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

Task 7.7: Statistical analysis plan (SAP) for generation and quality assessment of the list of completed and ongoing pregnancies in population based data sources

Version 5.0

Document history

	Version	Date	Edits
Rosa Gini (ARS)	0.1	12 April 2021	First draft
Giorgio Limoncella, Claudia Bartolini (ARS)	0.2	13 April 2021	Itemsets and Meaning of survey update
Rosa Gini (ARS)	0.3	13 April 2021	Further inclusions
RG and GL (ARS)	0.5	19 April 2021	Incorporating Romin comments, adding details to analysis, adding figures, moved appendix to excel format
RG	0.6	20 April 2021	created links to Google tables, completed draft
RG, GL, CB	0.7	21 April 2021	minor changes to T2.2
CB	0.8	22 april 2021	Minor changes in tables
working group	0.9	23 April 2021	Split conceptsets of 'birth' and of 'gestational age'
RG	1.0	25 April 2021	finalised draft to be submitted to DAPs
RG	1.2	5 May 2021	Incorporated DAPs' comments, improved reconciliation algorithm, included graphical representations of quality of pregnancy record
CB	1.3	6 May 2021	Updated pregnancy_end_date
CB, GL	1.4	12 May 2021	Added new concept sets
RG	1.5	16 May 2021	Added dummy table section
RG, GL, CB	1.6	4 Jun 2021	Incoporated case when CONCEPTSET is found in primary care. Filled out dummy table section
RG	2.0	4 Jun 2021	Released for new round of comments
RG, AG, CB, GL, GR	3.0	4 April 2022	Started update AG: revise introduction, add ongoing pregnancies (timely detection) and sensitivity of this approach; update

			references AG/GL: write a rebuttal documents for coauthors' comments GL: update description of reconciliation, including datasource-specific algorithms; update shell tables; add link to GitHub
GL, HN	3.0	7 July 2022	updated reconciliation: new dap-specific parameters inserted
GL, AG	4.0	Apr 2023	Update of Table 1 (inclusion of Epichron, KI, HSD),
GL, AG	5.0	August 2023	Updated reconciliation parameters and added predictive model
GL, HN, LM	5.0.1		DAP Norway adaption

Table of contents

1. Authors

Rosa Gini, Claudia Bartolini, Giorgio LImoncella, Anna Girardi, ARS

Carlos E. Duran, Vjola Hoxhaj, Caitlin Dodd, Miriam Sturkenboom, UMCU

Romin Pajouheshnia, UU

Tania Schink, BIPS

Consuelo Huerta, Ana Llorente Garcia, Patricia García Poza, Mar Martín Pérez, AEMPS

ConcePTION's DAPs:

Norway: Hedvig Nordeng, Luigi Maglanoc

2. Background and objectives

2.1 Background

ConcePTION aims to build an ecosystem that can use observational data sources to generate Real World Evidence (RWE) that may be used for clinical and regulatory decision making. Real world evidence is required to address the big information gap of medication safety in pregnancy.

Data sources participating in multi-database studies, such as the studies conducted in the WP1 of ConcePTION, do not necessarily come with a complete enumeration of the pregnancies experienced in their underlying populations. Data sources are different in terms of compositions of data banks[1], and many data banks collect information that is pertinent to the purpose of identifying pregnancies: birth registries, primary care medical records, hospital administrative records, termination registries, and others. All combinations of such data banks are documented among the data sources of ConcePTION, see the Deliverable 7.5 of ConcePTION (Dodd C et al, 2020), and Table 1 below. Therefore, in order to identify the list of pregnancies, algorithms that only retrieve pregnancies using diagnostic codes, or that process records independently of their origin, are suboptimal (Matcho A et al, 2018; Sarayani A , et al, 2020). On the other hand, algorithms that only retrieve pregnancies from birth registries may fail to detect pregnancies that end prematurely, and fail to exploit the full range of information that may be available from primary care medical records (Ortiz SS et al, 2020) or records of specialist visits (Schink T et al, 2020). 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). A recent work, focused on a single national primary care database, addressed the issue of pregnancy episodes with no outcomes or conflicting records which normally are excluded from the algorithm generating the pregnancy register (Campbell J et al., 2022). Here researchers provided specific potential scenarios for uncertain pregnancy episodes (recorded outcome missing or conflicting records), which accounted for one out of three pregnancy episodes detected by the algorithm. Most of the pregnancy records with missing outcome were likely to be true pregnancies with outcomes not captured in the database and such pregnancies would be missed if episodes with recorded outcome missing were excluded by the algorithm. Authors also pointed out that pregnancies ending in miscarriage were more likely to have conflicting records and excluding such uncertain episodes may lead to an underestimation of miscarriage as a pregnancy outcome; however, specific guidelines must be defined for the inclusion of uncertain episodes.

To incorporate all such scenarios in a transparent and common framework, a novel algorithm must be designed. Such an algorithm may also enable estimation of the moment when a completed pregnancy had left its first identifiable sign in the data source, allowing the extraction of information on pregnancies also while they are ongoing or end prematurely, even if their outcome is unknown. As a matter of fact, including only completed pregnancies will result in both an underestimation of the actual number of pregnancies and potentially exclude periods when a woman is pregnant.

The ConcePTION algorithm for pregnancies will be tailored both to capture completed pregnancies and to detect ongoing pregnancies at early stages. This will enable the identification of pregnancies with premature end and allow timely investigation of emerging issues (new drugs/vaccines).

Institutions from multiple studies beyond ConcePTION are contributing to this study: the EMA/2017/09/PE/04 retinoids project (Sturkenboom M et al, 2019), the ACCESS Study (Sturkenboom M et al, 2020), the CONSIGN Study (Nordeng H et al, 2021).

[1] The definition of data bank and data source are contained in (Thurin, CPT 2021) and included in the Glossary of this document

This Statistical Analysis Plan builds on section 5.4.2 of the Data Characterisation Protocol of the ConcePTION Project, version 1.0.

2.2 Objective

The objectives of this study are

- **Objective a:** to define, implement, and test an algorithm to identify pregnancies that have been completed in the data sources contributing to this study. The algorithm will be structured in standard components, each associated with a combination of data banks used to retrieve the information
- **Objective b:** to characterise the subpopulations of pregnancies identified by the algorithm.
- **Secondary objective:** predict when a completed pregnancy would have been first detected in each data source, depending on the component and the characteristics of the pregnancy

2.3 Strategy

The algorithm will comprise standard components: for instance, pregnancy retrieved from a birth registry, or pregnancies retrieved from a hospital administrative data bank using a diagnostic code associated to delivery. Composition of such components will be tailored to each data source in a hierarchical manner: e.g. in order to associate date of start of pregnancy to a pregnancy, estimate from ultrasound recorded in a birth registry may be deemed more reliable than diagnostic codes of pregnancy-related events retrieved from a hospital administrative data bank. Pregnancies derived from all available data banks will be included, and consequently, the population of identified pregnancies will be comprised of several subpopulations. For instance, pregnancies whose start and end date are obtained by sources other than registries, despite often having a lower degree of certainty, will be identified and be available for inclusion in the main analysis or in sensitivity analyses. In some data sources, according to their composition in data banks and to the information recorded in the specific data bank, a smaller range of options may be available.

3. Methods

3.1 Study design

This is a descriptive study of a cohort. The preliminary part of the study is the design and testing of the algorithms aimed at defining the cohort of pregnancies.

3.2 Data sources

The Data Access Providers (DAPs) participating in the study are listed in Table 1, along with the data sources they have access to. The families of data banks[2] composing each data source are also presented, limited to those data banks that are used in this study: birth registry, termination registry, spontaneous abortion registry, congenital anomaly registry, hospital administrative records, primary care medical records. Records for birth registry are prompted by a delivery or by a home visit. For hospital administrative records, record of a diagnosis is prompted by discharge from a hospital admission, from a specialist visit, or from both. The data sources are described more in detail in an annex of the Deliverable 7.5 (Dodd C et al, 2020) and in the ConcePTION Data Catalogue (Swertz et al, 2021).

[2] The definition of “families of data banks” is clarified in the Glossary

Table 1. Families of data banks and prompts enclosed in the data sources participating in the study.

		Data banks enclosed in the data source to be used in this study								
family of data bank		birth registry		termination registry	spontaneous abortion registry	congenital anomaly registry	hospital administrative records		primary care medical records	other
prompt of the data bank		live birth or stillbirth	first home visit				discharge from hospitalisation	specialist visit (with diagnosis)		
DAP	**Data source**									
University of Oslo (UOSL)	Norwegian Health Data	X					X	x	x	
Aarhus University (AARHUS)	Danish Health Data	X					X	X		
University of Ulster (ULST)	EUROmediCAT					X				
University Hospital in Toulouse (CHUT)	EFEMERIS		X	X		X	X			
Bordeaux University	SNDS						X			
PHARMO Institute	PHARMO		X				X		X	
Leibniz Institute for Prevention Research and Epidemiology (BIPS)	GePaRD						X		x	
FISABIO	Valencia database	X		X		X	X			

Information System for Research in Primary Care (IDIAP)	SIDIAP		X				X		X	
University of Ferrara (FERR)	EMILIA ROMAGNA	X				X	X			
CNR-IFC	CATuscany	X				X	X			
ARS Toscana	ARS	X		X	X		X			
MESSINA	CASERTA	X					X			
Finnish Institute of Health and Welfare (THL)	THL	X		X		X	X	X		administrative data on primary care visits
Univeristy of Swansea (USWAN)	SAIL	X			X		X	X	X	
AEMPS	BIFAP						X		X	
GSK	CPRD_GOLD								X	
Tables of the ConcePTION CDM fed by such data banks		SURVEY_ID SURVEY OBSERVATIONS				EUROCAT SURVEY_ID SURVEY OBSERVATIONS		EVENTS PROCEDURES MEDICAL OBSERVATIONS		

DAP: Data Access Provider; CDM: Common Data Model

3.3 Common Data Model and Quality Control

To implement the study, all data sources will be mapped to the ConcePTION Common Data Model v2.2 (ConcePTION CDM v2.2), according to an Extraction, Transformation and Load (ETL) procedure whose design is documented in standard ETL documents and is available in the ConcePTION Catalogue (Swertz et al, 2021). The last row of Table 1 shows the destination of the included data banks: birth registry, termination registry, spontaneous abortion registries are loaded to the SURVEY_ID and SURVEY_OBSERVATIONS tables; the congenital anomaly registry is loaded to the EUROCAT, SURVEY_ID and SURVEY_OBSERVATIONS tables; hospital administrative records and primary care medical records are loaded to the EVENTS, PROCEDURES and MEDICAL_OBSERVATIONS tables. Pregnancies occur in the vast majority of cases in persons of female gender, however transgender men or non-binary persons may be pregnant, too. In the ConcePTION CDM the gender of the person is recorded at the time of creation of data, and classified as male, female, other, undetermined, or unknown, and this will be included in the analysis.

The ConcePTION CDM ensures that the origin of each record can be retrieved during data processing, since each table contains a *meaning* variable: such variables are associated with the origin data bank and the prompt.[3]

Quality control will be ensured by the Level 1 and 2 checks of the ConcePTION Data Characterisation Study.

3.4 Data processing for algorithm design

3.4.1 Overall design

Subjects are first selected as experiencing the end of a pregnancy or an ongoing pregnancy. The selection is done in parallel from four streams (note that in a data source only a subset of the streams may be activated)

stream PROMPTS: prompts of birth registries, terminations registries, and spontaneous abortion registries in SURVEY_ID: the existence of one of such record implies readily that a pregnancy has ended

stream EUROCAT: records of the EUROCAT table

stream CONCEPTSETS: diagnostic codes from the EVENTS or procedure codes from the PROCEDURES or codes from the MEDICAL_RECORDS file referring to an end or an ongoing pregnancy

stream ITEMSETS: variables from routinely collected healthcare data that are only populated when a woman is pregnant

For each record, the start of the corresponding pregnancy is inferred from existing information:

- in stream PROMPTS, the start is obtained from the corresponding record of the registry, as registered in the SURVEY_OBSERVATIONS file;
- in stream EUROCAT, the start is obtained from the corresponding record of the Congenital Anomaly registry, as registered in the EUROCAT file;
- in stream CONCEPTSETS, start of pregnancy is obtained using methods described in Matcho et al, 2020, or methods recommended by the DAP;
- in stream ITEMSETS, the method will be tailored to the itemset.

[3] The definition of prompt is contained in Deliverable 7.7 (Swertz et al, 2021).

The resulting sets of pregnancies for a same subject are then compared with each other in order to identify which of them could be attributable to a single pregnancy episode recorded on multiple occasions.

In the following example, three pregnancies are identified for the same subject from different streams

person_id	start_date_pregnancy	end_date_pregnancy	Stream	Conceptset	meaning	meaning_of_start_date
P01	2015-11-06	2016-08-01	EVENTS	ongoing_pregnancy	specialist_visit	timing_of_prenatal_care
P01	2015-11-15	2016-07-17	PROMPTS		birth_registry	GESTAGE_FROM_USOUNDS_DAYS
P01	2019-02-28	2019-12-19	PROMPTS		birth_registry	GESTAGE_FROM_USOUNDS_DAYS

The first two records would need to be resolved due to overlapping dates. The DAP-specific hierarchy of components would be used to determine which dates are preferred: in this case, the first record was detected during a prenatal specialist visit and start of pregnancy was assigned in an automated fashion; in the second record, the birth registry had a record of gestational age obtained from ultrasound. Clearly, the two pregnancies are the same, and the second start date has a better quality. The third record, instead, is a different pregnancy of the same woman. The final result will therefore be

pregnancy_id	person_id	start_date_pregnancy	end_date_pregnancy	meaning_of_start_date
P01_01	P01	2015-11-25	2016-08-17	GESTAGE_FROM_USOUNDS_DAYS
P01_02	P01	2019-02-28	2019-12-19	GESTAGE_FROM_USOUNDS_DAYS

Possible inconsistencies in the results from the previous steps (e.g. pregnancies overlapping, or non-premature birth following after 2 months a premature birth) are classified based on the expertise of the research team. A decision algorithm will be defined to resolve each of the inconsistencies. The final decision algorithm is decided after a round of testing.

Pregnancies whose inconsistencies fail to be resolved are recorded in a separate dataset.

The overall design is represented in Figure 1.

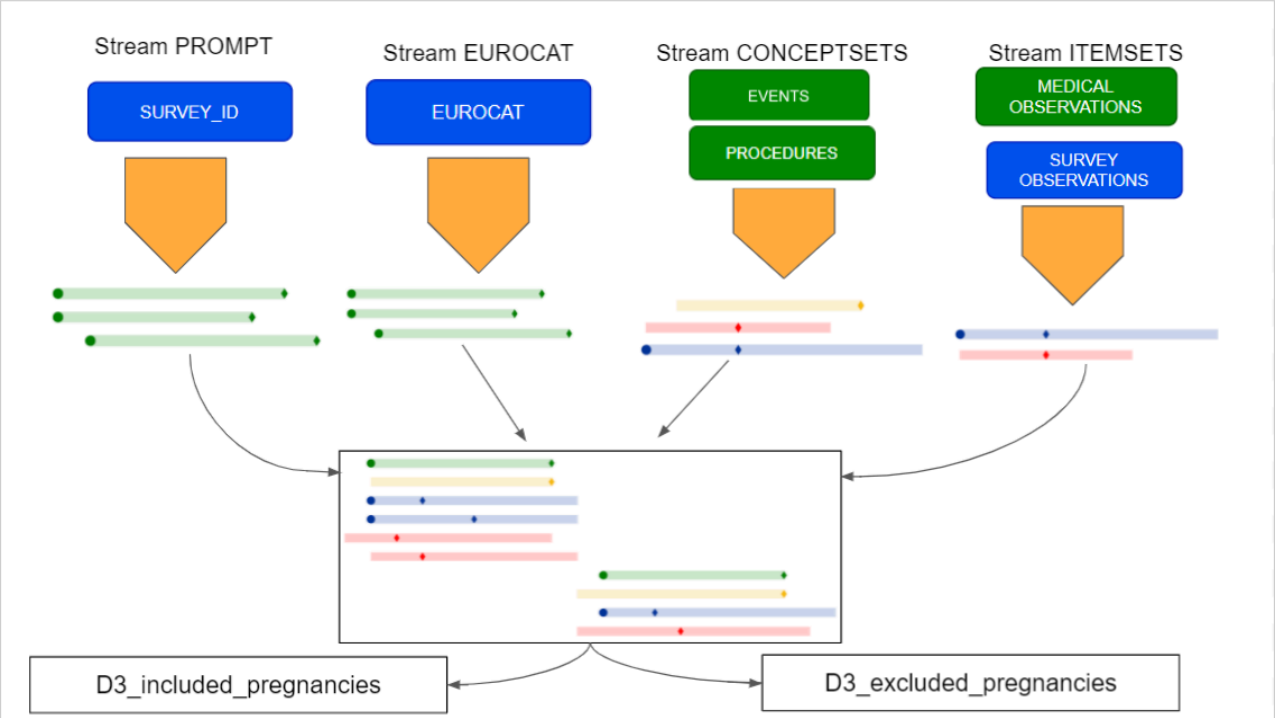


Figure 1. Flow of the overall design. In the graphical representation, a diamond represents the date of the record, a circle represents a record of a date of start of pregnancy, and the bar represents the interval between start and end. When the diamond is not at the end of the bar, it means that the 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 bar represents the quality of the records: green if there is no imputation, yellow if only start date is imputed, red if both start and end date are imputed; the blue bar represents a case when the record is recorded while the pregnancy is ongoing, therefore the end of pregnancy is imputed, but the start of pregnancy is available.

3.4.2 Inclusion criteria

Inclusion criteria are the following

Stream PROMPTS

Prompts aimed at identifying pregnancies for stream PROMPTS from SURVEY_ID are identified from the following lists of meanings recorded in survey_meaning

- *live birth or still birth:* birth_registry_mother, 8days_CHUT_questionnaire, childhealth_mother_registry, anomalies_baby_registry, congenital_anomalies
- *ongoing pregnancy:* primary_care_pregnancy_register
- *induced termination:* induced_termination_registry or birth_registry_mother TOPFA
- *spontaneous abortion:* spontaneous_abortion_registry
- *estimated_type_and_period_of_pregnancy:* this is a reconstituted registry from the French National Healthcare System that is created based on routine healthcare and accessed by the University of Bordeaux as such

In the Table 2 are presented the list of meanings associated to each outcome in each ETL document, as documented in the corresponding ETL documents.

https://docs.google.com/spreadsheets/d/1oh3N2PBCjKw-uj6UhKdvzLzCE-yEi4keEnJy0Jb7dbg/edit#gid=688046871

Table 2. List of meanings that are inclusion criteria for the stream PROMPTS

DAP	live birth or still birth	ongoing pregnancy	induced termination	spontaneous abortion	other
UOSL	birth_registry_mother	*none*	*none*	*none*	
AARHUS	birth_registry	*none*	*none*	*none*	
ULST					
CHUT	pregnancy_characteristics	*none*	pregnancy_characteristics	pregnancy_characteristics	
Bordeaux	estimated_type_and_period_of_pregnancy	none	estimated_type_and_period_of_pregnancy	estimated_type_and_period_of_pregnancy	
PHARMO					
BIPS	algorithm_pregnancy	*none*	algorithm_pregnancy	algorithm_pregnancy	
FISABIO	birth_registry_mother birth_registry_child	*none*	*none*	*none*	
SIDIAP	birth_registry	*none*	induced_termination_registry	spontaneous_abortion_registry	
FERR					
CNR-IFC	birth_registry_mother birth_registry_child	*none*	*none*	*none*	
ARS	birth_registry_mother	*none*	induced_termination_registry	spontaneous_abortion_registry	
MESSINA	birth_registry_mother	*none*	induced_termination_regist	spontaneous_abortion_registry	

			ry		
THL	birth_registry	*none*	induced_termination_registry	*none*	
USWAN	ONS birth registry				
AEMPS	algorithm_pregnancy	*none*	algorithm_pregnancy	algorithm_pregnancy	

in USWAN, access to data on induced terminations and spontaneous abortion is being sought

Stream CONCEPTSETS

In the Stream CONCEPTSETS, we query the EVENTS file for diagnostic codes (actually they may be found also in MEDICAL_OBSERVATIONS and SURVEY_OBSERVATIONS), and PROCEDURES files for procedure codes.

To query the EVENTS file, SNOMED codes have been extracted from Matcho et al, 2018, mapped to ICD10CM, ICD9CM, ICPC2P, RCD, SNOMED and RCD2 using Codemapper (Becker et al, 2017), and further refined by data partners according to local expertise. After the revision, the final list of concepts is the following:

event definition

Gestation less than 24 weeks - type of end unknown

24weeks pregnancy - type of end unknown

Gestation 25 26 weeks - type of end unknown

Gestation 27 28 weeks - type of end unknown

Gestation 29 30 weeks - type of end unknown

Gestation 31 32 weeks - type of end unknown

Gestation 33 34 weeks - type of end unknown

Gestation 35 36 weeks - type of end unknown

Gestation 37 weeks - type of end unknown

Gestation less than 24 weeks - type of end live birth

24weeks pregnancy - type of end live birth

Gestation 25 26 weeks - type of end live birth

Gestation 27 28 weeks - type of end live birth

Gestation 29 30 weeks - type of end live birth

Gestation 31 32 weeks - type of end live birth

Gestation 33 34 weeks - type of end live birth

Gestation 35 36 weeks - type of end live birth

Gestation 37 weeks - type of end live birth

Ongoing Pregnancy

Start of Pregnancy

Gestational diabetes

Fetal growth restriction

Preeclampsia

Pregnancy Bleeding

Birth classified as possible

Birth classified as narrow

At term delivery

Preterm Birth

Postterm delivery

Stillbirth

Induced Termination

Spontaneous abortion

Ectopic pregnancy

Gestation less than 24 weeks - child*

24weeks pregnancy - child*

Gestation 25 26 weeks - child*

Gestation 27 28 weeks - child*

Gestation 29 30 weeks - child*

Gestation 31 32 weeks - child*

Gestation 33 34 weeks - child*

Gestation 35 36 weeks - child*

Gestation 37 weeks - child*

*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

- Start of pregnancy,
- Gestational age:
    - Gestation_less24
    - Gestation_24
    - Gestation_25_26
    - Gestation_27_28
    - Gestation_29_30
    - Gestation_31_32
    - Gestation_33_34
    - Gestation_36_35
    - Gestation_more37
- Ongoing pregnancy,
- Birth,
    - pre-term birth
    - at_term_birth
    - post_term_birth
    - live birth

    - still birth
- Elective Termination,
- Spontaneous abortion,
- Ectopic pregnancy,
- Molar pregnancy

The list of codes included in each conceptset is available here.

The conceptsets used to extract useful procedures from PROCEDURES are listed in Table 3

https://docs.google.com/spreadsheets/d/1oh3N2PBCjKw-uj6UhKdvzLzCE-yEi4keEnJy0Jb7dbg/edit#gid=1973613615

Table 3. List of conceptsets of procedures executed in pregnancy and recorded in the PROCEDURE file, that is to be used in Stream CONCEPTSETS

Stream EUROCAT

All records in EUROCAT are included

Stream ITEMSETS

The itemsets used to extract records identifying pregnancies from MEDICAL_OBSERVATIONS are listed in Table 4

https://docs.google.com/spreadsheets/d/1oh3N2PBCjKw-uj6UhKdvzLzCE-yEi4keEnJy0Jb7dbg/edit#gid=178085618

Table 4. List of itemsets from MEDICAL_OBSERVATIONS to be used as inclusion criteria in the Stream ITEMSETS

3.4.3 Study variables

Process streams separately

In this step, each record of pregnancy extracted using the inclusion criteria is associated with a date of start of pregnancy and a date of end of pregnancy. The step is different according to the stream.

Stream PROMPTS

In the Stream PROMPTS, the following sequence of steps will be implemented and recorded in table D3_Stream_PROMPTS (see section 4.3.2)

- all the available information about end of pregnancy (Table 5) and type of end of pregnancy (Table 6) will be associated with each record; the information is then used to fill the pregnancy_end_date and meaning_of_end_of_pregnancy variables, in a hierarchical manner as follows
    - The variable “pregnancy_end_date” will be generated in the dataset “dataset_pregnancies”. This variable will be initially filled with the value given in the DATEENDPREGNANCY itemset, if DATEENDPREGNANCY is not available for the record it will be filled by the other itemsets,

in order: END_LIVEBIRTH, END_STILLBIRTH, END_INTERRUPTION or END_SPONTANOUS_ABORTION.

  - At the same time, the variable “meaning_of_end_of_pregnancy” will be created and stored in “dataset_pregnancies”, and it will contain the step of the hierarchy above where the information has been found
  - At the end, the variable “type_of_end_of_pregnancy” is generated, which will correspond to the itemset “type”, when this is available, otherwise it will be filled with the content of the variable ‘survey_meaning’

- the information about start of pregnancy (Table 6) will be associated with each record; if a column is missing, the corresponding variable will be missing as well; the columns will be used to compute the start of pregnancy. The variable “pregnancy_start_date” will be generated and stored in “dataset_pregnancies”, and it will take the value of the itemset DATESTARTPREGNANCY when this is available. Otherwise “pregnancy_start_date” will be estimated using the gestational ages stored in the others itemset, according to the hierarchy is listed below:

  1. DATESTARTPREGNANCY
  2. GESTAGE_FROM_DAPs_CRITERIA_DAYS
  3. GESTAGE_FROM_DAPs_CRITERIA_WEEKS
  4. GESTAGE_FROM_USOUNDS_DAYS
  5. GESTAGE_FROM_USOUNDS_WEEKS
  6. GESTAGE_FROM_LMP_DAYS
  7. GESTAGE_FROM_LMP_WEEKS

the step of the hierarchy that generated the best estimate will be stored in the variable meaning_of_start_of_pregnancy

- Note that some information may be recorded as information associated to the child, and not to the pregnancy; if this is the case, information across multiple records of pregnancies with more than one child must be summarised (eg end of pregnancy may be obtained as the maximum between the birth dates of the children) across children’s records

The itemsets used to extract start of pregnancy from SURVEY OBSERVATIONS are listed in Table 5, as documented in the corresponding ETL documents

ConcePTION: tables for pregnancy SAP

Table 5. List of itemsets associated with start of pregnancies in SURVEY_OBSERVATIONS, that is to be used in Stream PROMPTS to associate to each record of SURVEY_ID their corresponding date of start of pregnancy

The itemsets used to extract end of pregnancy from SURVEY OBSERVATIONS are listed in Table 6, as documented in the corresponding ETL documents

https://docs.google.com/spreadsheets/d/1oh3N2PBCjKw-uj6UhKdvzLzCE-yEi4keEnJy0Jb7dbg/edit#gid=1906577145

Table 6. List of itemsets associated with end of pregnancies in SURVEY_OBSERVATIONS, that is to be used in Stream PROMPTS

The itemsets used to extract type of end of pregnancy from SURVEY OBSERVATIONS are listed in Table 7, as documented in the corresponding ETL documents

https://docs.google.com/spreadsheets/d/1oh3N2PBCjKw-uj6UhKdvzLzCE-yEi4keEnJy0Jb7dbg/edit#gid=1906577145

Table 7. List of itemsets associated with type of end of pregnancies in SURVEY_OBSERVATIONS, that is to be used in Stream PROMPTS

In addition, a PROMPT dataset will be created from the PERSON_RELATIONSHIP table. A pregnancy episode will be generated using the meaning implying the relationship between mother and child. The end date of the pregnancy will correspond to the date of birth of the child, while the beginning will be imputed 280 days earlier.

Stream EUROCAT

In the Stream EUROCAT, the following sequence of steps will be implemented and stored in Table D3_Stream_EUROCAT (see section 4.3.2)

- information about end and start of pregnancy will be obtained, respectively, from variables 'birthdate' and 'gestlength', and about type of end of pregnancy from 'type'
- Note that some information may be recorded as information associated to the newborn, and not to the pregnancy; if this is the case, pregnancies with more than one newborns have multiple records. If this is the case, information about the pregnancy must be summarised across children's records, eg end of pregnancy may be obtained as the maximum between the birth dates of the children

Stream CONCEPTSETS

In the Stream CONCEPTSETS, to each of the pregnancies extracted from EVENTS or other files using the conceptsets, start and end of pregnancy is associated using the algorithms listed in Table 8 suggested by DAPs. The algorithms may cover the case of ongoing pregnancies. The output will be stored in table D3_Stream_ CONCEPTSETS (see section 4.3.2)

https://docs.google.com/spreadsheets/d/1oh3N2PBCjKw-uj6UhKdvzLzCE-yEi4keEnJy0Jb7dbg/edit#gid=1648347487

Table 8. Algorithms recommended by each DAP to retrieve end of pregnancy and start of pregnancy in Stream CONCEPTSETS.

In data sources whose DAPs do not suggest a specific algorithm, or only suggest an algorithm in specific cases, the algorithms from Matcho A et al, 2018 are used to impute the pregnancy start and end date, and implemented as follows

- if the pregnancy has ended, then the start date is imputed with the following rules[4]:

  if type_end_of_pregnancy=='LB' and CONCEPTSET == 'pre-term birth' then pregnancy_start_date = pregnancy_end_date – 245

[4] See supplementary file pone.0192033.s011.docx, https://doi.org/10.1371/journal.pone.0192033.s011

else if outcome.type=='LB' and CONCEPTSET == 'live birth' then pregnancy_start_date = pregnancy_end_date – 280

else if outcome.type=='SB' then pregnancy_start_date = pregnancy_end_date – 280

else if outcome.type=='SA' then pregnancy_start_date = pregnancy_end_date – 70

else if outcome.type=='ECT' then pregnancy_start_date = pregnancy_end_date – 55

else if outcome.type=='T' then pregnancy_start_date = pregnancy_end_date – 70

- if the pregnancy is ongoing and has a start date but has no end, then at term end of the pregnancy is assumed for the imputation, and

  pregnancy_end_date = pregnancy_start_date + 280,

- if the pregnancy is ongoing[5] and has not a start nor a end date, then at term end of the pregnancy is assumed, and

  pregnancy_start_date = date_record– 55,

  pregnancy_end_date = pregnancy_start_date + 280,

Stream ITEMSETS

In the Stream ITEMSETS, to each of the pregnancies extracted from MEDICAL_OBSERVATIONS using the itemsets as inclusion criteria, start and end of pregnancy is associated using the algorithms listed in Table 9 suggested by DAPs and stored in Table D3_Stream_ITEMSETS (see section 4.3.2). The algorithms may cover the case of ongoing pregnancies.

https://docs.google.com/spreadsheets/d/1oh3N2PBCjKw-uj6UhKdvzLzCE-yEi4keEnJy0Jb7dbg/edit#gid=1648347487

Table 9. Algorithms recommended by each DAP to retrieve end of pregnancy and start of pregnancy in Stream ITEMSETS.

Merge streams of the same person

The four tables D3_Stream_PROMPTS, D3_Stream_EUROCAT, D3_Stream_CONCEPTSETS and D3_Stream_ITEMSETS will be first analysed in terms of internal consistency: each table will be linked with PERSONS and spells of OBSERVATION_PERIODS, to compute binary variables

1) pregnancy_with_record_date_out_of_range: records whose record date is outside the dates documented in the instance of the data source, as described in the metadata table INSTANCE.
2) not linked with PERSONS
3) person not in fertile age (between 12 and 55) at start of pregnancy
4) person not in OBSERVATION_PERIODS at record date

[5] See supplementary file pone.0192033.s009.docx, https://doi.org/10.1371/journal.pone.0192033.s009

Female gender is not a requirement, see above section 3.3. Gender as collected from the PERSON table is included as a variable.

The four tables will then be appended with one another to reconcile pregnancies that are in fact the same and stored in table D3_groups_of_pregnancies (see 4.3.3) and then in the final D3s (see 4.3.1).

The reconciliation procedure will run as follows

A) An ordering of quality of records is established and stored in variable order_quality; records are of

- quality green if both pregnancy_start_date and pregnancy_end_date are recorded;
- quality yellow if pregnancy_end_date is recorded as the record date and pregnancy_start_date is imputed[6]
- quality blue if pregnancy_start_date is recorded and pregnancy_end_date is imputed;
- quality red if both pregnancy_start_date and pregnancy_end_date are imputed;

and within the same quality color, the ordering is: EUROCAT, PROMPT, ITEMSETS, CONCEPTSETS, due to the expected accuracy in date recordings; within CONCEPTSETS, concept sets referring to longer pregnancies are considered to be higher in the hierarchy. The resulting ordering is as follows

1) EUROCAT: both pregnancy_start_date and pregnancy_end_date are recorded

2) PROMPT: both pregnancy_start_date and pregnancy_end_date are recorded

3) ITEMSETS: both pregnancy_start_date and pregnancy_end_date are recorded

4) CONCEPSETS: pregnancy completed and pregnancy_start_date recorded

5) PROMPT: pregnancy completed and pregnancy_start_date not available and imputed

6) ITEMSETS: pregnancy completed and pregnancy_start_date not available and imputed

7) CONCEPSETS: pre-term birth, meaning non primary care, pregnancy_start_date not available and imputed

7) CONCEPSETS: at-term birth, meaning non primary care, pregnancy_start_date not available and imputed

7) CONCEPSETS: post-term birth, meaning non primary care, pregnancy_start_date not available and imputed

8) CONCEPSETS: live birth, meaning non primary care, pregnancy_start_date not available and imputed

9) CONCEPSETS: stillbirth, meaning non primary care, pregnancy_start_date not available and imputed

10) CONCEPSETS: interruption, meaning non primary care, pregnancy_start_date not available and imputed

[6] note that if the record is in primary care, the estimation of end date may be wrong of some weeks

11) CONCEPTSETS: spontaneous abortion, meaning non primary care, pregnancy_start_date not available and imputed

12) CONCEPTSETS: ectopic pregnancy, meaning non primary care, pregnancy_start_date not available and imputed

13) CONCEPSETS: stillbirth possible, meaning non primary care, pregnancy_start_date not available and imputed

14) CONCEPSETS: interruption possible, meaning non primary care, pregnancy_start_date not available and imputed

15) CONCEPSETS: spontaneous abortion possible, meaning non primary care, pregnancy_start_date not available and imputed

20) CONCEPTSETS: meaning implying primary care, pregnancy_start_date not available and imputed, end date estimated with record date

25) all Streams: ongoing pregnancy and pregnancy_start_date recorded

50) all Streams: ongoing pregnancy having pregnancy_start_date not available and imputed

Reconciliation of records

Within each person the records are sorted first per order_quality (DAPs may require a different hierarchy rules) and then in reverse order of record_date, from the most recent to the most ancient. 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.

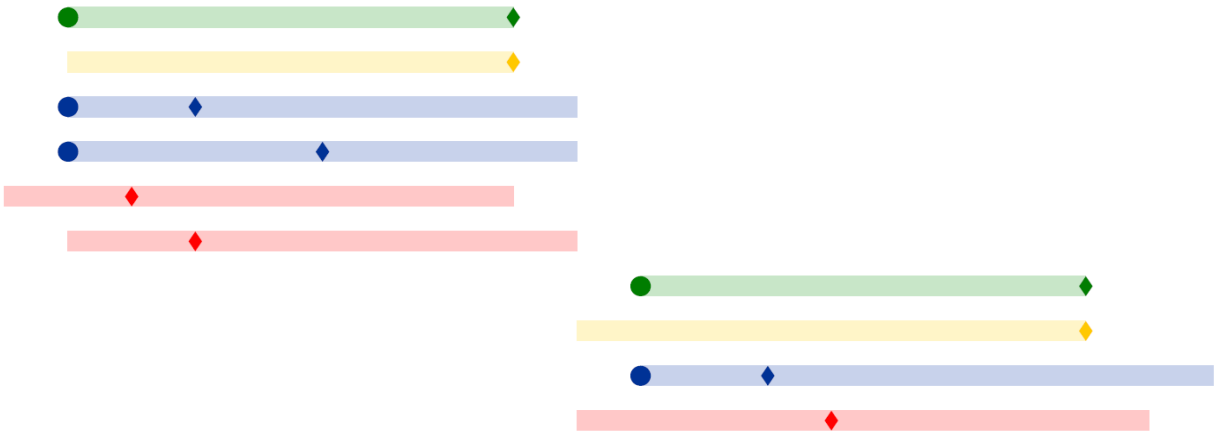


Figure 2. Graphical representation of two groups of records. In the graphical representation, a diamond represents the date of the record, a circle represents a record of a date of start of pregnancy, and the bar represents the interval between start and end. When the diamond is not at the end of the 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

Quality rules:

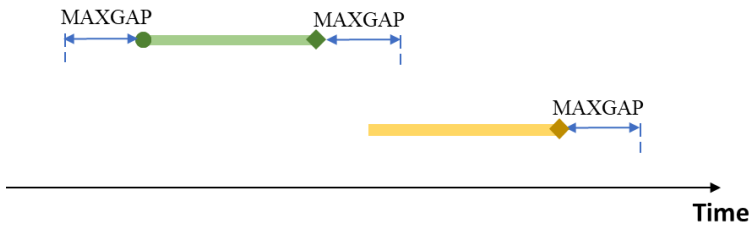
Each time a comparison is made, a string describing the reconciliation result will be added to the variable "algorithm_for_reconciliation". In the meantime, it will be checked whether the pregnancy information needs to be updated. The first review concerns the type of the end of pregnancy. If the type of the first record is different from the comparison record, and the 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

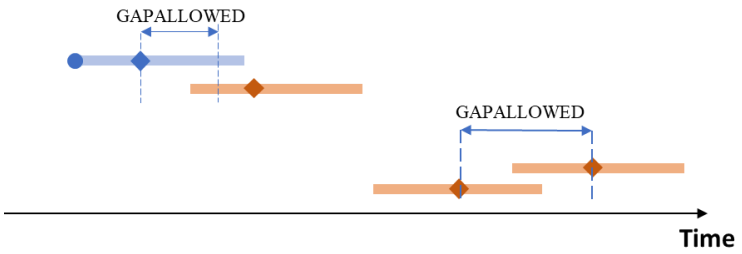
Rules for reconciliation are described in the following box.

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

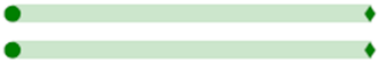
2) end date < start date next record

3) start date > end date next record

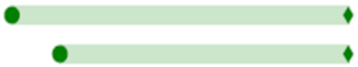
Reconciling Green with Green:

Thus, among records belonging to the same pregnancy:

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



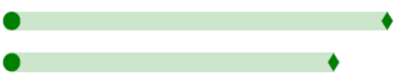
b) if they have different start dates:



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:

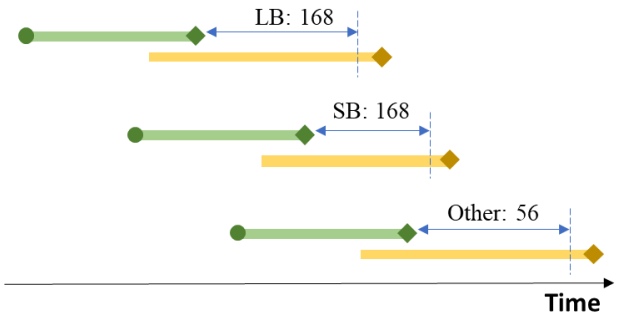
- When a non-LB record is compared with a LB record, LB pregnancies is selected
- Otherwise most recent pregnancy is selected
-

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

Reconciling Green with Yellow:

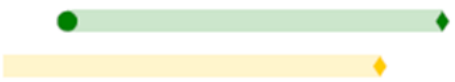
When a green record is compared with a yellow record, the records are assigned to different pregnancies if:

- 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



a) if the end dates are concordant, algorithm_for_reconciliation = “GY:concordant_”

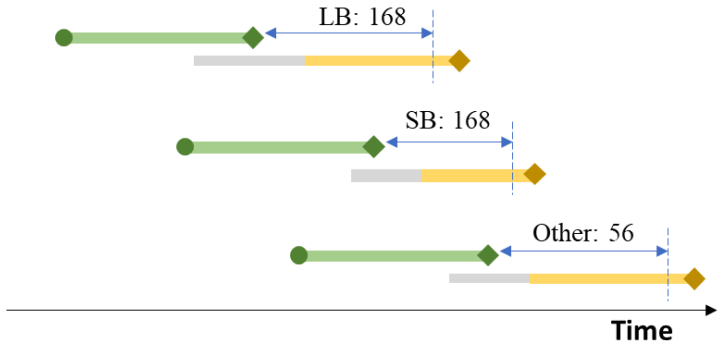
b) if the inconsistency is only on dates and they are of less than 7 days algorithm_for_reconciliation = “GY:SlightlyDiscordantEnd_”



c) 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:

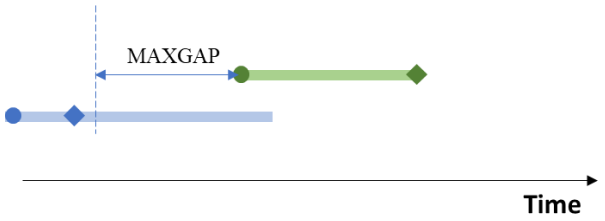
- If LB: End date - 154
- If SB: End date - 154
- If other: End date - 42



Reconciling Green with Blue:

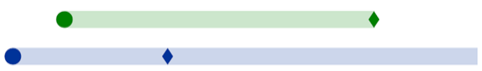
When a green record is compared with a blue record, the records are assigned to different pregnancies if:

- start date - MAXGAP > record date next record
-



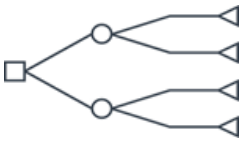
a) if the start dates are concordant, algorithm_for_reconciliation = "GB:concordant"

b) if the start dates are disconcordant,

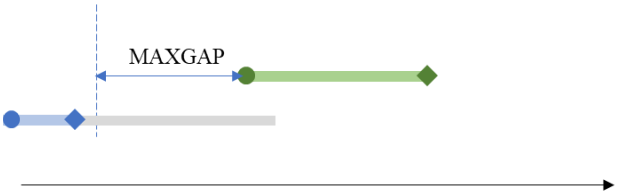


- as a default: update the value of pregnancy_start_date, the meaning_start_date, and set algorithm_for_reconciliation = "GB:StartUpdated"; the rationale is that the start of pregnancy recorded during pregnancy is of higher quality;

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



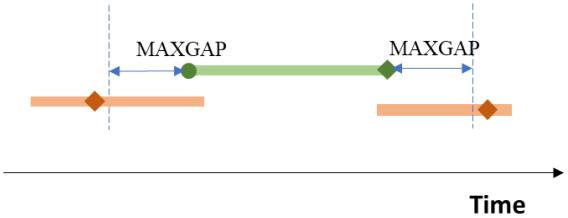
If pregnancies overlap, end of blue pregnancy is set at record date



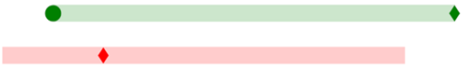
Reconciling Green with red:

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

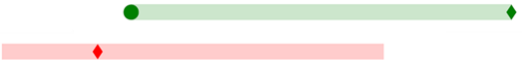
- End date green + MAXGAP < record date next record
- start date green - MAXGAP > record date next record



a) if 'record date' of the red record is between the start and end of the green record = "GR:NoInconsistecies"

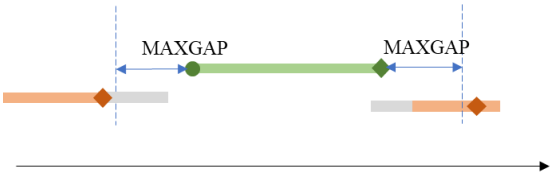


b) if 'record date' of the red record is not between the start and end of the green record = "GR:Inconsistecies"



If pregnancies overlap:

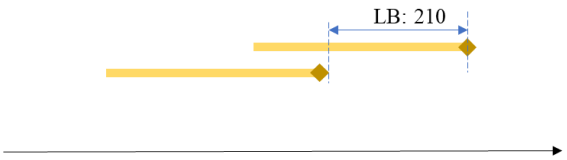
- If red pregnancy overlap on the left, end of red pregnancy is set at most recent record date
- If red pregnancy overlap on the right, start of red pregnancy is set at oldest record date – (MAXGAP/2)



Reconciling yellow with yellow:

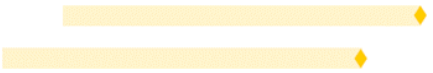
When a yellow record is compared with a yellow record, the records are assigned to different pregnancies if:

- 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



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

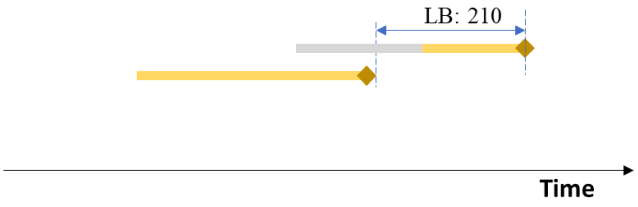
b) if the inconsistency is on dates and they are of less than 7 days apart, algorithm_for_reconciliation = "YY:SlightlyDiscordantEnd_"



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



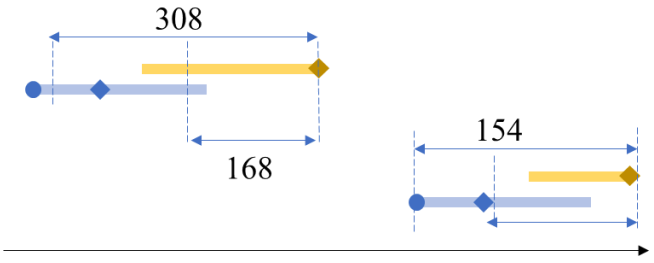
Reconciling yellow with blue:

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

- End date - 310 > start 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

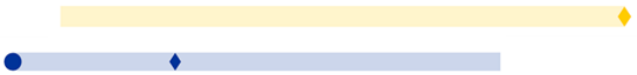


When a yellow record is reconciled with a blue record, the "meaning_start_date" and "imputed_start"

variables are updated, and:

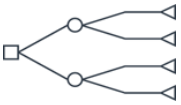
a) if the start dates are concordant, algorithm_for_reconciliation = “YB:concordant”

b) if the start dates are disconcordant,



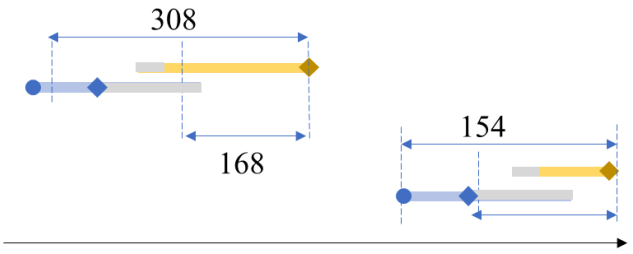
- as a default: update the value of pregnancy_start_date, the meaning_start_date and set algorithm_for_reconciliation = “YB:StartUpdated_”; the rationale is that the start of pregnancy recorded during pregnancy is of higher quality;

- 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



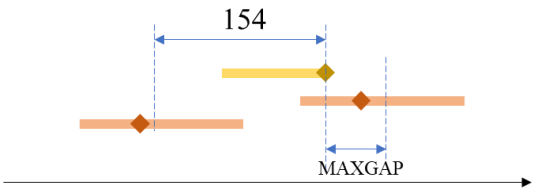
Reconciling yellow with red:

The records are assigned to different pregnancies if:

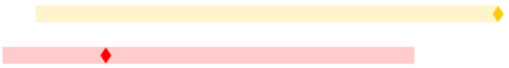
- end date + MAXGAP < record date next record

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

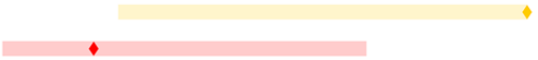
- end date – (154) > record date next record



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



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

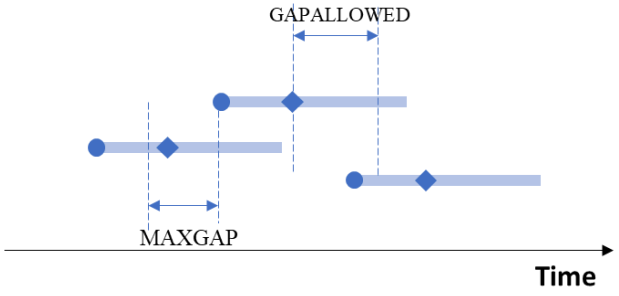


If pregnancies overlap: End of red pregnancy is set at most recent record date / start of red pregnancy is set at oldest record date – 14 days

Reconciling Blue with Blue:

Two records are classified as belonging to different pregnancies if:

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

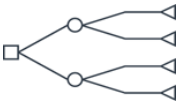


a) if the start dates are concordant, algorithm_for_reconciliation = "BB:concordant_"

b) 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

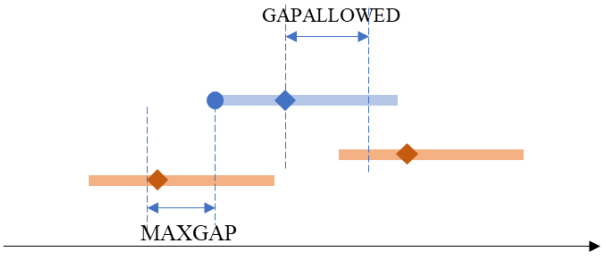


If pregnancies overlap: End of pregnancy is set at most recent record date

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



a) if 'record date' of the red record is between the start and end of the yellow record = “BR:NoInconsistecies_”

b) if 'record date' of the red record is not between the start and end of the green record =

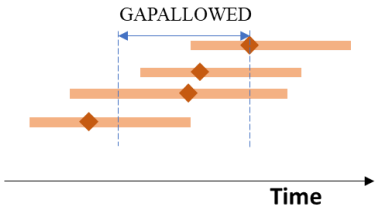
"BR:Inconsistecies_"

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

Reconciling Red with Red:

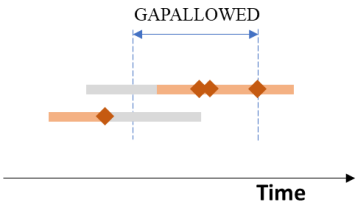
Two records are classified as belonging to different pregnancies if:

- 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. 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

The dataset resulting from the reconciliation, where the unit of observation is the record of pregnancy, will be saved as D3_groups_of_pregnancies_reconciled_before_excl, and the dataset containing the resulting pregnancies (unit of observation: pregnancy) will be saved as D3_pregnancy_reconciled_before_excl.

The alternative algorithms considered in step 3, 4 and 6 are indicated in Table 10

https://docs.google.com/spreadsheets/d/1oh3N2PBCjKw-uj6UhKdvzLzCE-yEi4keEnJy0Jb7dbg/edit#gid=16

48347487

Table 10. Algorithms recommended by each DAP to reconcile pregnancies

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 for birth_narrow diagnosis codes, record date - 55 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 will be 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 will be 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 will be 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 will be 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.

Create dummy for quality criteria

To datasets “D3_pregnancies_reconciled_before_excl” and “D3_groups_of_pregnancies_reconciled_before_excl” will be added with dummy variables that the daps will use to select pregnancies to include/exclude.

Variables will be defined as follows:

1. GGDE: set to 1 if the pregnancy is composed by two green record that are discordant on the date of end of pregnancy, 0 otherwise
2. GGDS: set to 1 if the pregnancy is composed by two green record that are discordant on the date of start of pregnancy, 0 otherwise
3. gestage_greater_44: set to 1 if the gestational age is greater than 44 weeks

Finally, the tables will be renamed as “D3_pregnancies_reconciled” and “D3_groups_of_pregnancies_reconciled”.

Verification

A random sample of 35 records of D3_pregnancies_reconciled will be extracted and stored in tables sample_from_pregnancies_anon.csv. For each pregnancy, start, end, and type of end are defined by information obtained from multiple records retrieved from different sources. When records carry different information, the algorithm selects the highest quality (in accordance with the hierarchy) and saves the results of the process. Within each quality color, records of the same pregnancy can be concordant (equal start and end dates), slightly discordant (start and end date different by no more than 7 days) or discordant (start and end date different more than 7 days). The sampling method chosen for the verification process is stratified sampling. The strata were initially defined according to the color quality, and then, within each color quality, divided by concordance. The following mutually exclusive strata were obtained:

- Green_Concordant
- Green_Discordant
- Yellow_Concordant
- Yellow_CP_SlightlyDiscordant
- Yellow_Discordant
- Blue
- Red

Each pregnancy in the sample is linked to its whole corresponding group of pregnancies from D3_groups_of_pregnancies_ reconciled and the result will be stored in sample_from_pregnancies.RData. The researchers of the DAP will be requested to review the original records that compose the group and store a judgement on the quality of the decision of the algorithms in Box 1 and Table 10. If needed, Table 10 will be updated by the DAPs to suggest modifications of the algorithm. The tale sample_from_pregnancies_anon.csv will be updated with the judgement, and the pregnancy identifier will be erased. The resulting dataset will be shared aggregated and with the central group for analysis. If a DAP is not allowed to share small numbers, only the suggestions will be shared.

3.6 Data analysis

Groups of pregnancies in D3_groups_of_pregnancies_ reconciled will be described across years of start of pregnancy as

- number of pregnancies
- which combination of streams detected the pregnancy
- which combination of record quality detected the pregnancy
- distribution of type of end (whether it is a life birth, a still birth…)

It will be also described:

- the distance between the end of the pregnancy and the first record retrieved
- maternal age at start of pregnancy: median + quartiles, mean + standard deviation
- distribution of meaning of start of pregnancy (how the start of pregnancy date was computed)
- distribution of meaning of end of pregnancy (how the end of pregnancy date was computed)

3.7 Limitations

A systematic validation of the list of pregnancies emerging from the iterations of the analysis is not feasible in this study. This implies that start of pregnancy may be misclassified, and that some pregnancies may be excluded because information is not available, For a comprehensive list of limitations of using secondary data sources to identify pregnancies, see Margulis A et al, 2015.

However, validation of a sample of records from selected data sources (see section 3.5) is expected to provide actionable information.

The output of this study will consist of several subpopulations of pregnancies, classified according to the expected sensitivity and accuracy of the most relevant study variables, eg the subpopulation detected by the various streams, or the meaning of the start of pregnancy. It is not within the scope of this study to provide recommendations on which population of pregnancy is most suitable for a specific class of research questions. However, this study will create the conditions for decisions to be made with respect to this at the stage of protocol design, by balancing power, sensitivity to a wide range of outcomes, accuracy, and external validity, and by supporting the design of sensitivity analyses, if appropriate.

4. Data models and study script

As specified in the Deliverable 7.11 (Thayer D, et al, 2021), the scripts of the ConcePTION pipeline are hosted in a GitHub repository, see Bartolini et al, 2021.

4.1 Dummy tables and figures

The dummy tables and figures are listed in the spreadsheet at this link

4.2 Datasets of results

4.3 Analytic datasets (D4)

Since this is a descriptive study, the analytic datasets coincide with the D3s.

4.4 Main datasets of study variables (D3)

4.4.1 Datasets of study variables of included and excluded pregnancies

Name: D3_included_pregnancies

Unit of observation: pregnancies that were completed during the period of time included in the datasource instance

https://docs.google.com/spreadsheets/d/19ulBqvPBEJ6dpUQh39397pLyKPw10Dl3yn0hR5wHHPQ/edit#gid=184804680

Name: D3_excluded_pregnancies

Unit of observation: pregnancies that were extracted but had to be discarded due to inconsistent information

https://docs.google.com/spreadsheets/d/19ulBqvPBEJ6dpUQh39397pLyKPw10Dl3yn0hR5wHHPQ/edit#gid=458907268

4.4.2 Datasets from the Streams

D3_Stream_PROMPTS

Unit of observation: Records from SURVEY_ID

https://docs.google.com/spreadsheets/d/19ulBqvPBEJ6dpUQh39397pLyKPw10Dl3yn0hR5wHHPQ/edit#gid=0

D3_Stream_EUROCAT

Unit of observation: Records from EUROCAT

https://docs.google.com/spreadsheets/d/19ulBqvPBEJ6dpUQh39397pLyKPw10Dl3yn0hR5wHHPQ/edit#gid=59209496

D3_Stream_CONCEPTSETS

Unit of observation: Records from EVENTS, MEDICAL_OBSERVATIONS, PROCEDURES

https://docs.google.com/spreadsheets/d/19ulBqvPBEJ6dpUQh39397pLyKPw10Dl3yn0hR5wHHPQ/edit#gid=495713974

D3_Stream_ITEMSETS

Unit of observation: Records from MEDICAL_OBSERVATIONS detected by ITEMSETS

https://docs.google.com/spreadsheets/d/19ulBqvPBEJ6dpUQh39397pLyKPw10Dl3yn0hR5wHHPQ/edit#gid=1824327908

4.4.3 Dataset of groups of pregnancies

D3_groups_of_pregnancies

Unit of observation: Records from all the streams

https://docs.google.com/spreadsheets/d/19ulBqvPBEJ6dpUQh39397pLyKPw10Dl3yn0hR5wHHPQ/edit#gid=2109944600

4.4.4 Datasets for test

sample_from_pregnancies_anon

Unit of observation: a pregnancy in the sample for test

	record 1	…
preg_id		
n		
pregnancy_start_date		
pregnancy_end_date		

type_of_pregnancy_end		
pregnancy_start_date_correct	to be filled by validator	…
pregnancy_start_date_difference	to be filled by validator	…
pregnancy_end_date_correct	to be filled by validator	…
pregnancy_end_date_difference	to be filled by validator	…
type_of_pregnancy_end_correct	to be filled by validator	…
records_belong_to_multiple_pregnancy	to be filled by validator	…
comments	to be filled by validator	…
record_date		
origin		
meaning		
codvar		
coding_system		
conceptset		
source_column		
source_value		
itemsets		
link		

4.5 Study script

The study script is available and documented in the following repository

https://github.com/ARS-toscana/ConcePTIONAlgorithmPregnancies

References

Bartolini et al, 2021. Repository of the script of the ConcePTION Algorithm for Pregnancies. Accessible from: https://github.com/ARS-toscana/ConcePTIONAlgorithmPregnancies/wiki. Accessed April 2021

Becker B 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.

Charlton RA et al, 2014. Charlton RA, Neville AJ, Jordan S, Pierini A, Damase-Michel C, Klungsøyr K, et al. Healthcare databases in Europe for studying medicine use and safety during pregnancy. Pharmacoepidemiol Drug Saf. 2014 Mar 24;

ConcePTION CDM v2.2. Table specifications. Accessible from : https://docs.google.com/spreadsheets/d/1hc-TBOfEzRBthGP78ZWIa13C0RdhU7bK/edit#gid=144408873 Accessed April 2021.

Dodd C et al, 2020. Dodd C, et al. D7.5 Report on existing common data models and proposals for ConcePTION. https://www.imi-conception.eu/wp-content/uploads/2020/10/ConcePTION-D7.5-Report-on-existing-common-data-models-and-proposals-for-ConcePTION.pdf. Accessed April 2021.

Margulis A et al, 2015. Margulis AV, Palmsten K, Andrade SE, Charlton RA, Hardy JR, Cooper WO, et al. Beginning and duration of pregnancy in automated health care databases: review of estimation methods and validation results. Pharmacoepidemiology and Drug Safety. 2015;24(4):335–42.

Matcho A et al, 2018. Matcho A, Ryan P, Fife D, Gifkins D, Knoll C, Friedman A. Inferring pregnancy episodes and outcomes within a network of observational databases. PLOS ONE. 2018 Feb 1;13(2):e0192033.

Nordeng H et al, 2021. CONSIGN study: COVID-19 infection and medicines in pregnancy - a multinational registry based study. EU PAS Register EUPAS39438. Accessible from: http://www.encepp.eu/encepp/viewResource.htm?id=39439. Accessed May 2021

Ortiz SS 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.

Thayer D, et al, 2021. D7.11 Wiki page with description of tools and GITHUB repository. https://www.imi-conception.eu/wp-content/uploads/2021/06/ConcePTION_D7.11_github.pdf

Sarayani A , et al, 2020. Sarayani A, Wang X, Thai TN, Albogami Y, Jeon N, Winterstein AG. Impact of the Transition from ICD–9–CM to ICD–10–CM on the Identification of Pregnancy Episodes in US Health Insurance Claims Data. Clin Epidemiol. 2020 Oct 15;12:1129–38.

Schink T 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. Front Public Health. 2020;8:350.

Sturkenboom M et al, 2019. Impact of EU label changes and revised pregnancy prevention programme for oral retinoid containing medicinal products: utilization and prescribing trends. EU PAS Register EUPAS31095. Accessible from : http://www.encepp.eu/encepp/viewResource.htm?id=31096. Accessed May 2021

Sturkenboom M et al, 2020. Background rates of Adverse Events of Special Interest for monitoring COVID-19 vaccines. EU PAS Register EUPAS37273. Accessible from: http://www.encepp.eu/encepp/viewResource.htm?id=40361. Accessed April 2021.

Swertz et al, 2021. D7.7 Prototype of FAIR data catalogue-2nd. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5e0fd67dd&appId=PPGMS