfeeding at 4-8 weeks
2. Breastfeeding at birth

4.4.1.1 Wales

In the original extraction variables BREASTFEED_BIRTH_FLG and variable BREASTFEED_8_WKS_FLG data entry is coded as follow:

1	Exclusive Milk
2	Combined Milk Feeding - Predominantly Breast
3	Combined Milk Feeding - Partially Breast
4	Artificial Milk Feeding
	Yes any of 1-3
	No none
No entry	Missing data

The variables were converted to binary formats during the ETL process. Values "1", "2", "3" and "Yes any of 1-3" were collapse to value "1", values "4" and "No none" to "0" and missing data remained "NA"

Table 4.2. Summary of the variables to retrieve information on breastfeeding

Data source	CDM table	Description of the origin of information	Rule to select record	Dictionary	Unit of observation
SAIL Databank	SURVEY_OBSERVATIONS	Breastfeeding at birth	so_source_table =NCCHD so_source_column =BREASTFEED_BIRTH_FLG	so_source_value: “0”,”1” , NA	Child
	SURVEY_OBSERVATIONS	Breastfeeding at 8 weeks	so_source_column =BREASTFEED_8_WKS_FLG	“0”,”1”	Child
EFEMERIS	SURVEY_OBSERVATIONS	Breastfeeding at 8 days	so_source_column =J8_ALLAITEMENT	so_source_value: “0”,”1”	Child
	SURVEY_OBSERVATIONS	Breastfeeding at 9 months	so_source_column =M9_ALLAITEMENT	so_source_value: “0”,”1”	Child
	SURVEY_OBSERVATIONS	Breastfeeding duration in weeks at 9 months	so_source_column =M9_DUREE_ALLAIT_SEIN	so_source_value: numeric variable describing duration in weeks	Child
	SURVEY_OBSERVATIONS	Breastfeeding duration in weeks of exclusively breastfed baby at 9 months	so_source_column =M9_DUREE_ALLAIT_SEIN_EXCLU	so_source_value: numeric variable describing duration in weeks	Child
	SURVEY_OBSERVATIONS	Breastfeeding at 24 months	so_source_column =M24_ALLAITEMENT	so_source_value: “0”,”1”	Child

	SURVEY_OBSERVATIONS	Breastfeeding duration in weeks at 24 months	so_source_column =M24_DUREE_ALLAIT_SEIN	so_source_value: numeric variable describing duration in weeks	Child
	SURVEY_OBSERVATIONS	Breastfeeding duration in weeks of exclusively breastfed baby at 24 months	so_source_column =M24_DUREE_ALLAIT_SEIN_EXCLU	so_source_value: numeric variable describing duration in weeks	Child
ARS	SURVEY_OBSERVATIONS	Breastfeeding at discharge from birth hospitalization	so_source_table=CAP2 so_source_column =ALLATTA	so_source_value : 01 = only breast milk 02 = breast milk with the addition of water or other liquids other than milk 03 = breast milk and infant formula 04 = infant formula 00= newborn is born dead 99 = not detected or incorrect	Child
	PROCEDURES_SPC	Woman attending training and support groups for breastfeeding at a primary care community centre	meaning_of_procedure = “service_for_breastfeeding”	NA	Mother

4.4.1.2 ARS - ALLATTA

The "01" modality - during the entire hospital stay the baby was only given breast milk (from mother or milk bank).
The "02" mode - during the entire hospital stay, in addition to breast milk, even occasionally, water or other liquids were administered (eg glucose solution, chamomile, herbal teas).
The "03" modality - during the whole hospital stay, in addition to breast milk, infant formula has also been administered occasionally, regardless of the addition of other liquids.
The "04" mode was used if only formula milk was administered, with no breastmilk, with or without other liquids.

To align with Wales and Haute-Garonne, the Tuscany data will be recorded by merging 01, 02 and 03 to give 'any breastfeeding'. 04 remains as 'no breastfeeding'.

4.4.1.3 EFEMERIS

M9_ALLAITEMENT, 0= no breastfeeding at 9 months /1= breastfeeding at 9 months
M9_DUREE_ALLAIT_SEIN in weeks at 9 months
M9_DUREE_ALLAIT_SEIN_EXCLU in weeks of exclusively breastfed baby at 9 months
M24_DUREE_ALLAIT_SEIN_EXCLU in weeks of exclusively breastfed baby at 24 months

To align Haute-Garonne and Wales data sources we must lose data from France:
If duration ≥ 6-8/52 code *M9_DUREE_ALLAIT_SEIN in weeks at 9 months* as 1 = breastfeeding any.
If duration missing (estimation 50%), code breastfeeding at 9 months as breastfeeding at 6-8 weeks, leaving 'not feeding at 9 months' as unknowns (not missing data). We'll need a sensitivity analysis without this substitution.

Wales has some data at 6 months (60-75% completion) and France has information at 24 months, but these will not be explored in the demonstration projects.

The data sources are compared in **Table 4.3**.

Table 4.3 Strengths and limitations (for each variable)

Datasource	Variable	Strengths	Limitations
ARS	so_source_table="CAP2" so_source_column ="ALLATTA" from SURVEY_OBSERVATIONS	Complete for the almost 100% of children Probably valid (collected at hospital) Very detailed	Collected at birth, so may signify intention, rather than practice.
SAIL Databank	so_source_table =NCCHD so_source_column =BREASTFEED_BIRTH_FLG from SURVEY_OBSERVATIONS	Complete for almost >90%	Collected at birth, so may signify intention, rather than practice. Vulnerable to social desirability response bias.
	so_source_column =BREASTFEED_8_WKS_FLG from SURVEY_OBSERVATIONS	Detailed, even if converted into a binary variable	~75% complete

EFEMERIS	so_source_column =J8_ALLAITEMENT from SURVEY_OBSERVATIONS	The 'Protection service' collects health certificates from hospitals, private paediatricians, GPs and sometimes mothers. We are missing children for whom we don't have any of the three health certificates. Estimated coverage of about 80% of Haute-Garonne. Concerning quality, data recorded in the health certificates are partly self-reported by the woman, and partly collected by the health professional
	so_source_column =M9_ALLAITEMENT from SURVEY_OBSERVATIONS	
	so_source_column =M9_DUREE_ALLAIT_SEIN from SURVEY_OBSERVATIONS	
	so_source_column =M9_DUREE_ALLAIT_SEIN_EXCLU from SURVEY_OBSERVATIONS	

4.4.2 Exposures of DP2 (taken from SAP 1.3.7)

These are defined in demonstration project 2, and will remain unchanged, see **Table 4.4**.

Table 4.4. Exposures affecting breastfeeding

Variable	Definition in words	Categorization
Prescribed medicines in pregnancy	We shall define the main exposure of interest, and co-exposures of concern.	Medicine ATC groups to be defined in turn
Co-prescriptions	Opioids, benzodiazepines, AEDs, antipsychotics, gabapentinoids, anti-cancer therapies	In ATC codes
Indication for prescription*		
Medicated depression		
Unmedicated depression		
Co-morbidities		As in the DP
Discontinuation of prescription in trimester 1		
Discontinuation of prescription pre-pregnancy		
Prescriptions in T2 or T3		
Prescriptions during breastfeeding weeks 1-8		

** Note. This may not be possible for all conditions, and is for discussion. We have previously published using 'depression medicated' and 'depression unmedicated' (Jordan et al 2019). This is confounded by severity of indication, and is predicated on the depression diathesis hypothesis; however, it represents one strategy to explore the contributions of both the medicines and the condition.*

4.4.3 Breastfeeding Restrictions

Dyads where breastfeeding may be compromised or against medical advice will be described separately, and excluded from the main analysis, see **Table 4.5**. Many children with major congenital anomalies are breastfed, but we have no information regarding the overall picture (*Silva et al, 2021*).

Table 4.5. Conditions and medicines complicating breastfeeding

Infant		**Comments**
Any major anomaly	All EUROCAT births	
	ICD10	
Cleft palate	Q35	
Both clefts	Q37	
Galactosaemia -ever	E74.2	Only in infant
Brain injury NOS to infant	S06.9	we shall not check the mother for "brain injury"
Cerebral injury at birth / intracranial laceration	P10	
CNS birth injuries	P11	peripheral nerve injuries have not been excluded, as they may not affect breastfeeding
Mother		**Comments**
TB of respiratory system	A15	before or during pregnancy or before 4-8week data collection. There is a possibility of open TB.
Breast cancer	C50	before or during pregnancy or before 4-8week data collection
Personal history of breast cancer by time of birth	Z80.3	
Antineoplastic agents	ATC L01***	Administered during pregnancy or up to 8 weeks post-partum.
Clozapine (ATC N05AH02) in pregnancy	Clozapine (ATC N05AH02)	Administered during pregnancy or up to 8 weeks post-partum.
Lithium (ATC N05AN / N05AN01) in pregnancy	Lithium (ATC N05AN / N05AN01)	Administered during pregnancy or up to 8 weeks post-partum.

4.4.4 Covariates

For both breastfeeding and neurodevelopmental outcomes, we shall account for several covariates, as in SAP 1.3.7 (*Jordan et al., 2022*). These have been defined and coded in other demonstration projects,and will remain unchanged.

4.5 Discussion

In this section of the deliverable, we described how to collect information on breastfeeding from three different data sources: EFEMERIS (Haute-Garonne), SAIL Databank (Wales) and ARS (Tuscany, Italy). We used information from the DP2 protocol, Subtask 1.3.7 Breastfeeding section, the SAP for 1.3.7 (including excel sheets), the ConcePTION catalogue and conception CDM specification document. Information collected is currently used in Demonstration Project 2 to describe breastfeeding at birth and at 4-8 weeks after birth.

A systematic scoping review to May 2022 identified 11 papers from ten established databases with information on medicines exposure and breastfeeding; we found no multi-centre studies (Jordan et al 2023). This review identified the urgent need for more data to inform women, healthcare professionals, and pharmaceutical companies of:

1) as yet unquantifiable, but probably rare, serious harms to infants exposed to medicines via breastmilk,
2) infants who are phenotypically vulnerable to medicines exposure via breastmilk
3) unknown long-term harms (we are able to follow up for developmental outcomes), and
4) the more insidious but more pervasive harm in terms of reduced breastfeeding rates following medicines exposure in late pregnancy and peri-partum. Data from socio-economically vulnerable communities is key to this.

Both breastfeeding and neurodevelopment depend on many inter-related factors, necessitating databanks with data on a full range of covariates, and multivariable analyses, as described in SAP 1.3.7. As in all pharmaco-epidemiological analyses, demonstration of association is not tantamount to causation, but it allows professionals to identify vulnerable dyads and target support.

To conclude, we need to retrieve the data to:

1. ensure infants are monitored appropriately for any adverse drug reactions
2. inform breastfeeding patients using long-term medicines as to whether the benefits of breastfeeding outweigh exposure to medicines via breastmilk and
3. target additional support to breastfeeding patients whose medicines may affect breastfeeding.

4.6 References

(Davies et al, 2020) Davies G, Jordan S, Thayer D, Tucker D, Humphreys I (2020) Medicines prescribed for asthma, discontinuation and perinatal outcomes, including breastfeeding: A population cohort analysis. PLOS ONE 15(12): e0242489. https://doi.org/10.1371/journal.pone.0242489

(Ghozy et al, 2020) Ghozy S, Tran L, Naveed S, Quynh TTH, Helmy Zayan A, Waqas A *et al*: Association of breastfeeding status with risk of autism spectrum disorder: A systematic review, dose-response analysis and meta-analysis. *Asian J Psychiatr* 2020, 48:101916101916.

(Jordan et al, 2013) Jordan S, Watkins A, Storey M, Allen SJ, Brooks CJ, Garaiova I, Heaven ML, Jones R, Plummer SF, Russell IT, Thornton CA, Morgan G. (2013) Volunteer Bias in Recruitment, Retention, and Blood Sample Donation in a Randomised Controlled Trial Involving Mothers and Their Children at Six Months and Two Years: A Longitudinal Analysis. PLoS ONE 8(7): e67912. doi:10.1371/journal.pone.0067912

(Jordan et al, 2022a) Jordan S, Bromley R, Damase-Michel C, Given J, Komninou S, Loane M, Marfell N, Dolk H. Breastfeeding, pregnancy, medicines, neurodevelopment, and population databases: the information desert. Int Breastfeed J. 2022 Aug 2;17(1):55. doi: 10.1186/s13006-022-00494-5. PMID: 35915474. https://internationalbreastfeedingjournal.biomedcentral.com/articles/10.1186/s13006-022-00494-5

(Jordan et al 2022b) Glossary of Terms used in Pharmacovigilance and Pharmacoepidemiology (wp1 glossary in the wp1 members' area) https://members.imi-conception.eu/Member-Area/Work-Package-1?folderId=5702&view=gridview&pageSize=10 or ConcePTION - Work Package 1 (imi-conception.eu)

(Jordan et al, 2019) Jordan S, Davies GI, Thayer DS., Tucker D., Humphreys I. 2019 Antidepressant prescriptions, discontinuation, depression and perinatal outcomes, including breastfeeding: a population cohort analysis. PLOS ONE 14(11): e0225133. https://doi.org/10.1371/journal.pone.0225133

(Jordan et al, 2023) Jordan S, Komninou S, Lopez Leon S (2023) Where are the data linking infant outcomes, breastfeeding and medicine exposure? A systematic scoping review. PLOS ONE 18(4): e0284128. https://doi.org/10.1371/journal.pone.0284128

(Jordan et al 2024) Task 1.3.7 protocol_V4_18.2.21 v5.docx (live.com)

(Kramer et al, 2008) Kramer MS, Aboud F, Mironova E, Vanilovich I, Platt RW, Matush L *et al*: Breastfeeding and child cognitive development: New evidence from a large randomized trial. *Arch Gen Psychiatry* 2008, 65(5):578-584

(Lacroix et al, 2009) Lacroix I, Hurault C, Sarramon MF, Guitard C, Berrebi A, Grau M, Albouy-Cossard C, Bourrel R, Elefant E, Montastruc JL, Damase-Michel C. Prescription of drugs during pregnancy: a study using EFEMERIS, the new French database. Eur J Clin Pharmacol. 2009 Aug;65(8):839-46. doi: 10.1007/s00228-009-0647-2. Epub 2009 Apr 14. PMID: 19365629.

(Loane et al, 2020) Spreadsheet containing all additional data sources for the ConcePTION Data Source Catalogue Available: https://www.imi-conception.eu/wp-content/uploads/2019/09/ConcePTION_D1.1_spreadsheet-containing-all-additional-data-sources-for-the-ConcePTION-Data-Source-Catalogue.pdf (accessed 15.4.24)

(McAndrew et al, 2012) McAndrew F, Thompson J, Fellows L, Large A, Speed M, Renfrew M: The infant feeding survey 2010. In. Edited by NHS Information Centre for Health and Social Care. Office of National Statistics; 2012.

(Whitney et al, 2023) Whitney MD, Holbrook C, Alvarado L, Boyd S. Length of Maternity Leave Impact on Mental and Physical Health of Mothers and Infants, a Systematic Review and Meta-analysis. Matern Child Health J. 2023 Aug;27(8):1308-1323. doi: 10.1007/s10995-022-03524-0. Epub 2023 Apr 12. PMID: 37043071.

(WHO 2008) WHO 2008 Indicators for Assessing Infant and Young Child Feeding Practices – Part I: Definitions. Conclusions of a Consensus Meeting Held 6–8 November 2007 in Washington D.C. https://apps.who.int/iris/bitstream/handle/10665/43895/9789241596664_eng.pdf;jsessionid=DB32B0C8C42A0F61174ECAF42D8FC8FD?sequence=1

(Silva et al, 2021) Silva et al, Perinatal morbidities, congenital malformations and breastfeeding outcomes, Journal of Neonatal Nursing,Volume 27, Issue 6, 2021,Pages 412-418, https://doi.org/10.1016/j.jnn.2021.05.003

Chapter 5. Misclassification

5.1 Introduction

Electronic healthcare data sources are a key resource in assessing the occurrence of events in human populations, playing a major role in public health and regulatory decision-making. During the recent coronavirus crisis, calculation of background rates of events that may have later occurred as adverse reactions to the new vaccines were requested by the European Medicines Agency ahead of the vaccination campaign and, in March 2021, were used to rapidly assess potential causality of safety signals emerging from spontaneous reports. In pregnancy studies, the occurrence of adverse maternal or infant outcomes in the general population and in populations exposed to medications can be used to provide signals on the safety of utilization for pregnant women, because clinical trials in this population are rarely conducted. Information on the prevalence of conditions such as hypertensive disorders during pregnancy is also important for pregnancy studies.

In database studies, the occurrence of the event of interest is measured using an indicator. Due to the nature of such databases, the indicator may be affected by errors: *false positives* are cases that are detected by the indicator but are not true cases; *false negatives* are true cases that are missed by the indicator. The degree of misclassification of an indicator is measured by its *validity indices*, including *sensitivity*, which is the proportion of true positives among true cases in the population, *specificity*, which is the proportion of true negatives among non-cases in the population, and *positive predictive value* (PPV), which is the proportion of true positives among cases detected by the indicator. Guidelines in the conduct and reporting of database studies recommend estimating validity indices, which would allow to adjust observed rates using a straightforward correction (PPV/SE).

However, validation is often not feasible due to cost, time constraints, or ethical reasons. When validation studies are conducted, they often estimate only the PPV. Indeed, estimation of the PPV requires extracting a sample of cases positive for the indicator and their assessment through comparison with a gold standard. On the contrary, the true outcome is not observed, and it is often a rare event, thus estimating the sensitivity requires large validation samples to get enough true positive cases. Therefore, the rate of occurrence is overestimated if $\mathrm{SE} > \mathrm{PPV}$, or underestimated if adjusted only with the PPV or if $\mathrm{PPV} > \mathrm{SE}$.

A second problem is when the objective of the database study is to assess a causal relationship between an exposure and an outcome. Misclassification of the indicator of the outcome may introduce bias both in the risk ratio and in the risk difference. Each validity index can be defined with respect to the exposure groups. The sensitivity or specificity of the indicator is said to be *non-differential* if it does not depend on the exposure group. Risk difference is prone to bias irrespective of whether sensitivity is differential. If sensitivity and specificity are non-differential, the risk ratio is biased towards the null if $\mathrm{PPV} < 1$. When $\mathrm{PPV} = 1$ and sensitivity is non-differential, then the risk ratio is unbiased: for this reason, guidelines recommend choosing indicators that minimize false positives. However, a test for non-differentiality is not available in the literature, so this assumption usually goes untested.

In this paper, we build on a strategy suggested by Lanes et al. Alongside a primary indicator with a high PPV, we suggest searching for an auxiliary *screening indicator* aimed at capturing all (or nearly all) of the cases not captured by the specific algorithm so that their union has near-perfect sensitivity,

while still being specific enough to exclude many non-cases. Estimating the PPV of both indicators then becomes feasible, possibly stratified per exposure.

This paper considers the situation when a screening algorithm is available. We have an estimate of PPV for both indicators, and we are able to make at least one in a range of assumptions. We first show that it is possible to estimate the sensitivity of the primary indicator or a lower bound thereof. Furthermore, in the case when an exposure is involved in the study, and the PPV of both indicators is estimated across exposure strata, we provide a hypothesis test to verify whether the sensitivity of the main indicator is non-differential. A simulation study is conducted to verify the performance of the test in multiple scenarios. Finally, we demonstrate how to use the PPVs to adjust, or to obtain bounds on, the number of cases, risks, risk ratios and risk differences.

5.2 Notation

We denote by Y the binary variable for the true value of the outcome of interest, with A an observed, possibly misclassified, binary variable that is the primary indicator for Y, and by B a screening indicator for the same outcome. Combining A and B yields further indicators, such as $\mathrm{A} \cup \mathrm{B}$, namely the indicator retrieving all cases retrieved by either A or B, and $\mathrm{A} \cap \mathrm{B}$ which is the *intersection* indicator retrieving all cases retrieved by both A and B.

We assume a framework with a finite target population with N subjects. For an indicator I, we denote by $\mathrm{TP_I}$, $\mathrm{FN_I}$ and $\mathrm{FP_I}$ the true positives, false negatives and false positives in the target population, respectively. Then, $\mathrm{N_Y} = \mathrm{TP_I} + \mathrm{FN_I}$ is the number of subjects in the study population for whom the outcome has occurred and $\mathrm{N_I} = \mathrm{TP_I} + \mathrm{FP_I}$ is the number of subjects identified by the indicator I.

We indicate with $\pi = \mathrm{N_Y}/\mathrm{N}$ the true, unobserved prevalence of Y and with $\mathrm{P_I} = \mathrm{N_I}/\mathrm{N}$ the observed prevalence. Note that π does not depend on the indicator.

The validity indices are defined for each indicator I as follows: *sensitivity* $\mathrm{SE_I} = \mathrm{TP_I}/(\mathrm{TP_I} + \mathrm{FN_I})$; *specificity* $\mathrm{SP_I} = \mathrm{TN_I}/(\mathrm{TN_I} + \mathrm{FP_I})$; *positive predictive value* $\mathrm{PPV_I} = \mathrm{TP_I}/(\mathrm{TP_I} + \mathrm{FP_I})$.

We focus on the case of a binary exposure, denoting with e the exposed and with $\bar{\mathrm{e}}$ the non-exposed. We denote by $\mathrm{N^e}$ and $\mathrm{N^{\bar{e}}}$ the number of subjects in the study population who are in the exposed and non-exposed groups, respectively. Then, $\mathrm{SE_I^e}$ is the sensitivity for the exposed and $\mathrm{SE_I^{\bar{e}}}$ for the non-exposed. Similarly, $\mathrm{SP_I^e}$ is the specificity for the exposed and $\mathrm{SP_I^{\bar{e}}}$ for the non-exposed. An indicator I has non-differential sensitivity if $\mathrm{SE_I^e} = \mathrm{SE_I^{\bar{e}}}$ and non-differential specificity if $\mathrm{SP_I^e} = \mathrm{SP_I^{\bar{e}}}$.

Lastly, let's adopt the notation in which the "hat" symbol above a parameter (e.g., $\widehat{\mathrm{PPV}}$) indicates that we are referring to a sample statistic rather than the population value of the parameter.

5.3 Sensitivity of indicators

5.3.1 Estimating the sensitivity

Let's first consider the situation when we can assume that $SE_{A\cup B} = 1$, i.e., B is a perfect screening indicator. The following formula, proven in Appendix B, shows how the sensitivity of the primary indicator can be derived from observed prevalences and positive predictive values:

$$SE_A = 1 - P_B \times PPV_{B-}P_{A\cap B} \times PPV_{A\cap B} / P_A \times PPV_A + P_B \times PPV_{B-}P_{A\cap B} \times PPV_{A\cap B} \quad (1)$$

As an example, consider a study on myocardial infarction under the assumption that all cases that are not promptly admitted to a hospital die. In this case, A is the indicator retrieving hospital admissions with a specific diagnostic code of infarction, whereas B is the indicator retrieving hospitalization with unspecified codes of infarction or death for any cause.

If $SE_{A\cup B}$ cannot be assumed to be 1, it is proven in Appendix B that the right-hand side of equation [eq1] is an upper bound for the sensitivity of A, namely

$$SE_A \leq 1 - P_B \times PPV_{B-}P_{A\cap B} \times PPV_{A\cap B} / P_A \times PPV_A + P_B \times PPV_{B-}P_{A\cap B} \times PPV_{A\cap B} \quad (2)$$

If no case is retrieved by both A and B, namely $P_{A\cap B} = 0$, equation (2) simplifies to

$$SE_A \leq P_A \times PPV_A / P_A \times PPV_A + P_B \times PPV_B \quad (3)$$

The approach summarised by formulas (1)-(3) has the advantage of providing an estimate of the sensitivity based on PPV; the drawback is that the estimator of the sensitivity is based on the estimators of three PPV s, namely those of the primary indicator A, the screening indicator B, and their intersection $A \cap B$. This could imply a loss of accuracy due to conveying multiple sources of uncertainty. Formulas (2)-(3) can be written using the absolute number of subjects identified by an indicator (N_I) in place of the observed prevalence (P_I). Thus, by way of illustration, formula (3) becomes:

$$SE_A \leq N_A \times PPV_A / N_A \times PPV_A + N_B \times PPV_B \quad (4)$$

5.3.1.1 Example: estimating an upper bound for the sensitivity of an indicator of angioedema

This example was presented at the 38th International Conference on Pharmacoepidemiology and Therapeutic Risk Management. Angioedema is a local, circumscribed edema due to increased plasma leakage from capillaries into the deep layers of the skin and mucous membranes. The ENTRESTO study is a post-authorization safety study requested by the European Medicines Agency as part of the Risk Management Plan of Entresto, a medicinal product indicated for the treatment of heart failure. One of the objectives of the study was to estimate the incidence and relative risk of angioedema, which is a suspected adverse reaction to the medicinal product.

As part of the ENTRESTO study, a validation study was conducted in the data source accessed by ARS Toscana, to estimate the PPV and sensitivity of an indicator A for angioedema (ICD9CM code 995.1). A screening indicator B was identified, retrieving cases of hypersensitivity (ICD9CM codes 374.82, 376.33, 478.25, 478.6, 478.75, 508.8, 708.0, 708.1, 708.8, 708.9, 782.3,995.0, 995.2, 995.27). In the study population, A retrieved $\mathrm{N_A} = 34$ cases and B retrieved $\mathrm{N_B} = 451$ cases, while no case was found by both A and B, hence $\mathrm{N_{A\cap B}} = 0$. All cases retrieved by A were validated, and 12 were found to be true cases, hence $\widehat{\mathrm{PPV}}_\mathrm{A} = 35.3\%$. A sample of 96 cases was validated from those retrieved by B, and 8 were found to be true cases, hence $\widehat{\mathrm{PPV}}_\mathrm{B} = 8.3\%$. Therefore, from formula (4),

$$\widehat{\mathrm{SE}}_\mathrm{A} \leq 24.5\%.$$

Note that, even though the PPV of B is much lower than the PPV of A, since its observed prevalence is much higher, the true cases among those retrieved by B may be a substantial share.

5.4 A test for non-differential sensitivity

5.4.1 Methods

The non-differentiality hypothesis states that the sensitivity of the primary indicator is identical across exposure groups, namely

$$\mathrm{H_0}: \mathrm{SE_A^e} = \mathrm{SE_{\bar{A}}^e} \quad (5)$$

To carry out the test for the hypothesis of non-differential sensitivity, equation (5) can be rewritten as

$$\mathrm{H_0}: \frac{\mathrm{SE_A^e}}{\mathrm{SE_{\bar{A}}^e}} - 1 = 0 \quad (6)$$

To derive the test statistic, we estimate the sensitivities of A in the exposure strata by exploiting a screening indicator B. Specifically, in Appendix D we prove that, under the assumption of non-differentiality of the sensitivity of $\mathrm{A} \cup \mathrm{B}$ (i.e., $\mathrm{SE_{A\cup B}^e} = \mathrm{SE_{\bar{A}\cup B}^e}$), the hypothesis (6) can be expressed in terms of observed prevalences (P) and positive predictive values (PPV) as follows:

$$\mathrm{H_0}: \frac{\mathrm{P_A^e PPV_A^e}\left(\mathrm{P_{\bar{A}}^e PPV_{\bar{A}}^e} + \mathrm{P_B^e PPV_B^e} - \mathrm{P_{\bar{A}\cap B}^e PPV_{\bar{A}\cap B}^e}\right)}{\mathrm{P_{\bar{A}}^e PPV_{\bar{A}}^e}\left(\mathrm{P_A^e PPV_A^e} + \mathrm{P_B^e PPV_B^e} - \mathrm{P_{A\cap B}^e PPV_{A\cap B}^e}\right)} - 1 = 0 \quad (7)$$

The estimate of the PPV, denoted as $\widehat{\mathrm{PPV}}$, is achieved by comparing each subject with a gold standard, which provides the true classification of the outcome. The validation study must be carefully designed to provide accurate estimates of the PPVs, and the sampling process must consider each exposure stratum and each indicator.

Once the required PPVs are estimated, given that the observed prevalences P are known, it is straightforward to calculate the value of test statistic. To carry out the test, it is required to determine

the distribution of the test statistic under the null. As this is not straightforward, we proceed with the bootstrap method, and reject the hypothesis if the 95% percentile confidence interval does not include the zero.

5.5 Simulation study

To assess the performance of the proposed non-differentiality test, we conducted a simulation study over multiple scenarios in which the indicators' validity indices and the population characteristics were varied. We generated two binary vectors: one representing the true outcome of interest Y and the other representing the exposure groups e, $\bar{e}$. The proportion of subjects exposed was fixed at 5%. We explored three scenarios for the true prevalence of the outcome ($Y = 1$) in the unexposed group: $\pi^{\bar{e}} = 0.01$ (rare), $\pi^{\bar{e}} = 0.05$, and $\pi^{\bar{e}} = 0.1$ (common). Moreover, we considered three scenarios for the relative risk $RR = \pi^{e}/\pi^{\bar{e}}$: 1, 1.2 and 2.

Then, we generated two indicators of Y denoted with A (primary) and B (screening), which are misclassified with specific error rates for each exposure group. $SE_{A\cup B}$ was fixed to 0.85 or 0.95. In each scenario, $SE_{A\cup B}$ remains equal in both exposure groups to comply with the assumption of non-differentiality of the sensitivity of $A \cup B$.

The values of the sensitivities are chosen to produce a sensitivity ratio $SE_A^e/SE_A^{\bar{e}}$ in {0.6, 0.8, 1, 1.25, 1.67}, where values above 1 are reciprocals of those below 1 (in fact, $1.25 = 1/0.6$ and $1.67 = 1/0.8$). In particular, when $SE_{A\cup B} = 0.85$, then $SE_A^{\bar{e}}$ was set at 0.5 and SE_A^e varied in {0.3, 0.4, 0.5, 0.625, 0.835}; alternatively, when $SE_{A\cup B} = 0.95$, then SE_A^e was fixed at 0.75 and $SE_A^{\bar{e}}$ varied in {0.45, 0.60, 0.75, 0.94} (in this case the fifth value is missing since it is impossible to produce a sensitivity ratio equal to 1.67).

To study the possible scenarios in real-life situations, we varied the intersection among A and B, considering two values for $SE_{A\cap B}$: 0 and 0.2. When $SE_{A\cap B}$ was equal to 0, $SP_{A\cap B}$ was set to 1 to create a *non-intersection scenario*, where no subject tests positive for both indicators. Non-intersection scenarios are plausible, as seen in the ENTRESTO study.

The specificity SP was set to $1-\pi^{\bar{e}}/10$ and $1-\pi^{\bar{e}}$ for A and B, respectively, in both exposure groups; this implies that the PPV has a maximum of 90% for A and 50% for B. We further assumed that $1-SP_{A\cap B} = (1-SP_A) \times (1-SP_B)$, i.e., the number of false positives of A is independent of the number of false positives of B.

For each scenario, we generated a finite population of size 1 million with the above characteristics. Then, repeating 1000 times, we drew a validation sample and carried out the test. We considered three different sizes of the validation sample: 250, 500 and 750.

We carried out random sampling across indicators and exposure. Specifically, we sampled 40% of the subjects from those with $A = 1$, 40% from $B = 1$ and 20% from $A = 1$ and $B = 1$ jointly. Each of the above groups has 50% exposed and 50% unexposed.

In equation (7), the PPV was replaced by its estimate $\widehat{PPV}$, given by the proportion of true positives ($Y = 1$) in the subset of subjects flagged by the indicator ($A = 1$ or $B = 1$).

The non-differentiality test was conducted using the bootstrap method. Specifically, 500 bootstrap samples were generated from the validation sample. For each bootstrap sample, the test statistic was calculated, and its empirical distribution was derived. The 95% percentile confidence interval was then computed, and the null hypothesis was rejected if the interval did not encompass zero. The power of the test is the proportion of times the test rejected the null hypothesis, denoted as Pr_{rej}. When the null hypothesis is true (sensitivity ratio equal to one), Pr_{rej} is the empirical significance level of the test.

The simulation was conducted using the R software. The code is freely available on GitHub at the following link: https://github.com/GiorgioLimoncella/NonDifferentialityTest.

The findings of the simulation study are summarised in Figure 1 and Figure 2, which show how the rejection rate Pr_{rej} varies according to the true sensitivity ratio, in the case when $\overline{SE^e} = 0.5$ and $\overline{SE^e} = 0.75$, respectively; graphs are stratified by $\overline{\pi^e}$, $SE_{A\cap B}$ and RR.

First, under the null hypothesis of a sensitivity ratio equal to 1, the proposed test has a rejection proportion Pr_{rej} close to the significance level 0.05. Then, the power of the test depends on the factors of the simulation design in the expected way:

- positive association with the size of the validation sample (strong)
- positive association with the distance of the sensitivity ratio $SE^e/\widetilde{SE^e}$ from the null hypothesis (strong)
- positive association with the baseline sensitivity in the unexposed (strong)
- negative association with the sensitivity of the intersection $SE_{A\cap B}$ (strong)
- positive association with the distance of the risk ratio RR from 1 (moderate)
- negative association with the prevalence in the unexposed $\overline{\pi^e}$ (weak)

Let us explore the cases when the power is large, namely exceeding 0.8. For a validation sample of size 750, the power is always large if the sensitivity ratio is 0.6 or 1.67, whereas for values closer to 1 the power is large only for some configurations with $SE_{A\cap B} = 0$. For a validation sample of size 500, the performance worsens, though the power is still large for most configurations if the sensitivity ratio is 0.6 or 1.67. Finally, for a validation sample of size 250, the performance further worsens, so the power is large only for some configurations if the sensitivity ratio is 0.6 or 1.4, while it is quite low if the sensitivity ratio is closer to 1.

Figure:5.1 Proportion of times when the null hypothesis is rejected, Pr_{rej}. 12 different scenarios are presented in which $SE_{A\cap B}$, RR and $\pi^{\overline{e}}$ vary. Each panel presents the power of the test power across three different sample sizes: 250, 500, and 750. $SE^{\overline{e}}$ is set to 0.5.

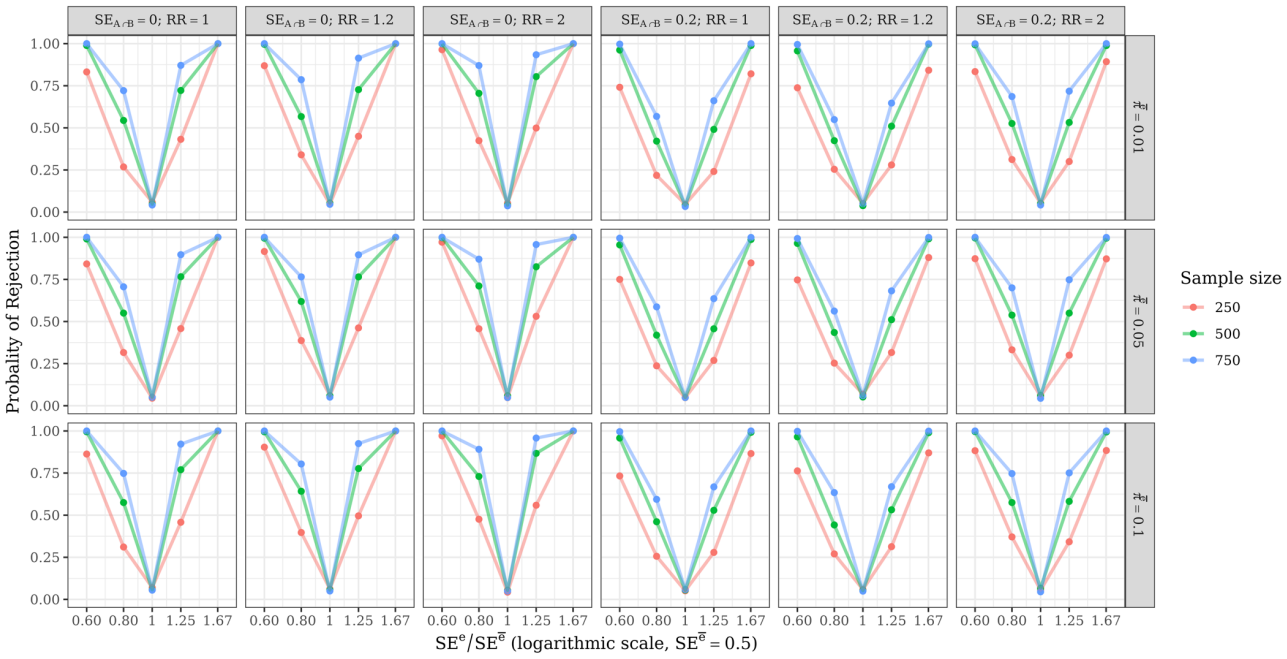
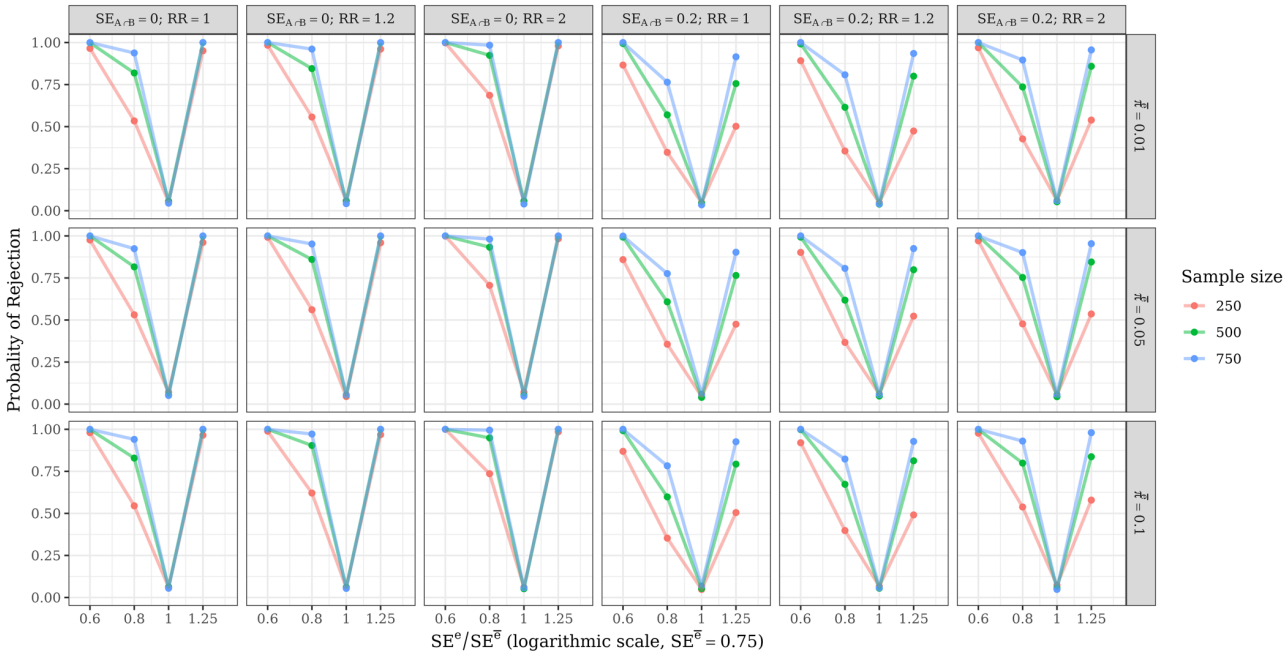


Figure 5.2: Proportion of times when the null hypothesis is rejected, Pr_{rej}. 12 different scenarios are presented in which $SE_{A\cap B}$, RR and $\pi^{\overline{e}}$ vary. Each panel presents the power of the test power across three different sample sizes: 250, 500, and 750. $SE^{\overline{e}}$ is set to 0.75.



5.6 Adjusting the misclassification bias for number of cases and risk, risk ratio and risk difference

In this section, we will discuss a situation where we have an observed outcome Y and have conducted validation studies for an indicator A with high specificity and a screening algorithm B, obtaining estimates of the positive predictive value (PPV) for both indicators. The issue is how this information can be used to adjust measures of the number of cases and, as long as the PPVs have been estimated across exposure strata, to adjust the risk ratio and the risk difference.

In this section, we use the term 'risk' to indicate a measure of risk where the numerator is the number of cases, and the calculation of the denominator does not depend on the detection of cases. This is the case for the prevalence, the cumulative incidence, and (with approximation) the rate of a rare event: in the latter case, indeed, the risk is the number of new cases divided by the person time still at risk, which depends on the detection of cases; however if the event is rare, a small variation in the detection of new cases has a negligible impact on the person time at risk.

5.6.1 Number of cases and risk

The number of cases observed in the population $\mathrm{N_A}$ differs from the true number $\mathrm{N_Y}$. If we know $\mathrm{PPV_A}$ and $\mathrm{SE_A}$, a straightforward correction is possible:

$$N_Y = N_A \times \frac{PPV_A}{SE_A} \quad (8)$$

This is because

$$\begin{aligned} N_Y &= TP + FN \\ &= (TP + FN) \times \frac{(TP + FP)}{TP} \times \frac{TP}{(TP + FP)} \\ &= (TP + FP) \times \frac{(TP + FN)}{TP} \times \frac{TP}{(TP + FP)} \\ &= N_A \times \frac{1}{SE_A} \times PPV_A \end{aligned}$$

Denoting with R_Y a measure of the risk for Y in the population and with R_A its measure through A, and dividing each member of equation (8) by the common denominator, we obtain the following formula:

$$R_Y = R_A \times \frac{PPV_A}{SE_A} \quad (9)$$

When we can assume that $SE_{A \cup B} = 1$, the number of cases and the true risk are then obtained as follows:

$$N_Y = N_A \times PPV_A / 1 - P_B \times PPV_B - P_{A \cap B} \times PPV_{A \cap B} / P_A \times PPV_A + P_B \times PPB_B - P_{A \cap B} \times PPV_{A \cap B} \quad (10)$$

and

$$R_Y = R_A \times PPV_A / 1 - P_B \times PPV_B - P_{A \cap B} \times PPV_{A \cap B} / P_A \times PPV_A + P_B \times PPB_B - P_{A \cap B} \times PPV_{A \cap B} \quad (11)$$

In cases where it cannot be assumed that $SE_{A \cup B} = 1$, a lower bound for the number of cases and for the risk is derived:

$$N_Y \geq N_A \times PPV_A / 1 - P_B \times PPV_B - P_{A \cap B} \times PPV_{A \cap B} / P_A \times PPV_A + P_B \times PPB_B - P_{A \cap B} \times PPV_{A \cap B} \quad (12)$$

and

$$R_Y \geq R_A \times PPV_A / 1 - P_B \times PPV_B - P_{A \cap B} \times PPV_{A \cap B} / P_A \times PPV_A + P_B \times PPB_B - P_{A \cap B} \times PPV_{A \cap B} \quad (13)$$

For example, in the ENTRESTO study, we can adjust N_Y as follows. Since $A \cap B = \emptyset$, the correction factor in formula (12) is

$$\widehat{PPV_A} / 1 - N_B \times \widehat{PPV_B} / N_A \times \widehat{PPV_A} + N_B \times \widehat{PPV_B} = 1.45 \quad (14)$$

Thus, the true number of cases is at least 1.45 times higher than $N_A = 34$:

$$N_Y \geq N_A \times 1.45 = 49.3 \quad (15)$$

Similarly, an estimate of risk for the study based on A would underestimate the true risk by at least 45%.

5.6.2 Risk ratio

The true risk ratio is based on the outcome Y:

$$RR_Y = \frac{N_Y^e / N^e}{N_Y^{\bar{e}} / N^{\bar{e}}} \quad (16)$$

In database studies, it is approximated using the observed indicator A:

$$RR_A = \frac{N_A^e / N^e}{N_A^{\bar{e}} / N^{\bar{e}}} \quad (17)$$

Brenner and Gefeller proved the following relationship:

$$RR_Y = RR_A \frac{PPV_A^e}{PPV_A^{\bar{e}}} \frac{SE_A^{\bar{e}}}{SE_A^e} \quad (18)$$

Note this formula also holds for rate ratios as long as the event is rare.

The common approach to adjust for measurement error is to assume that the sensitivity is non-differential across exposure groups ($SE_A^e = SE_A^{\bar{e}}$) so that the last factor in equation (18) disappears and the correction only needs the PPVs of A. If the screening algorithm B is available and it can be assumed that $SE_{A \cup B}$ is non-differential ($SE_{A \cup B}^e = SE_{A \cup B}^{\bar{e}}$), the statistical test introduced in the previous section can be used to test the hypothesis $SE_A^e = SE_A^{\bar{e}}$ of non-differential sensitivity and make the results more robust.

If the test fails, but it still can be assumed that $SE_{A \cup B}$ is non-differential, then the correction (18) becomes

$$RR_Y = RR_A \times \frac{P_A^e PPV_A^e + P_B^e PPV_B^e - P_{A \cap B}^e PPV_{A \cap B}^e}{P_A^{\bar{e}} PPV_A^{\bar{e}} + P_B^{\bar{e}} PPV_B^{\bar{e}} - P_{A \cap B}^{\bar{e}} PPV_{A \cap B}^{\bar{e}}} \quad (19)$$

5.6.3 Risk difference

The risk difference is defined as

$$RD_Y = \frac{N_Y^e}{N^e} - \frac{N_Y^{\bar{e}}}{N^{\bar{e}}}$$

By equation (8), the risk difference can be calculated through the validity indexes of an indicator A as follows

$$RD_Y = \frac{PPV_A^e}{SE_A^e} \times \frac{N_A^e}{N^e} - \frac{PPV_A^{\bar{e}}}{SE_A^{\bar{e}}} \times \frac{N_A^{\bar{e}}}{N^{\bar{e}}} \quad (20)$$

If A is chosen to have a very large PPV, this may come at the expense of the sensitivity. Formula (20) shows that this may induce a significant underestimation of the risk difference: the PPV of the screening indicator B can help quantify this.

Multiple scenarios are possible depending on the assumptions that can be made on A and on the PPVs that are available or can be estimated.

We first describe a simple scenario (Scenario 1) and prove a lower bound for the underestimation of the risk difference based on PPV_B. We then discuss a set of formulas that can be used when this simple scenario is not valid (Scenarios 2-5) and prove them in Appendix (Section 5.9)

All scenarios, except the second one, assume the existence of the screening indicator B. In Scenarios 1 and 3, the condition that SE_A is non-differential is required. This assumption can be justified on theoretical grounds or, if the PPV_B can be estimated across exposure strata, it can be verified through the non-differentiality test.

Scenario 1: SE_A non-differential ($SE_A^e = SE_A^{\bar{e}}$) and $PPV_A^e = PPV_A^{\bar{e}} = PPV_A = 1$ and $P_{A \cap B} = 0$

This is the case when indicator A has been chosen to be highly specific. The statement that $PPV_A = 1$ in both exposure strata may be an assumption or may be supported by a validation study.

Replacing the assumptions in formula (20) proves the following formula:

$$RD_Y = \frac{1}{SE_A} \times RD_A$$

where RD_A is the observed risk difference

$$RD_A = \frac{N_A^e}{N^e} - \frac{N_A^{\bar{e}}}{N^{\bar{e}}}$$

Hence, using the upper bound (3) for SE_A, we get the following lower bound for the estimator of the risk difference:

$$|RD_Y| \geq |P_A + P_B \times PPV_B/P_A \, RD_A| \quad \begin{matrix} \text{sign}(RD_Y) \\ = \\ = \end{matrix} \begin{matrix} \text{sign}(RD_A) \\ \\ |(1 + P_B \times PPV_B/P_A)RD_A| \end{matrix} \quad (21) / (22)$$

Note that this formula does not require that PPV_B is estimated across exposure strata.

Scenario 2: SE_A non-differential ($SE_A^e = SE_A^{\bar{e}} = SE_A$)

This is a scenario where the auxiliary indicator B is not available, but estimation stratified by exposure of the PPV of A can be obtained. Then, the following lower bound can be used:

$$\begin{aligned} \text{sign}(RD) = &\quad \text{sign}\left(PPV_A^e \times P_A^e - PPV_A^{\bar{e}} \times P_A^{\bar{e}}\right) \\ |RD| \geq &\quad \left|\left(PPV_A^e \times P_A^e - PPV_A^{\bar{e}} \times P_A^{\bar{e}}\right)\right| \end{aligned} \quad (23)/(24)$$

Scenario 3: SE_A non-differential ($SE_A^e = SE_A^{\bar{e}}$) and $PPV_A^e = PPV_A^{\bar{e}} = PPV_A = 1$

This is the generalization of Scenario 1 to the case when $P_{A\cap B} > 0$. Again, a lower bound can be provided where PPV of B can be estimated in the general population (and therefore, be retrieved from a previous study, if available):

$$\begin{aligned} \text{sign}(RD) = &\quad \text{sign}\left(P_A^e - P_A^{\bar{e}}\right) \\ |RD| \geq &\quad \left|\frac{P_A + P_B PPV_B - P_{A\cap B}}{P_A - 2P_{A\cap B}}\left(P_A^e - P_A^{\bar{e}}\right)\right| \end{aligned} \quad (25)$$

If $A \cap B = \emptyset$, then:

$$|RD| \geq \left|\frac{P_A + P_B PPV_B}{P_A}\left(P_A^e - P_A^{\bar{e}}\right)\right| \quad (26)$$

Scenario 4: $SE_{A\cup B}$ non-differential ($SE_{A\cup B}^e = SE_{A\cup B}^{\bar{e}} = SE_{A\cup B}$)

This is the generalization of Scenario 1 to the case when the assumption that there are no false positives in A cannot be made, and $P_{A\cap B} \geq 0$. Then, PPV of A, B and (if $P_{A\cap B} > 0$) $A \cap B$ must be estimated across exposure strata, and the following lower bound holds:

$$\begin{aligned} \text{sign}(RD) &= \text{sign}\left(\left(P_A^e \times PPV_A^e + P_B^e \times PPV_B^e - P_{A\cap B}^e \times PPV_{A\cap B}^e\right)\right. \\ &\quad \left.-\left(P_A^{\bar{e}} \times PPV_A^{\bar{e}} + P_B^{\bar{e}} \times PPV_B^{\bar{e}} - P_{A\cap B}^{\bar{e}} \times PPV_{A\cap B}^{\bar{e}}\right)\right) \\ |RD| &\geq \left|\left(P_A^e \times PPV_A^e + P_B^e \times PPV_B^e - P_{A\cap B}^e \times PPV_{A\cap B}^e\right)\right. \\ &\quad \left.-\left(P_A^{\bar{e}} \times PPV_A^{\bar{e}} + P_B^{\bar{e}} \times PPV_B^{\bar{e}} - P_{A\cap B}^{\bar{e}} \times PPV_{A\cap B}^{\bar{e}}\right)\right| \end{aligned} \quad (27)$$

If, on top of this, the assumption $SE_{A\cup B} = 1$ can be made, we obtain an equation:

$$\begin{aligned} RD = &\quad \left(P_A^e \times PPV_A^e + P_B^e \times PPV_B^e - P_{A\cap B}^e \times PPV_{A\cap B}^e\right) \\ &-\left(P_A^{\bar{e}} \times PPV_A^{\bar{e}} + P_B^{\bar{e}} \times PPV_B^{\bar{e}} - P_{A\cap B}^{\bar{e}} \times PPV_{A\cap B}^{\bar{e}}\right) \end{aligned} \quad (28)$$

Scenario 5: positive predictive value of A greater than sensitivity in the exposed group ($PPV_A^e > SE_A^e$)

We finally give a formula that can be used when PPV of A and B are available only among non-exposed: when exposure is rare, such parameters can be assumed to be equal to those in the general population, that may be available from previous studies. The formula holds under the assumption that the PPV of A among exposed is higher than its sensitivity, which is plausible if A can be built with convincingly high specificity.

$$\begin{aligned} \text{sign}(\text{RD}) &= \text{sign}\left(\text{P}_{\text{A}}^{\text{e}} - 1/\text{SE}_{\text{A}\cup\text{B}}\left(\text{P}_{\tilde{\text{A}}}^{\tilde{\text{e}}} \times \text{PPV}_{\tilde{\text{A}}}^{\tilde{\text{e}}} + \text{P}_{\tilde{\text{B}}}^{\tilde{\text{e}}} \times \text{PPV}_{\tilde{\text{B}}}^{\tilde{\text{e}}} - \text{P}_{\tilde{\text{A}}\cap\text{B}}^{\tilde{\text{e}}} \times \text{PPV}_{\tilde{\text{A}}\cap\text{B}}^{\tilde{\text{e}}}\right)\right) \\ |\text{RD}| &\geq \left|\text{P}_{\text{A}}^{\text{e}} - 1/\text{SE}_{\text{A}\cup\text{B}}\left(\text{P}_{\tilde{\text{A}}}^{\tilde{\text{e}}} \times \text{PPV}_{\tilde{\text{A}}}^{\tilde{\text{e}}} + \text{P}_{\tilde{\text{B}}}^{\tilde{\text{e}}} \times \text{PPV}_{\tilde{\text{B}}}^{\tilde{\text{e}}} - \text{P}_{\tilde{\text{A}}\cap\text{B}}^{\tilde{\text{e}}} \times \text{PPV}_{\tilde{\text{A}}\cap\text{B}}^{\tilde{\text{e}}}\right)\right| \end{aligned} \quad (29)$$

where $1/\text{SE}_{\text{A}\cup\text{B}} \geq 1$.

5.7 Discussion

When an event Y is the outcome of a database study, it is common practice to choose a very specific indicator A for Y, and, if validation is possible, to estimate PPV_{A}, to account for false positives captured by A. In this paper, we elaborated on the notion of a screening indicator B, aimed at capturing all (or nearly all) of the cases so that it has near-perfect sensitivity, while still being specific enough to exclude many non-cases. We showed that, by including the estimation of PPV_{B} in the validation study, we can at least partially account for false negatives that A fails to identify. Recently, the estimation of the sensitivity of indicators was recommended by regulatory authorities to improve the quality of evidence generated for regulatory purposes. While this recommendation is considered very arduous to comply with, we introduce a method that partially meets the recommendation.

When estimating the occurrence of Y, we provided a formula to adjust, or provide a lower bound for, the number of cases, the prevalence, the cumulative incidence, and the incidence rate of Y (the latter being valid only for rare events). An important application would be the safety of vaccines. Background rates of adverse events of special interest in the populations targeted by vaccines can be estimated before the deployment of the vaccine and be used during the vaccination campaign to monitor the occurrence of adverse events in the vaccinated population, in an observed-to-expected analysis. This occurred in Europe during the recent pandemic. If the background rates are underestimated, an observed-to-expected analysis is biased and may trigger false alerts. An updated strategy would be to pair each 'primary' indicator with a screening indicator, and to estimate the positive predictive values of both indicators to reduce the bias. This activity can be conducted in the inter-pandemic periods.

For association studies, we introduced several tools to address differential misclassification. Differential sensitivity is not a trivial matter. In safety studies, it may occur when studying suspected adverse reactions to a medicinal product since a clinician assessing the symptoms of a patient may use their exposure to the medicine in the diagnostic process, and the knowledge that the outcome is a possible adverse reaction may lead to record a more specific diagnostic code. Our tools are based on estimates of the PPVs of A and B across exposure strata. We introduced a statistical test for non-differential sensitivity based on the assumption that $\text{SE}_{\text{A}\cup\text{B}}$ is non-differential. Our test showed good power in detecting modest departures from non-differentiality with a validation sample of size 500, retrieved from cases detected by A, B and $\text{A} \cap \text{B}$ (if non-empty), independently on the prevalence of Y. This test can be used to decide whether association measures between Y and exposure are unbiased if adjusted by the PPV of A. If the test fails, we introduced a formula to adjust relative risks, and rate ratios for a rare Y, based on the same PPVs that were used for the test.

Our tools allow the application of the following strategy. When estimating the fitness-for-purpose of a data source to assess the occurrence of an event Y, a specific indicator A for Y can be sought alongside a screening indicator B. The sensitivity of a A in the population can be estimated by validating both a sample of A and a sample of B, using the methodology described in Section 3. This information can be used immediately to assess the number of cases of Y in the population and in subpopulations. Moreover, it can be used in a later stage when association studies are conducted with Y as an outcome.

If the risk ratio is the measure of association, knowledge of the sensitivity of A in the population can be used to evaluate the impact of differential misclassification: if A has low sensitivity in the population, the impact of differential sensitivity can be high, since sensitivity may become substantially higher in special populations, such as those exposed to a new medication. The knowledge of P_A, P_B, $P_{A \cap B}$ across exposure strata can be used to run a simulation of the power of our test, as done in Section 4, and to decide the sampling strategy for a validation study stratified by exposure. If the test fails, the results from the validation study can also be used to adjust the estimates.

If the association measure of interest is the risk difference, then knowledge of the sensitivity of A and the context of Y can be used to assess whether one of the scenarios of Section 5.3 applies in order to adjust for misclassification. If required by the scenario, a validation study stratified per exposure may also be conducted in this case.

A limitation of our study is that we did not provide estimates of the uncertainty around our formulas. However, the uncertainty can always be assessed through simulation-based approaches such as the bootstrap. Another limitation is that, in the simulations, we did not explore a range of options for the specificity of A and B. Such scenarios need further research.

Areas for further research could include extending the proposed method to non-rare events and hazard ratios, as well as incorporating the method in statistical modelling possibly with weighted estimation.

5.8 Conclusions

Whenever an auxiliary screening indicator can be defined to complement a primary indicator for an event, estimates of the PPV of both indicators provide tools to reduce the bias in estimating the number of cases, prevalence, cumulative incidence, rate (if the event is rare) and, in the case of association studies, risk ratio and risk difference. They also allow to test for non-differential sensitivity.

While direct estimation of the sensitivity is often infeasible, this novel methodology improves evidence based on data obtained from re-use of existing databases, which may prove critical for regulatory and public health decisions.

5.9 Appendices to Chapter 5

Interrelation between validity indices

We briefly summarize here the method introduced by Bollaerts et al. Prevalence can be defined as $\pi = TP + FN/N$, where N represents the magnitude of the study population. It can, therefore, be rewritten as a function of observed prevalence $(P = TP + FP/N)$, positive predictive value ($PPV = TP/TP + FP$), and sensitivity ($SE = TP/TP + FN$):

$$\begin{aligned}\pi &= TP + FN/N \\ &= TP + FN/N \times TP + FP/TP \times TP/TP + FP \\ &= TP + FP/N \times TP + FN/TP \times TP/TP + FP \\ &= P \times 1/SE \times PPV\end{aligned} \quad (30)$$

therefore:

$$\pi = \frac{P \times PPV}{SE} \quad (31)$$

$$SE = \frac{P \times PPV}{\pi} \quad (32)$$

$$PPV = \frac{SE \times \pi}{P} \quad (33)$$

Then, since the sum of true positives and false negatives is equal for all indicators (A, B, and AUB), that is, $TP_A + FN_A = TP_B + FN_B = TP_{AUB} + FN_{AUB} = N\pi$, the following holds:

$$SE_{AUB} = \frac{TP_{AUB}}{N\pi} = \frac{TP_A}{N\pi} + \frac{TP_B}{N\pi} - \frac{TP_{A\cap B}}{N\pi} = SE_A + SE_B - SE_{A\cap B} \quad (34)$$

then, using equation ([SE]):

$$SE_{AUB} = \frac{P_A \times PPV_A}{\pi} + \frac{P_B \times PPV_B}{\pi} - \frac{P_{A\cap B} \times PPV_{A\cap B}}{\pi} \quad (35)$$

and finally:

$$\pi = \frac{P_A \times PPV_A}{SE_{AUB}} + \frac{P_B \times PPV_B}{SE_{AUB}} - \frac{P_{A\cap B} \times PPV_{A\cap B}}{SE_{AUB}} \quad (36)$$

Deriving sensitivity from PPVs

In order to achieve an upper bound for sensitivity, equation (35) can be rewritten as:

$$SE_A = SE_{AUB} - \frac{P_B \times PPV_B}{\pi} + \frac{P_{A\cap B} \times PPV_{A\cap B}}{\pi} \quad (37)$$

then, replacing π with (36) we obtain:

$$SE_A = SE_{AUB} - \frac{P_B \times PPV_B - P_{A\cap B} \times PPV_{A\cap B}}{P_A \times PPV_A / SE_{AUB} + P_B \times PPV_B / SE_{AUB} - P_{A\cap B} \times PPV_{A\cap B} / SE_{AUB}} \quad (38)$$

if the sensitivity of $A \cap B$ is unknown, the best-case scenario estimator can be retrieved:

$$SE_A \leq 1 - \frac{P_B \times PPV_B - P_{A\cap B} \times PPV_{A\cap B}}{P_A \times PPV_A / SE_{AUB} + P_B \times PPV_B / SE_{AUB} - P_{A\cap B} \times PPV_{A\cap B} / SE_{AUB}} \quad (39)$$

and if there is no intersection among the indicators, i.e. no individual tests positive for both indicators:

$$SE_A \leq P_A \times PPV_A / P_A \times PPV_A + P_B \times PPV_B \quad (40)$$

Equations (39) can be rewritten using the absolute numbers of individuals detected by an indicator i, N_i, instead of observed prevalences:

$$SE_A \leq 1 - N_B \times PPV_B - N_{A\cap B} \times PPV_{A\cap B} / N_A \times PPV_A + N_B \times PPV_B - N_{A\cap B} \times PPV_{A\cap B} \quad (41)$$

and (40) can be rewritten as:

$$SE_A \leq N_A \times PPV_A / N_A \times PPV_A + N_B \times PPV_B \quad (42)$$

Risk difference

We illustrate multiple scenarios where the risk difference (RD) can be estimated based on P, PPV, and SE. RD is defined as:

$$RD = \pi^e - \pi^{\bar{e}} \quad (43)$$

RD can also be written as a function of observed prevalence, positive predictive value, and sensitivities; using equation (31), we obtain

$$RD = \frac{PPV_A^e}{SE_A^e} P_A^e - \frac{PPV_A^{\bar{e}}}{SE_A^{\bar{e}}} P_A^{\bar{e}} \quad (44)$$

Scenario 2: SE_A non-differential ($SE_A^e = SE_A^{\bar{e}} = SE_A$)

Assuming the non-differentiality of SE_A we can derive a lower bound for the RD, calculated using PPV and P of A across exposure strata. Equation (44) becomes:

$$RD = \frac{1}{SE_A}\left(PPV_A^e \times P_A^e - PPV_A^{\bar{e}} \times P_A^{\bar{e}}\right) \quad (45)$$

since $SE_A \in (0;1)$, the following relationships hold:

$$\begin{aligned} \text{sign}(RD) &= \text{sign}\left(PPV_A^e \times P_A^e - PPV_{\bar{A}}^e \times P_{\bar{A}}^e\right) \\ |RD| &\geq \left|\left(PPV_A^e \times P_A^e - PPV_{\bar{A}}^e \times P_{\bar{A}}^e\right)\right| \end{aligned} \quad (46) / (47)$$

Scenario 3: SE_A non-differential ($SE_A^e = SE_{\bar{A}}^e$) and $PPV_A^e = PPV_{\bar{A}}^e = PPV_A = 1$

In the case when SE_A is non-differential, the estimate does not depend on exposure strata: by replacing (39) in (45) in the case of $PPV_A^e = PPV_{\bar{A}}^e = 1$ we obtain:

$$\begin{aligned} \text{sign}(RD) &= \text{sign}\left(P_A^e - P_{\bar{A}}^e\right) \\ |RD| &\geq \left|\frac{P_A + P_B PPV_B - P_{A\cap B}}{P_A - 2P_{A\cap B}}\left(P_A^e - P_{\bar{A}}^e\right)\right| \end{aligned} \quad (48)/(49)$$

If $A \cap B = \emptyset$, then:

$$|RD| \geq \left|\frac{P_A + P_B PPV_B}{P_A}\left(P_A^e - P_{\bar{A}}^e\right)\right| \quad (50)$$

Scenario 4: $SE_{A\cup B}$ non-differential ($SE_{A\cup B}^e = SE_{\bar{A}\cup B}^e = SE_{A\cup B}$)

Assuming the non-differentiality of $SE_{A\cup B}$, we can derive a lower bound for the RD. Substituting (36) in equation (43) we obtain:

$$\begin{aligned} RD = &\left(\frac{P_A^e \times PPV_A^e}{SE_{A\cup B}^e} + \frac{P_B^e \times PPV_B^e}{SE_{A\cup B}^e} - \frac{P_{A\cap B}^e \times PPV_{A\cap B}^e}{SE_{A\cup B}^e}\right) \\ &-\left(\frac{P_{\bar{A}}^e \times PPV_{\bar{A}}^e}{SE_{\bar{A}\cup B}^e} + \frac{P_{\bar{B}}^e \times PPV_{\bar{B}}^e}{SE_{\bar{A}\cup B}^e} - \frac{P_{\bar{A}\cap B}^e \times PPV_{\bar{A}\cap B}^e}{SE_{\bar{A}\cup B}^e}\right) \end{aligned} \quad (51)$$

and, if we assume the non-differentiality of $SE_{A\cup B}$, (51) simplifies as:

$$\begin{aligned} RD = \frac{1}{SE_{A\cup B}} &\left[\left(P_A^e \times PPV_A^e + P_B^e \times PPV_B^e - P_{A\cap B}^e \times PPV_{A\cap B}^e\right)\right. \\ &\left.-\left(P_{\bar{A}}^e \times PPV_{\bar{A}}^e + P_{\bar{B}}^e \times PPV_{\bar{B}}^e - P_{\bar{A}\cap B}^e \times PPV_{\bar{A}\cap B}^e\right)\right] \end{aligned} \quad (52)$$

from (52) we obtain the sign of RD. In addition, it is possible to derive a lower bound for RD:

$$\begin{aligned} |RD| \geq &\left|\left(P_A^e \times PPV_A^e + P_B^e \times PPV_B^e - P_{A\cap B}^e \times PPV_{A\cap B}^e\right)\right. \\ &\left.-\left(P_{\bar{A}}^e \times PPV_{\bar{A}}^e + P_{\bar{B}}^e \times PPV_{\bar{B}}^e - P_{\bar{A}\cap B}^e \times PPV_{\bar{A}\cap B}^e\right)\right| \end{aligned} \quad (53)$$

If $SE_{A\cup B} = 1$:

$$\begin{aligned} |RD| = &\left|\left(P_A^e \times PPV_A^e + P_B^e \times PPV_B^e - P_{A\cap B}^e \times PPV_{A\cap B}^e\right)\right. \\ &\left.-\left(P_{\bar{A}}^e \times PPV_{\bar{A}}^e + P_{\bar{B}}^e \times PPV_{\bar{B}}^e - P_{\bar{A}\cap B}^e \times PPV_{\bar{A}\cap B}^e\right)\right| \end{aligned} \quad (54)$$

Scenario 5: positive predictive value of A greater than sensitivity in the exposed group ($PPV_A^e > SE_A^e$)

If A is chosen to have a very large PPV, but not necessarily 1, we can still explore the scenario when it can be assumed that $PPV \geq SE$. If we can assume that $PPV_A^e \geq SE_A^e$, formula (44) becomes:

$$RD \geq P_A^e - \frac{PPV_A^{\tilde{e}}}{SE_A^{\tilde{e}}} P_A^{\tilde{e}} \quad (55)$$

and, using formula (56) and formula (55):

$$SE_A^{\tilde{e}} = P_A^{\tilde{e}} \times PPV_A^{\tilde{e}} / \pi = P_A^{\tilde{e}} \times PPV_A^{\tilde{e}} / 1/SE_{A \cup B} \left(P_A^{\tilde{e}} \times PPV_A^{\tilde{e}} + P_B^{\tilde{e}} \times PPV_B^{\tilde{e}} - P_{A \cap B}^{\tilde{e}} \times PPV_{A \cap B}^{\tilde{e}} \right) \quad (56)$$

thus, replacing ([se_k]) in formula ([RSgeq]) we obtain:

$$RD \geq P_A^e - PPV_A^{\tilde{e}} \times P_A^{\tilde{e}} \times 1/SE_{A \cup B} \left(P_A^{\tilde{e}} \times PPV_A^{\tilde{e}} + P_B^{\tilde{e}} \times PPV_B^{\tilde{e}} - P_{A \cap B}^{\tilde{e}} \times PPV_{A \cap B}^{\tilde{e}} \right) / P_A^{\tilde{e}} \times PPV_A^{\tilde{e}} \quad (57)$$

and finally:

$$RD \geq P_A^e - 1/SE_{A \cup B} \left(P_A^{\tilde{e}} \times PPV_A^{\tilde{e}} + P_B^{\tilde{e}} \times PPV_B^{\tilde{e}} - P_{A \cap B}^{\tilde{e}} \times PPV_{A \cap B}^{\tilde{e}} \right) \quad (58)$$

where $1/SE_{A \cup B} \geq 1$.

Test statistic

The null hypothesis of the non-differentiality test $H_0: SE_A^e = SE_A^{\tilde{e}}$ can be rewritten as:

$$H_0: \frac{SE_A^e}{SE_A^{\tilde{e}}} - 1 = 0 \quad (59)$$

By exploiting the interrelationships between validation incidences, the SE of the indicator A can be rewritten as a function of observed prevalence and PPV of the indicators A, B and $A \cap B$ and SE of the indicator AUB. Using equation (32):

$$SE_A = \frac{P_A \times PPV_A}{\pi} \quad (60)$$

then, using equation (36):

$$SE_A = \frac{P_A \times PPV_A}{\frac{P_A \times PPV_A}{SE_{AUB}} + \frac{P_B \times PPV_B}{SE_{AUB}} - \frac{P_{A \cap B} \times PPV_{A \cap B}}{SE_{AUB}}} \quad (61)$$

thus, the test statistic can be rewritten as:

$$t = \frac{P_A^e \times PPV_A^e / P_A^e \times PPV_A^e + P_B^e \times PPV_B^e - P_{A\cap B}^e \times PPV_{A\cap B}^e}{P_{\bar{A}}^{\bar{e}} \times PPV_{\bar{A}}^{\bar{e}} / P_{\bar{A}}^{\bar{e}} \times PPV_{\bar{A}}^{\bar{e}} + P_{\bar{B}}^{\bar{e}} \times PPV_{\bar{B}}^{\bar{e}} - P_{\bar{A}\cap B}^{\bar{e}} \times PPV_{\bar{A}\cap B}^{\bar{e}}} \times \frac{SE_{AUB}^e}{SE_{\bar{A}UB}^{\bar{e}}} - 1 \quad (62)$$

and finally, if $SE_{AUB}^{E} = SE_{AUB}^{\bar{E}}$:

$$t = \frac{P_A^e PPV_A^e \left(P_{\bar{A}}^{\bar{e}} PPV_{\bar{A}}^{\bar{e}} + P_{\bar{B}}^{\bar{e}} PPV_{\bar{B}}^{\bar{e}} - P_{\bar{A}\cap B}^{\bar{e}} PPV_{\bar{A}\cap B}^{\bar{e}}\right)}{P_{\bar{A}}^{\bar{e}} PPV_{\bar{A}}^{\bar{e}} \left(P_A^e PPV_A^e + P_B^e PPV_B^e - P_{A\cap B}^e PPV_{A\cap B}^e\right)} - 1 \quad (63)$$