OMOP Common Data Model (CDM V5.4)

CIDACS Partnership with Western Cape Health Provincial Data Centre Mapping Specification

Junho/2025

Valentina Martufi, Juliana Prata, Danilo Dias
Karine Beck, Ricardo Félix, Julio Barbour

Contents

1.0 Abbreviations:

CDM	Common Data Model
ETL	Export Transform Load
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
IDAF	Data Interoperability and Analyses Group Federated CIDACS Data Platform

2.0 Introduction

This document reflects the requirements, assumptions, business rules and transformations for the implementation of the Common Data Model Version 5.4 (CDM) as implemented by CIDACS.

The purpose of this document is to describe the ETL mapping of the proprietary or licensed data from CIDACS Partnership with Western Cape Provincial Health Data Centre into the OMOP Common Data Model.

It is based on the OMOP ETL Specifications. General information that is covered by the OMOP ETL Specification will not be covered in this document, but a detailed discussion of the CIDACS Partnership with Western Cape Provincial Health Data Centre specific aspects of mapping and converting data to the standard CDM is provided.

Data model design decisions

- **Person-centric**
- Standardized **structure** (schema)
- Standardized **contents** (using Standardized Vocabularies)
- Optimized for **data analysis** (not ops / data transfer e.g. FHIR/OpenEHR)
- **Extensible**

Below is a depiction of the OMOP relational (SQL) database schema that adheres to these principles.

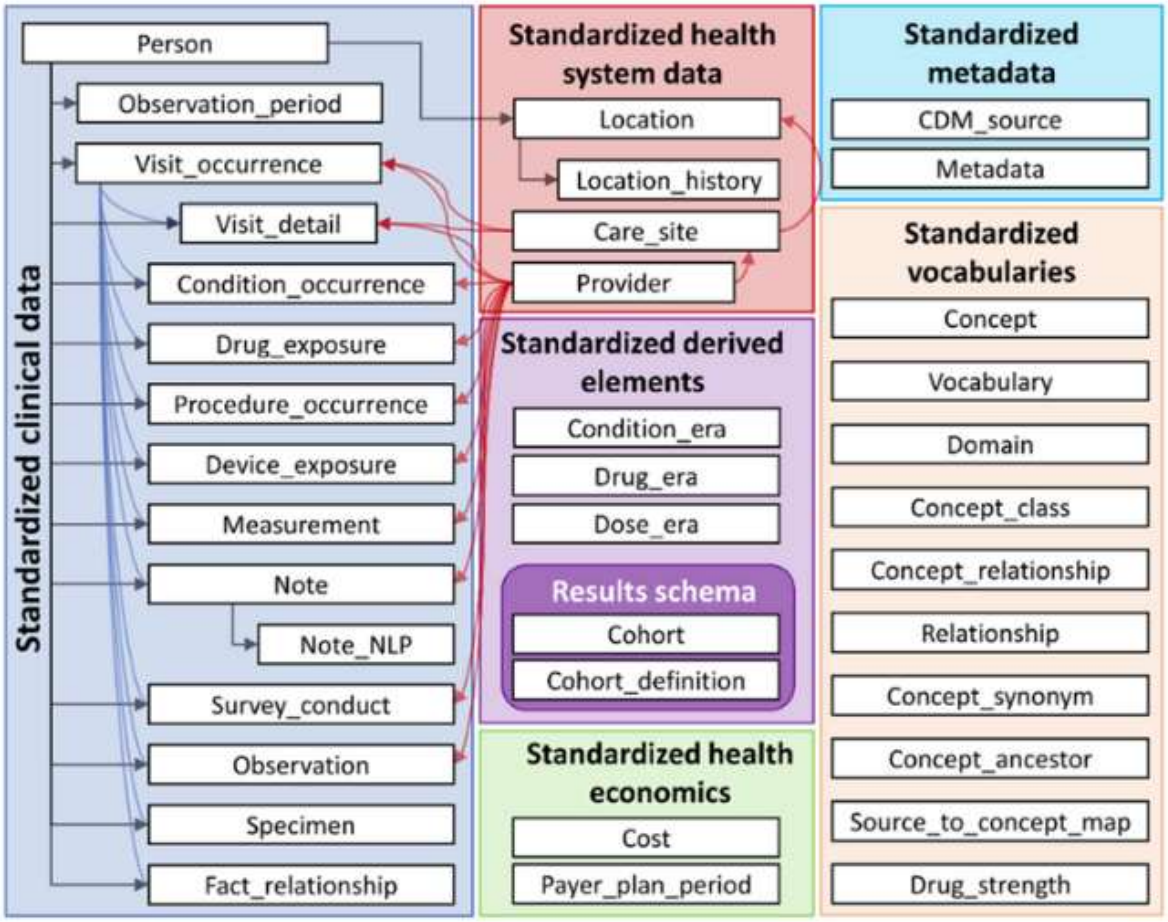


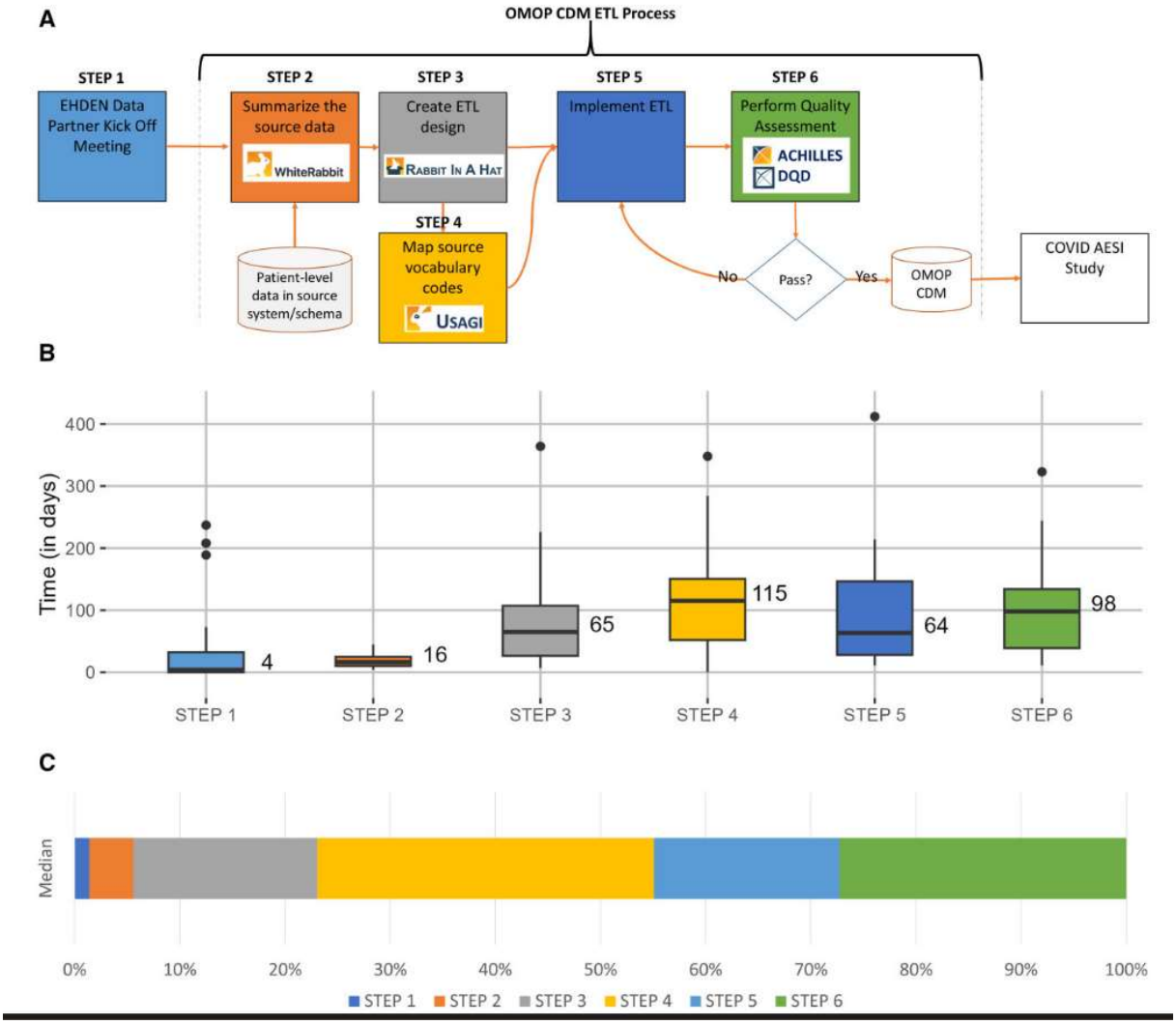
Figure 1: Overview of all tables in the CDM version 6.0. Note that not all relationships between tables are shown

OHDSI Community. (2021). The Book of OHDSI. Observational Health Data Sciences and Informatics. Available at: https://ohdsi.github.io/TheBookOfOhdsi/

OMOP CDM ETL – Time Estimates
Evidence from 25 data partners in the European Health Data & Evidence Network (EHDEN) suggest that there is a typical 6 step process to build and carry out a generic OMOP CDM extract transform load (ETL) data pipeline, where on average all steps together takes approximately 1 year.

Step 1 (4) + Step 2 (16) + Step 3 (65) + Step 4 (115) + Step 5 (65) + Step 6 (98) ~ 363-day average

The times and sequencing of activities were estimated and based on this process.



https://doi.org/10.1093/jamia/ocad214

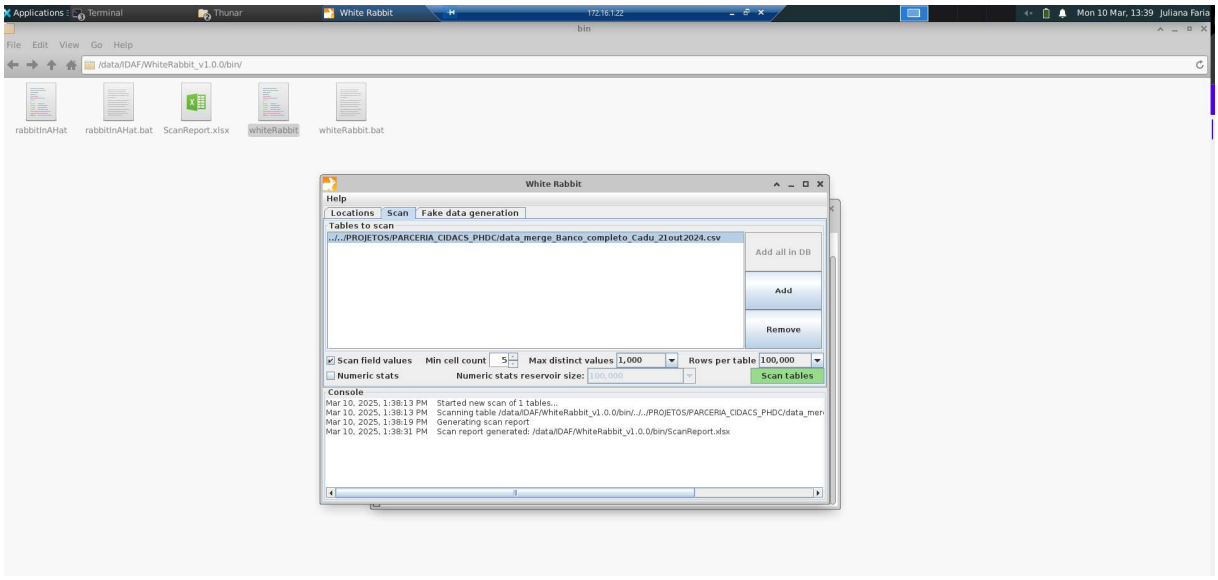
Figure 1: OMOP CDM ETL development process: (A) represents the ETL process map, (B) is a box plot of median length in days for each step across all data partners, and (C) is a stacked bar chart showing the percentage of median time each step took. CDM, common data model; COVID AESI Study, “Adverse Events of Special Interest within COVID-19 Subjects” study; DQD, DataQualityDashboard; EHDEN, European Health Data & Evidence Network; ETL, extract, transform, and load; OMOP, outcomes partnership common data model.

The implementation of the process of transforming the original data into structured and standardised data followed the above model, in accordance with good practices, and the total number of days for transforming the source data into OMOP CDM is also shown in the figure.

Step 1 - 06 days
This stage takes into account the time between the formation of the CIDACS Data Platform's Interoperability and Federated Analyses group (IDAF) and the project's kick-off meeting.

Step 2 - 251 days

WhiteRabbit was initially applied to the synthetic database to build the OMOP CDM, but as the project progressed and the learning curve developed, it was decided to use a database populated with original data. WhiteRabbit was re-run on the original database. As a result, the time allocated to this stage was extended to take account of the change in DataSource.



WhiteRabbit run in the ON-PREMISE environment

Step 3 and 4: 179 days

The stages of creating the ETL and defining and mapping the vocabulary codes took place concurrently.

Step 5: 223 days

The implementation of the ETL was started before the previous stage was completed because, due to the design of two ETLs for the construction of the OMOP, it was possible to start one of them before finalising both.

Step 6: 13 days

The quality assessment stage began after the construction of an infrastructure with improved robustness and unprecedented processing power in the exclusive analysis environment for the development of the work of IDAF and WG1 (Work Group 1 responsible for building the common OMOP data model; Side CIDACS). The design of the solution was studied and implemented by the Cidacs Data Platform to support the OHDSI tools on local servers.

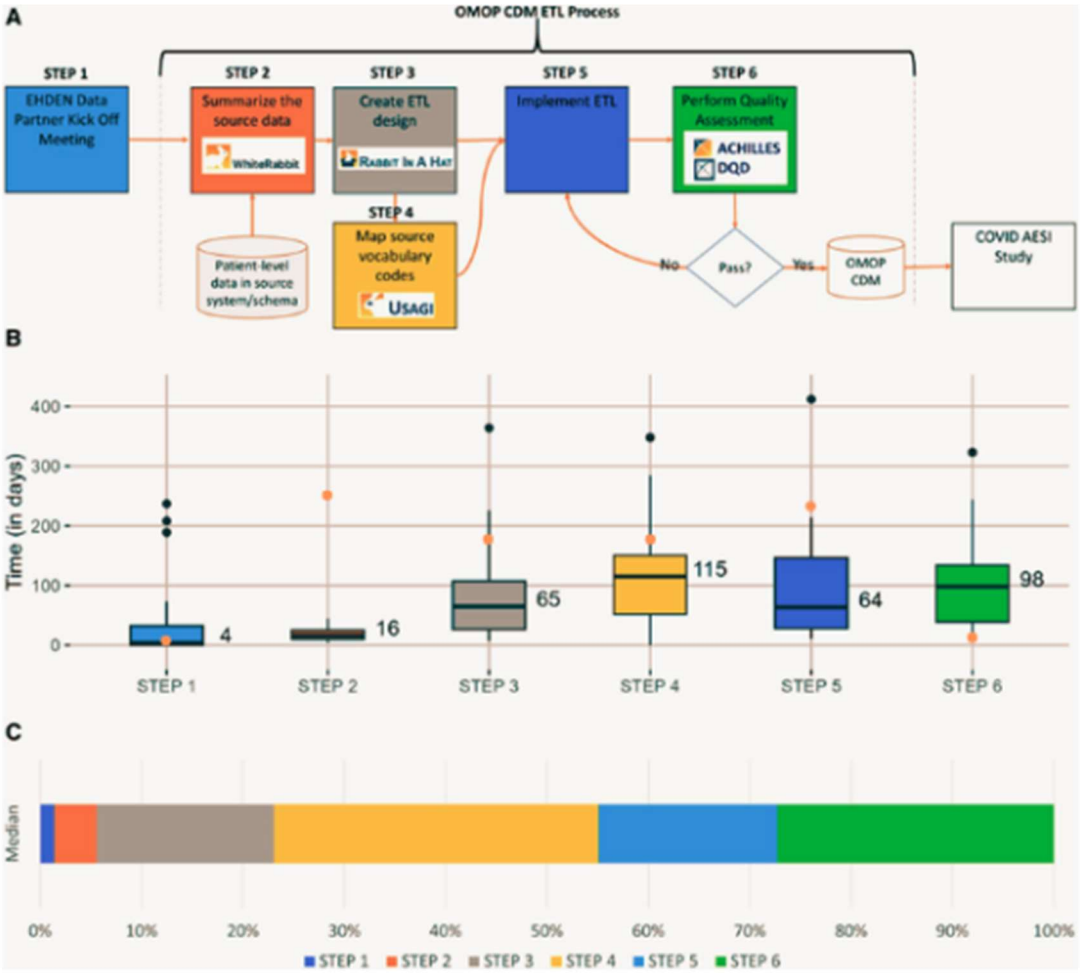


Figure 1: OMOP CDM ETL development process: (A) represents the ETL process map, (B) is a box plot of the average duration in days for each step across all data partners highlighting in orange the times obtained by the Wg1 team (orange points; Side CIDACS) and (C) is a stacked bar chart showing the percentage of the average time each step took. CDM, common data model; COVID AESI Study, "Adverse Events of Special Interest within COVID-19 Subjects" study; DQD, DataQualityDashboard; EHDEN, European Health Data & Evidence Network; ETL, extract, transform and load; OMOP, outcomes partnership common data model.

The document is composed of three main sections:

- **Source Data Mapping Approach.** Describes the data source health information selected for data modeling
- **Source Data Mapping of Mother.** Describes the main tables of the CDM schema and the special Mother data handling required for each table.
- **Source Data Mapping of Infant.** Describes the main tables of the CDM schema and the special Infant data handling required for each table.

In each section, the tables and their mapping are individually reviewed along with any source specific rules and exceptions.

3.0 Source Data Mapping Approach

For the creation of the Common Data Model for the study of infectious diseases affecting pregnant women in the Global South, this modeling work was made by using anonymized administrative data available in the CIDACS analysis environment under controlled access, following the Center's Data Governance guidelines. The variables were selected and extracted from datasets made available by Work Group 2, as described in Tables 1 and 2.

Table 1 presents the datasets used for this selection, and Table 2 lists the 42 variables that were selected by WG2 (CIDACS Side) and are the subject of this standardized data modeling work for the OMOP CDM Version 5.4.

Data Source	Definition
SINASC	Sistema de Informações sobre Nascidos Vivos | *Live Birth Information System*
SINAN- Sífilis	Casos notificados de sífilis gestacional (SG). Sistema de Informação sobre Agravos/para Doenças de Notificação | *Reported Cases of Gestational Syphilis (GS). Disease Notification Information System*
SIM	Sistema de Informação sobre Mortalidade | *Mortality Information System*
CADU	Informações socioeconômicas renovadas periodicamente nos conjuntos de dados do Cadastro Único. Criado em 2001 e em 2003 tornou-se o principal registro para a implementação e gestão dos programas sociais novos e existentes | *Periodically Updated Socioeconomic Information in the Unified Registry (Cadastro Único) datasets. Created in 2001 and became the main registry in 2003 for the implementation and management of new and existing social programs.*
Merges: SINASC-SINAN-Sífilis | SINASC-SIM | CADU-SINASC-Sífilis DataSource:. Basefinal_gest_limp.parquet Database size: 410 MB Parquet 16 million records Time frame: 2011 a 2020	

Table 1

Information on the variables selected for mapping				
Count	**field_name**	**table_name**	**data_type**	**field_description**
1	APGAR 1	sinasc	int	Apgar at the 1st minute
2	APGAR 5	sinasc	int	Apgar at the 5th minute
3	CODANOMAL	sinasc	string	Code of congential anomaly found (ICD 10)
4	CODMUNRES	sinasc	int	Municipality of residence of the newborn (code)
5	CONSPRENAT	sinasc	int	Number of ANC visits completed
6	CONSULTAS	sinasc	int	Number of prenatal consultations (0, 1-3, 4-6, 7+, unknown)
7	DT_NOTIFIC_MAE	sífilis gestacional	date	Notification form date
8	DT_OBITO	sim	date	Date of death (dd mm yyyy)
9	DTNASC	sinasc	date	Date of birth (dd mm yyyy)
10	DTNASCMAE	sinasc	date	Mother's date of birth (dd mm yyyy)
11	ESCMAE	sinasc	int	Newborn's mother's schooling (in years categories: 0, 1-3, 4-7, 8-11, 12+)
12	ESTCIVMAE	sinasc	int	Mother's marital status (single, married, widower, legally separated/divorces, civil union)
13	GRAVIDEZ	sinasc	int	Type of pregnancy (single, double, triple, etc)
14	ID_AGRAVO_MAE	sífilis gestacional	string	Name and code of the reported ailment according to ICD-10
15	ID_CIDACS_MAE	sinasc	bigint	Unique sequential identifier of the mother.
16	ID_CIDACS	sinasc	bigint	Unique sequential identifier of the record
17	IDADEMAE	sinasc	int	Newborn's mother's age
18	IDANOMAL	sinasc	int	Birth anomaly detected
19	MESPRENAT	sinasc	int	Month of pregnancy in which ANC was commenced
20	PARTO	sinasc	int	Type of birth (vaginal, c-section, etc)
21	PESO	sinasc	int	Weight at birth (gr)
22	QTDFILMORT	sinasc	int	Number of fetal losses and abortions
23	QTDFILVIVO	sinasc	int	Number of live children
24	RACACORMAE	sinasc	int	Race/etnicity
25	SEMAGESTAC	sinasc	int	Weeks of gestation
26	SEXO	sinasc	int	Sex of the baby
27	TPAPRESENT	sinasc	int	Fetal presentation (occiput anterior, posterior, or breech, etc)
28	TPCONFIRMA	sífilis gestacional	int	Refers to the result of the confirmatory treponemal test, performed prenatally. Confirmatory treponemal tests are indicated for diagnostic confirmation and exclusion of false positive non-treponemal test results. Treponemal tests are classified as: FTA-Abs (Fluorescent Treponemal Antibody-absorption), MHA-Tp (Microhemagglutination Treponema pallidum Assay), TPHA (Treponema pallidum Hemagglutination Assay), ELISA (Enzyme-Linked Immunosorbent Assay). (Null, Reactive, Non-reactive, Not performed, Ignored, Inconsistency)
29	TPESQUEMA_MAE	sífilis gestacional	int	Treatment scheme recommended by the Ministry of Health and available in the Guidelines Manual for the control of congenital syphilis (Null,penicillin g benzathine 2,400,000iu, penicillin g benzathine 4,800,000iu, penicillin g benzathine 7,200,000iu, other scheme, not performed)

30	TPEVIDENCI_MAE	sífilis gestacional	int	Clinical classification of syphilis in pregnancy 1-primary syphilis-hard chancre; 2-secondary syphilis-skin-mucosal lesions (papular syphilitic roseola, flat condyloma, alopecia); 3-tertiary syphilis-skin-mucosal lesions (tubercles or gummas); neurological alterations (tabes dorsales, dementia); cardiovascular alterations (syphilitic aortitis, aortic aneurysm); joint alterations (Charcot arthropathy); 4-latent syphilis-asymptomatic phase the diagnosis is only obtained through serological reactions.
31	TPTESTE1_MAE	sífilis gestacional	int	Refers to the result of the non-treponemal test recommended for the first prenatal consultation. Non-treponemal tests are understood as VDRL and RPR. (reagent, non-reagent, not done)
32	TRA_DT_SC	sífilis congênita	int	Date of initiation of treatment for syphilis of the mother of the reported case.
33	MARC_PBF_EQ_V4	cadu pop 100	cat	Indicates whether you are a beneficiary of the Bolsa Família
34	QTDGESTANT	sinasc	int	Number of previous pregnancies
35	DTULTMENST	sinasc	date	Date of last menstrual period

Table 2

From the source data, it is necessary to create an extraction, transformation, and loading (ETL) process. This process will structure the data for the CDM and add mapping to the standardized vocabularies. Automated SQL scripts will be generated. The goal is to make the ETL repeatable so that it can be executed again whenever the source data is updated.

The ETL stages include design, mapping, implementation, quality control, and maintenance. The activities are described in the following stages:

Design: The ETL design stage utilized OHDSI tools, White Rabbit, and Rabbit in a Hat. This stage involves a thorough understanding of the source data and the CDM. Additionally, technical knowledge is required to make the ETL computationally efficient.

Create Code Mappings: This stage involves relating the source concept to the target concept. There was an effort to map the most frequently used codes and to use the Observation table for mapping socioeconomic variables and other concepts from the source data not covered in the standardized tables.

Implementation: The implementation was carried out using SQL in the on-premise environment.

Quality Control: An iterative process within the ETL process. The CDM will be tested using best practices recommended by the OHDSI community.

CDM and ETL Maintenance: There is a maintenance cycle for the ETL, which is an ongoing process after the first CDM is built. Some common triggers requiring maintenance include changes in the source data, an ETL bug, a new OMOP Vocabulary release, or the CDM itself being altered or updated.

Since the CDM is a person-centered model, and for this study, the focus is on the Mother and Child, we opted for the design of the following three ETLs:

1. Mapping of maternal data

2. Mapping of child data

3. Mapping that links Mother and Infant (Fact_Relationship)

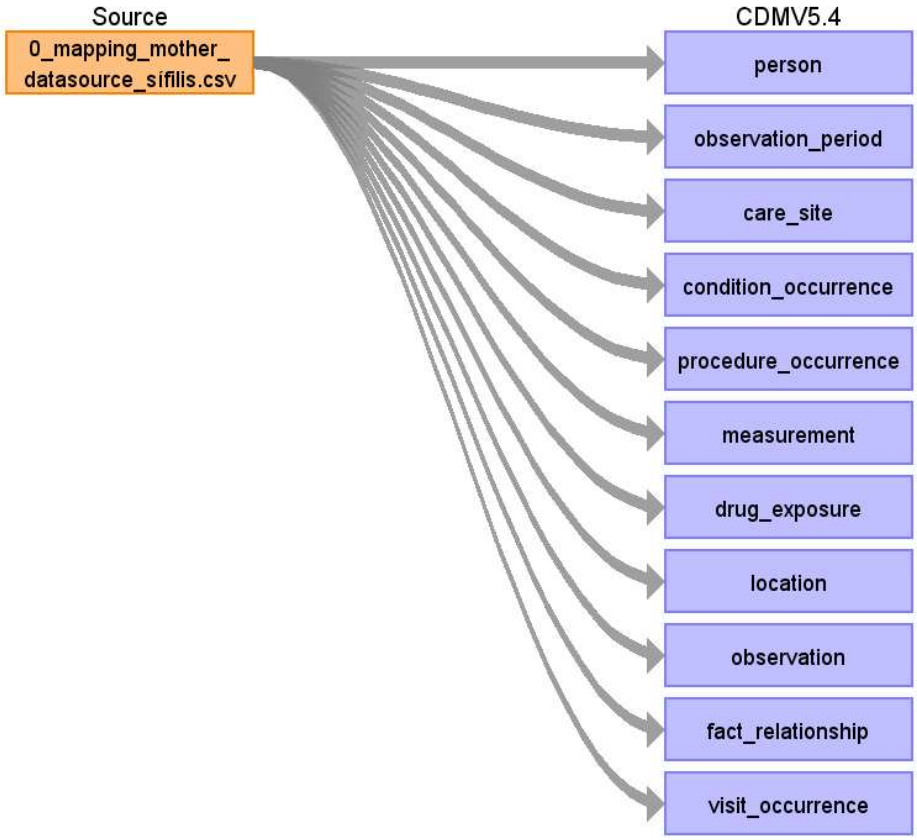
A UNION ALL will be applied to Tables 1 and 2, and the mother and child IDs will be linked through Table 3.

The Common Data Model was implemented in SQL on the Data Source and access to the standardized data is provided through analytical tools chosen by the research group.

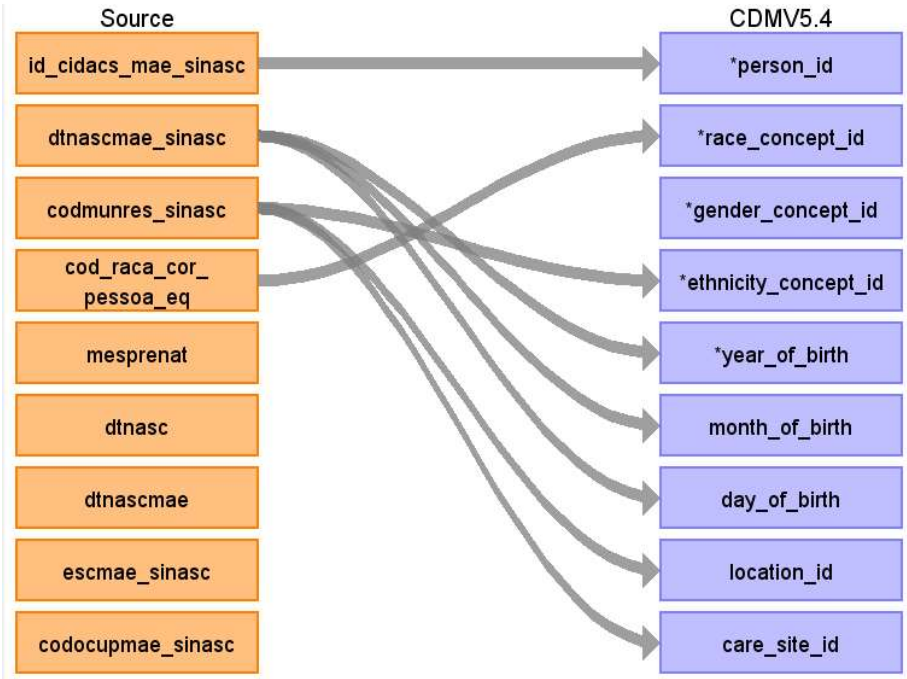
Tools like R e Python are commonly used.

4.0 Source Data Mapping of Mother

Below, the design of the mapping process related with the ETL of maternal data is presented under the format of images and tables extracted from the OHDSI tool named *Rabbit in a Hat.* The maternal mapping included 11 template tables.



4.1 Table Name: PERSON

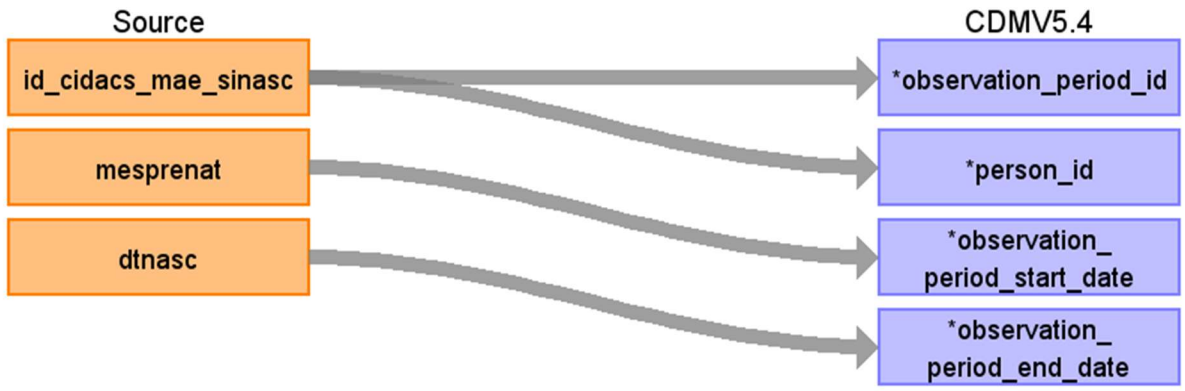


Destination Field	Source Field	Logic
person_id	id_cidacs_mae_sinasc	id_cidacs_mae_sinasc = person_id
race_concept_id	cod_raca_cor_pessoa_eq	if cod_raca_cor_pessoa_eq = 0 then race_concept_id = 0 if cod_raca_cor_pessoa_eq = 1 then race_concept_id = 8527 if cod_raca_cor_pessoa_eq = 2 then race_concept_id = 38003598 if cod_raca_cor_pessoa_eq = 3 then race_concept_id = 8515 if cod_raca_cor_pessoa_eq = 4 then race_concept_id = 4212311 if cod_raca_cor_pessoa_eq = 5 then race_concept_id = 19387526 if cod_raca_cor_pessoa_eq = 99 then race_concept_id = 0
gender_concept_id	cod_sexo_pessoa_eq	if cod_sexo_pessoa_eq = 0 (null) then gender_concept_id = 0 if cod_sexo_pessoa_eq = 1 (male) then gender_concept_id = 8507

		if cod_sexo_pessoa_eq = 2 (female) then gender_concept_id = 8532 if cod_sexo_pessoa_eq = 88 (ignored) then gender_concept_id = 4214687 if cod_sexo_pessoa_eq = 99 (inconsistence) then gender_concept_id = 0
ethnicity_concept_id	naturalmae_sinasc codmunres	create variable sg_uf_mae --> extract two first digits from the variable 'codmunres_sinasc' if sg_uf_mae varies from 11 to 53, then ethnicity_concept_id = 38003563 (Hispanic or Latino) else ethnicity_concept_id = 0 State Code 0- Nulo 11- RO 12- AC 13- AM 14- RR 15- PA 16- AP 17- TO 21- MA 22- PI 23- CE 24- RN 25- PB 26- PE 27- AL 28- SE 29- BA 31- MG 32- ES

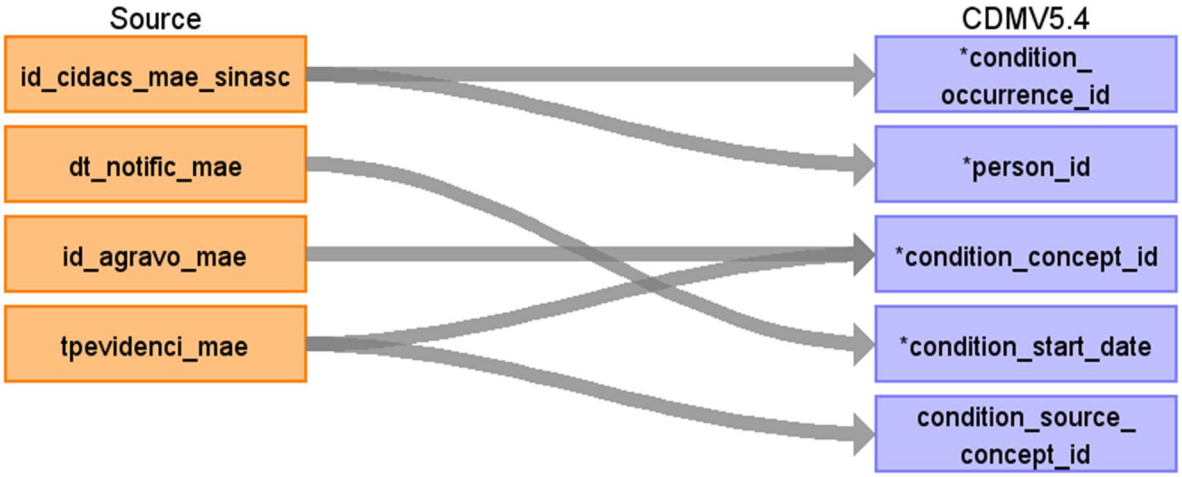
		33- RJ 35- SP 41- PR 42- SC 43- RS 50- MS 51- MT 52- GO 53- DF 99-Inconsistencia
year_of_birth	dtnascmae_sinasc idademae_sinasc	if dtnascmae_sinasc = null then dtnasc_sinasc - idade_mae_sinasc To build this data, is required the remotion of day and month from the variable dtnascmae_sinasc
month_of_birth	dtnascmae_sinasc	To build this data, is required the remotion of day and month from the variable dtnascmae_sinasc
day_of_birth	dtnascmae_sinasc	To build this data, is required the remotion of month and year from the variable dtnascmae_sinasc
location_id	codmunres_sinasc	Create ID (See table Location)
care_site_id	codmunres_sinasc	Create ID (See table Care_Site)

4.2 Table Name: OBSERVATION_PERIOD



Destination Field	Source Field	Logic	Comment
observation_period_id	id_cidacs_mae_sinasc	Create ID	Create ID
person_id	id_cidacs_mae_sinasc	See Person table	See Person table
observation_period_start_date	mesprenat	1.º extrair o mês da variável dtnasc 2.º reduzir do mês o valor encontrado na diferença: 9 - mesprenat 3.º extrair dia e ano do dtnasc 4.º observation_period_start_date == dia dtnasc/mês encontrado no modelo acima /ano dtnasc *o mês corresponde a um mês do ano	*** Utilizar a mesma regra aplicada na Visit_Occurrence para identificar a data da 1.ª visita pré-natal.
observation_period_end_date	dtnasc	observation_period_end_date == dtnasc_sinasc	
period_type_concept_id			Apply 32879

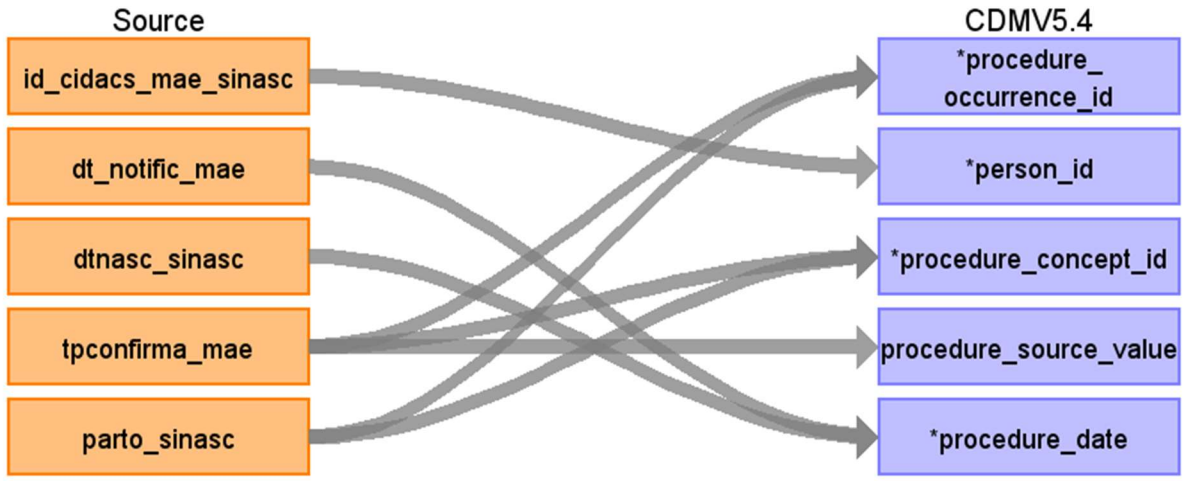
4.3 Table Name: CONDITION_OCCURRENCE



Destination Field	Source Field	Logic	Comment
condition_occurrence_id	id_cidacs_mae_sinasc	Create ID	Create ID
person_id	id_cidacs_mae_sinasc	See Person table	See Person table
condition_concept_id	tpevidenci_mae id_agravo_mae	if tpevidenci_mae == 1 then condition_concept_id == Primary (433410) if tpevidenci_mae == 2 then condition_concept_id == Secondary (437787) if tpevidenci_mae == 3 then condition_concept_id == Tertiary (435739) if tpevidenci_mae == 4 then condition_concept_id == Latent (40485064) if id_agravo_mae == 0981 then condition_concept_id == 443291 - Maternal syphilis during pregnancy, childbirth and the puerperium	Name and code of the reported ailment according to ICD-10 - 0981 Clinical classification of syphilis in pregnancy 1-primary syphilis-hard chancre; 2-secondary syphilis-skin-mucosal lesions (papular syphilitic roseola, flat condyloma, alopecia); 3-tertiary syphilis-skin-mucosal lesions (tubercles or gummas); neurological alterations (tabes dorsales, dementia); cardiovascular alterations (syphilitic aortitis, aortic aneurysm); joint alterations (Charcot arthropathy); 4-latent syphilis-asymptomatic phase the

		if id_agravo_mae == NA then condition_concept_id == 0	diagnosis is only obtained through serological reactions.
condition_start_date	dt_notific_mae	condition_start_date == dt_notific_mae for the concept mapping at id_agravo_mae and tpevidenci_mae	
condition_source_concept_id	tpevidenci_mae	Apply 4155395 - Infectious disease notification	
condition_type_concept_id			Apply 32879

4.4 Table Name: PROCEDURE_OCCURRENCE



Destination Field	Source Field	Logic	Comment
procedure_occurrence_id	parto_sinasc tpconfirma_mae	Create ID Create ID	Create ID
person_id	id_cidacs_mae_sinasc	See Person table	See Person table
procedure_concept_id	tpconfirma_mae parto_sinasc	4193412 - Treponema screening test if parto_sinasc == 0 then [target field OMOP] == 0 if parto_sinasc == 1 then [target field OMOP] == 44784097 (Vaginal delivery of fetus) if parto_sinasc == 2 then [target field OMOP] == 4015701 (Cesarean section) if parto_sinasc == 88 then [target field OMOP] == 0 if parto_sinasc == 99 then [target field OMOP] == 0	Refers to the result of the confirmatory treponemal test, performed prenatally. Confirmatory treponemal tests are indicated for diagnostic confirmation and exclusion of false positive non-treponemal test results. Treponemal tests are classified as: FTA-Abs (Fluorescent Treponemal Antibody-absorption) MHA-Tp (Microhemagglutination Treponema pallidum Assay) TPHA (Treponema pallidum Hemagglutination Assay) ELISA (Enzyme-Linked Immunosorbent Assay). (Null, Reactive, Non-reactive, Not performed, Ignored, Inconsistency)
procedure_type_concept_id			Apply 32879
procedure_source_value	tpconfirma_mae	if tpconfirma_mae == 0	

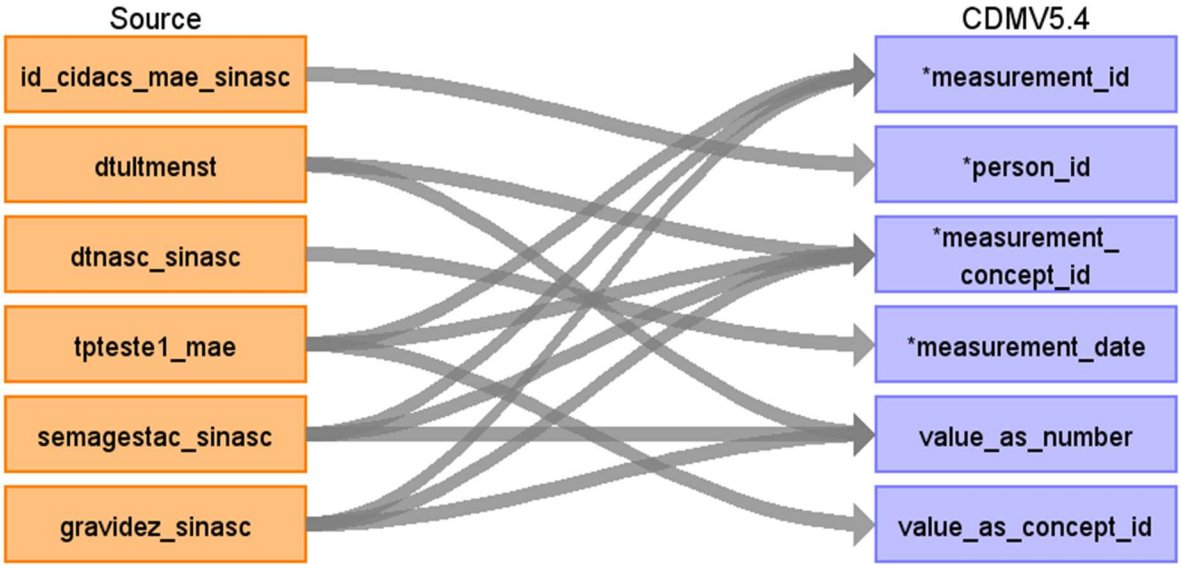
		then procedure_source_value = 0 if tpconfirma_mae == 1 then procedure_source_value = 4011063 (Reactive) if tpconfirma_mae == 2 then procedure_source_value = 4305306 (Non-Reactive) if tpconfirma_mae == 3 then procedure_source_value = 4118638 (Not-Performed) if tpconfirma_mae == 88 then procedure_source_value = 0 if tpconfirma_mae == 99 then procedure_source_value = 0	
procedure_date	dt_notific_mae dtnasc_sinasc	For concept tpconfirma_mae procedure_date == dt_notific_mae For concept parto_sinasc procedure_date == dtnasc_sinasc	

4.5 Table Name: CARE_SITE

Destination Field	Source Field	Logic	Comment
care_site_id	id_municip_mae	Extract two first digits from the variable 'id_municip_mae'. Code of the municipality where the health unit (or other reporting source) that made the notification is located.	Each instance of a Care_Site in the source data should be assigned this unique key.
care_site_source_value	id_municip_mae	care_site_source_value = id_municip_mae	The identifier of the care_site as it appears in the source data.

4.6 Table Name: MEASUREMENT

Para a criação do campo “measurement_concept_id”, foi necessário inserir um registro individual na tabela “measurement” para cada coluna de origem representando um conceito específico (concept). Esse processo foi realizado mantendo consistentes os valores dos demais campos, assegurando assim que cada entrada refletisse fielmente o conceito designado sem alterar as outras informações.



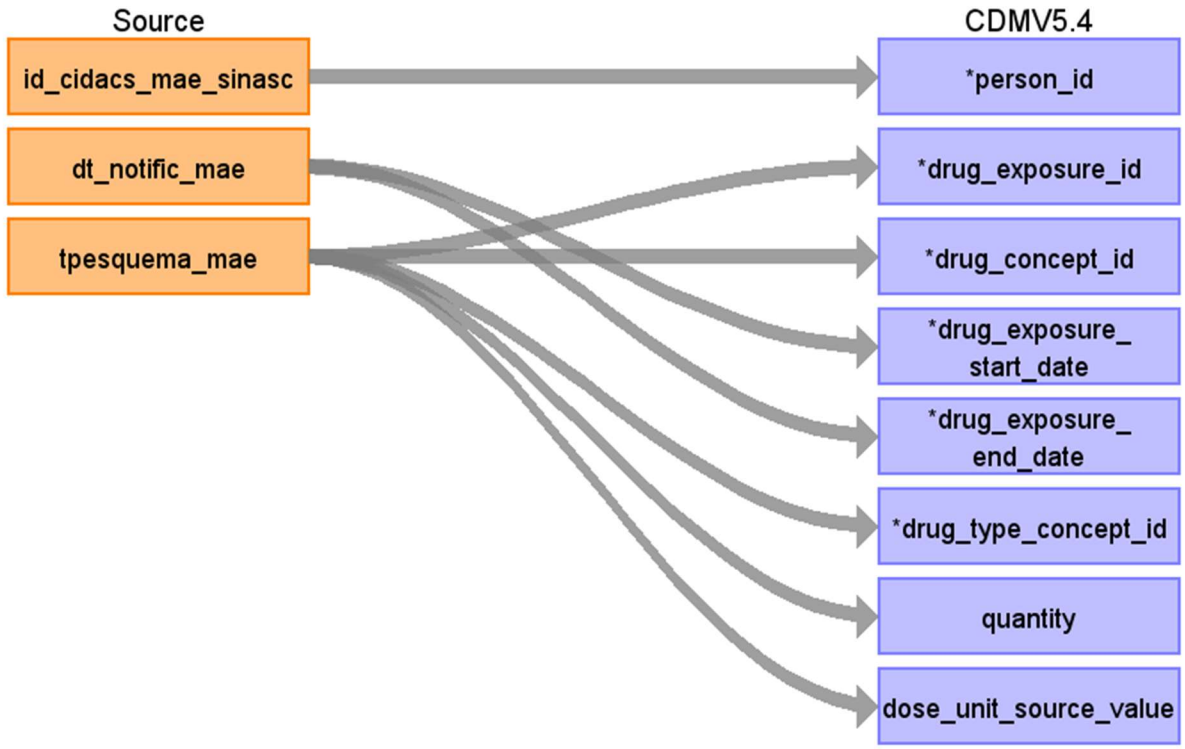
Destination Field	Source Field	Logic	Comment
measurement_id	tpteste1_mae semagestac_sinasc gravidez_sinasc	Create ID Create ID Create ID	Para criar o campo measurement_concept_id é preciso criar um registro na tabela measurement para cada coluna de origem (concept) mantendo o valor das demais iguais. Exemplo: a mesma lógica do Unnest do postgresql e bigquery / unpivot(melt) <Por Ricardo Félix> Create ID
person_id	id_cidacs_mae_sinasc	See Person table	See Person table
measurement_concept_id	semagestac_sinasc gravidez_sinasc tpteste1_mae dtultmenst	3012266 - Gestational age 4077859 - Number of fetuses 4299241 - VDRL test	0 – Nulo 1 – Única 2 – Dupla 3 – Tripla ou mais 88 – Ignorado

		dtultmenst == 4072438 (Date of last menstrual period)	Refers to the result of the non-treponemal test recommended for the first prenatal consultation. Non-treponemal tests are understood as VDRL and RPR. (reagent, non-reagent, not done)
measurement_date	dtnasc_sinasc	For concepts semagestac_sinasc and gravidez_sinasc use dtnasc_sinasc == measuremente_date For concept tpteste1_mae OR dtultmenst use mesprenat_sinasc == measuremente_date	
value_as_number	semagestac_sinasc gravidez_sinasc dtultmenst	semagestac_sinasc == value_as_number semagestac_sinasc == value_as_number if gravidez_sinasc == 0 then value_as_number == 0 if gravidez_sinasc == 1 then value_as_number == 1 if gravidez_sinasc == 2 then value_as_number == 2 if gravidez_sinasc == 3	0 – Nulo 1 – Única 2 – Dupla 3 – Tripla ou mais 88 – Ignorado 99 – Inconsistência

		then value_as_number == 3 if gravidez_sinasc == 88 then value_as_number == Null if gravidez_sinasc == 99 then value_as_number == Null IF measurement_concept_id == 4072438 THEN value_as_number == dtultmenst	
value_as_concept_id	tpteste1_mae	if tpteste1_mae == 0 then value_as_concept_id = 0 if tpteste1_mae == 1 then value_as_concept_id = 9191 (Positive) if tpteste1_mae == 2 then value_as_concept_id = 9189 (Negative) if tpteste1_mae == 3 then value_as_concept_id = 4118638 (Not-Performed) if tpteste1_mae == 88 then value_as_concept_id = 0 if tpteste1_mae == NA	

		then value_as_concept_id = 0	
measurement_type_concept_id			Apply 32879

4.7 Table Name: DRUG_EXPOSURE

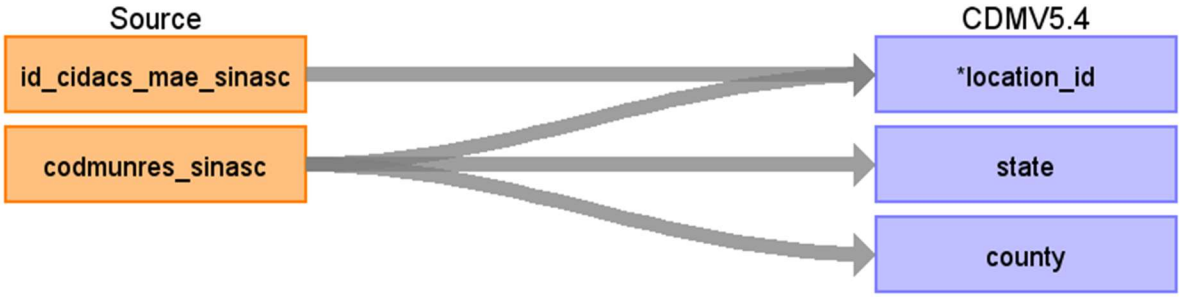


Destination Field	Source Field	Logic	Comment
person_id	id_cidacs_mae_sinasc	See Person table	See Person table
drug_exposure_id	tpesquema_mae		Create ID
drug_concept_id	tpesquema_mae	if tpesquema_mae == 1 or 2 or 3 then drug_concept_id == 1728416 (penicilina G) if tpesquema_mae == 4 then drug_concept_id == 1738521 (doxiciclina) if tpesquema_mae == 5	

		then drug_concept_id == 4118638 (not performed) if tpesquema_mae == 88 then drug_concept_id == 0 if tpesquema_mae == 99 then drug_concept_id == 0	
drug_exposure_start_date	dt_notific_mae	drug_exposure_start_date == dt_notific_mae	
drug_exposure_end_date	dt_notific_mae	if tpesquema_gen = 1 then drug_exposure_end_date = dt_notifc_mae + 1 (Dose unica) if tpesquema_gen = 2 then drug_exposure_end_date = dt_notifc_mae + 7 dias if tpesquema_gen = 3 then drug_exposure_end_date = dt_notifc_mae + 14 dias (03 doses com 1 semana de intervalo) if tpesquema_gen = 4 then drug_exposure_end_date = dt_notifc_mae+ 14 dias (pacientes com alergia à penicilina, tratados com doxiciclina) if tpesquema_mae = 5 then drug_exposure_end_date = dt_notifc_mae if tpesquema_mae = 88 then drug_exposure_end_date = dt_notifc_mae if tpesquema_mae = 99 then drug_exposure_end_date = dt_notifc_mae	
drug_type_concept_id	tpesquema_mae		Apply 32879
quantity	tpesquema_mae	if tpesquema_mae = 0	

		then quantity = Null if tpesquema_mae = 1 then quantity = 2400000 if tpesquema_mae = 2 then quantity = 4800000 if tpesquema_mae = 3 then quantity = 7200000 if tpesquema_mae = 4 then quantity = 100 if tpesquema_mae = 5 then quantity = 0 if tpesquema_mae = 88 then quantity = 0 if tpesquema_mae = 99 then quantity = 0	
dose_unit_source_value	tpesquema_mae	if tpesquema_mae = 0 then dose_unit_source_value = Null if tpesquema_mae = 1 OR 2 OR 3 then dose_unit_source_value = 8718 international unit (UI) if tpesquema_mae = 4 then dose_unit_source_value = 8576 (grams) if tpesquema_mae = 5 OR 88 OR 99 then dose_unit_source_value = Null	

4.8 Table Name: LOCATION

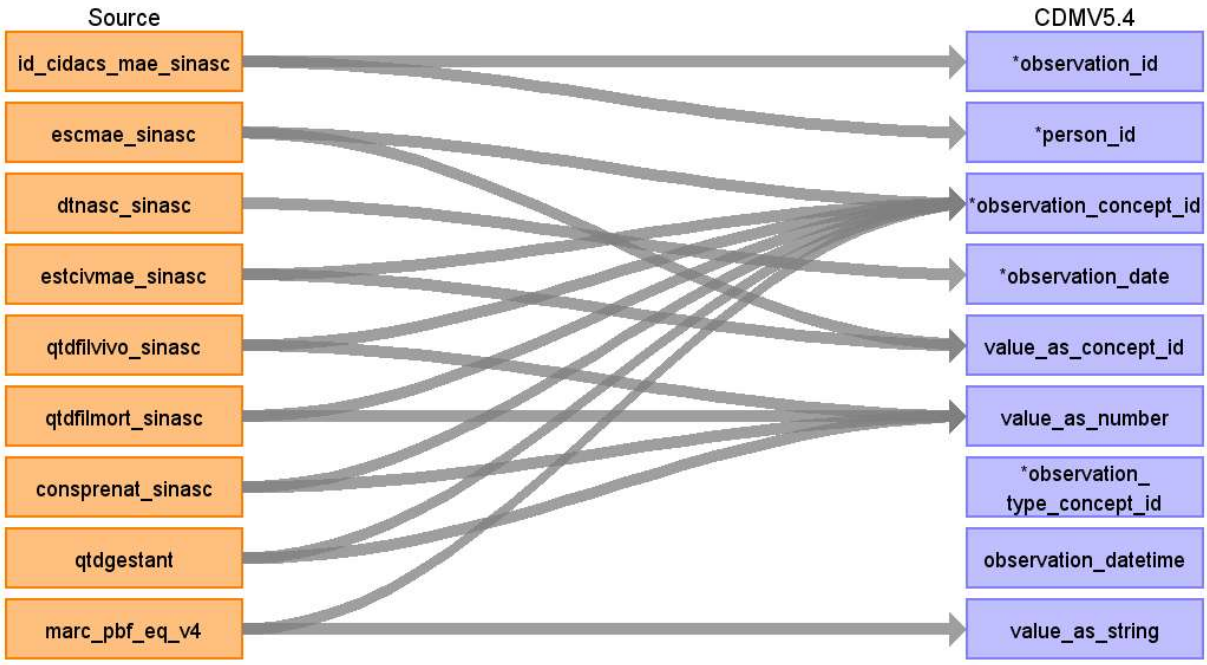


Destination Field	Source Field	Logic	Comment
location_id	id_cidacs_mae_sinasc codmunres_sinasc	Create ID Create ID	Each instance of a Location in the source data should be assigned this unique key.
state	codmunres_sinasc	Extract two first digits from the variable 'codmunres_sinasc'.	Unique identifier for brazilian municipalities. In this case the two first digits correponds to the state.
county	codmunres_sinasc	county == codmunres_sinasc	

4.9 Table Name: OBSERVATION

To create the “observation_concept_id” field, it was necessary to insert an individual record in the “observation” table for each source column representing a specific concept

(*concept*). This process was carried out while keeping the values of the other fields consistent, thus ensuring that each entry faithfully reflected the designated concept without altering the other information.



Destination Field	Source Field	Logic	Comment
observation_id	id_cidacs_mae_sinasc	Create ID	
person_id	id_cidacs_mae_sinasc	See Person table	See Person table
observation_concept_id	escmae_sinasc estcivmae_sinasc qtdfilvivo_sinasc qtdfilmort_sinasc consprenat_sinasc qtdgestant marc_pbf_eq_v4	4144287 - Schooling 4053609 - Marital status 45770046 (Número de nascidos vivos anteriores) 4090055 - Number of miscarriages or induced terminations of pregnancy 4313474 - Prenatal visit IF gestacao == 0	0 – Nulo 1 – Nenhuma 2 – 1 a 3 anos 3 – 4 a 7 anos 4 – 8 a 11 anos 5 – 12 e mais 88 – Ignorado 99 – Inconsistência 0 – Nulo 1 – Solteira

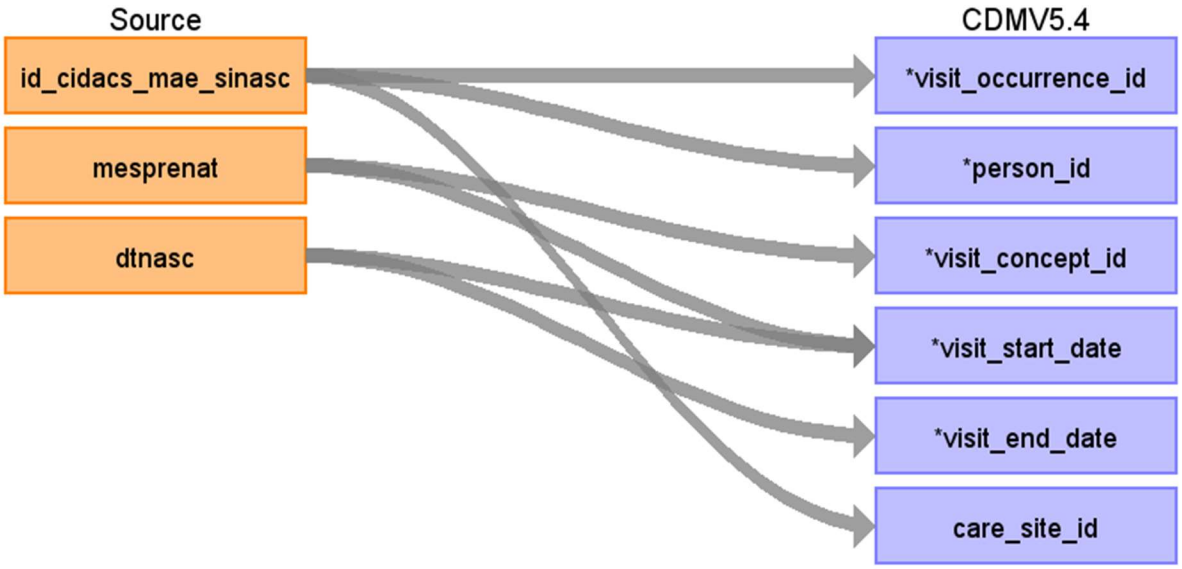
		THEN observation_concept_id == None marc_pbf_eq_v4 == 44804295 (Government policy)	2 – Casada 3 – Viúva 4 – Separada judicialmente/divorciada 5 – União estável 88 – Ignorada 99 – Inconsistência Number of previous live births 0 – Nulo 1 – Menos de 22 semanas 2 – 22 a 27 semanas 3 – 28 a 31 semanas 4 – 32 a 36 semanas 5 – 37 a 41 semanas 6 – 42 semanas e mais 88 – Ignorado Observation_id aglomera N conceitos distintos Para criar o campo observation_concept_id precisamos criar um registro na tabela observation para cada coluna de origem (concept) mantendo o valor das demais iguais. Ex: a mesma lógica do

			Unnest do postgresql e bigquery / unpivot(melt)
observation_date	dtnasc_sinasc	dtnasc_sinasc = observation_date	
value_as_concept_id	escmae_sinasc estcivmae_sinasc	if escame_sinasc = 0 then value_as_concept_id = 0 if escame_sinasc = 1 then value_as_concept_id = 45877495 (He/she never went to school) if escame_sinasc = 2 then value_as_concept_id = 37172863 (Less than a high school diploma) if escame_sinasc = 3 then value_as_concept_id = 37172863 (Less than a high school diploma) if escame_sinasc = 4 then value_as_concept_id = 37172863 (Less than a high school diploma) if escame_sinasc = 5 then value_as_concept_id = 45884461 (12th grade, no diploma) if escame_sinasc = 88 then value_as_concept_id = 0	

		if escame_sinasc = 99 then value_as_concept_id = 0 if estcivmae_sinasc = 0 then value_as_concept_id = 0 if estcivmae_sinasc = 1 then value_as_concept_id = 45879879 (Single) if estcivmae_sinasc = 2 then value_as_concept_id = 45876756 (Married) if estcivmae_sinasc = 3 then value_as_concept_id = 45883711 (Widowed) if estcivmae_sinasc = 4 then value_as_concept_id = 1620675 (Separated or Divorced) if estcivmae_sinasc = 5 then value_as_concept_id = 1620470 (Cohabitation) if estcivmae_sinasc = 88 then value_as_concept_id = 4052929 (Marital state unknown) if estcivmae_sinasc = 99	

		then value_as_concept_id = 0	
value_as_number	qtdfilvivo_sinasc qtdfilmort_sinasc consprenat_sinasc qtdgestant	IF observation_concept_id == 45770046 THEN value_as_number = qtdfilvivo_sinasc If observation_concept_id == 4090055 then value_as_number = qtdfilmort_sinasc if observation_concept_id == 4313474 then value_as_number == consprenat_sinasc IF observation_concept_id == 4132561 THEN value_as_number == qtdgestant	75 valores únicos
observation_type_concept_id			Apply 32879
value_as_string	marc_pbf_eq_v4	IF marc_pbf_eq_v4 == 1 THEN value_as_string == 4188539 (Yes) IF marc_pbf_eq_v4 == 2 THEN value_as_string == 4188540 (No)	

4.10 Table Name: VISIT_OCCURRENCE

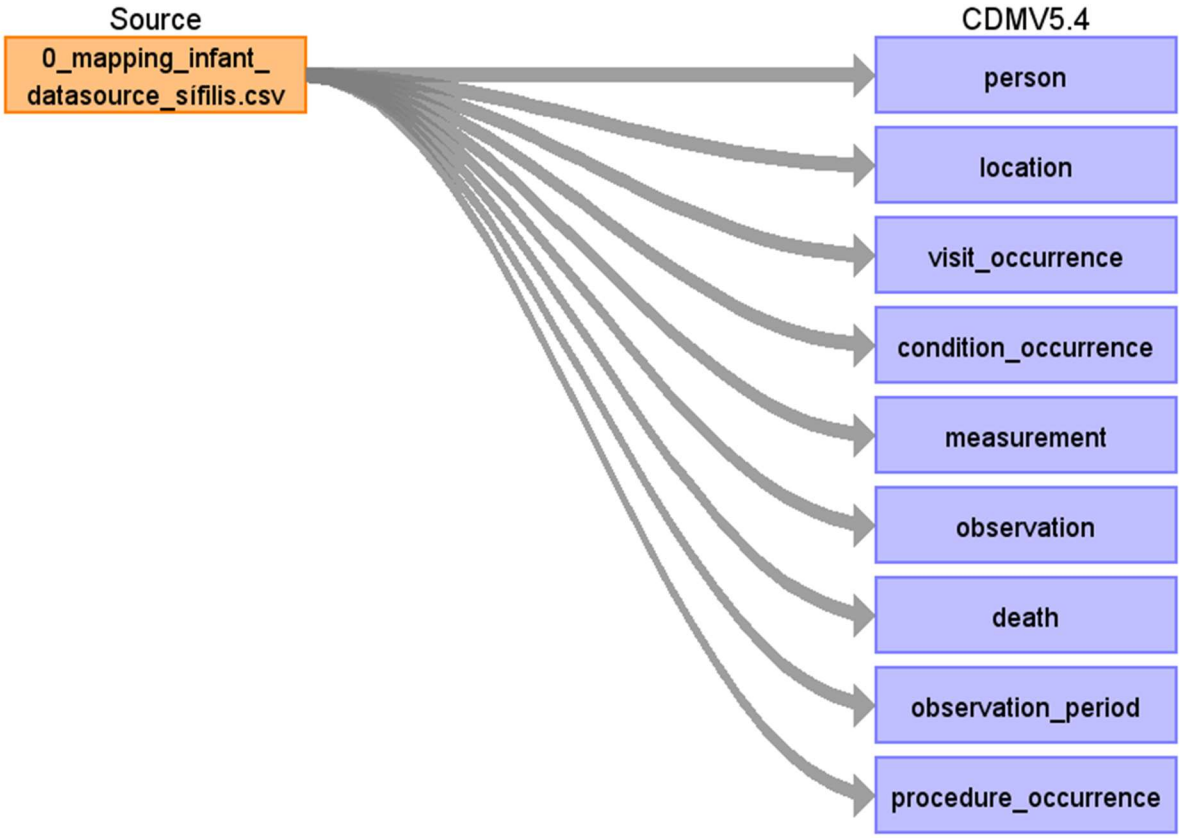


Destination Field	Source Field	Logic	Comment
visit_occurrence_id	id_cidacs_mae_sinasc	See Create ID	
person_id	id_cidacs_mae_sinasc	See Person table	
visit_concept_id	mesprenat	visit_concept_id == 9202 Concept ==> Outpatient Visit	
visit_start_date	dtnasc mesprenat	1.º extrair o mês da variável dtnasc 2.º reduzir do mês o valor encontrado na diferença: 9 - mesprenat 3.º extrair dia e ano do dtnasc 4.º Visit_start_date == dia dtnasc/mês encontrado no modelo acima /ano dtnasc *o mês corresponde a um mês do ano	
visit_end_date	dtnasc	visit_end_date == visit_start_date if visit_start_date == dtnasc	

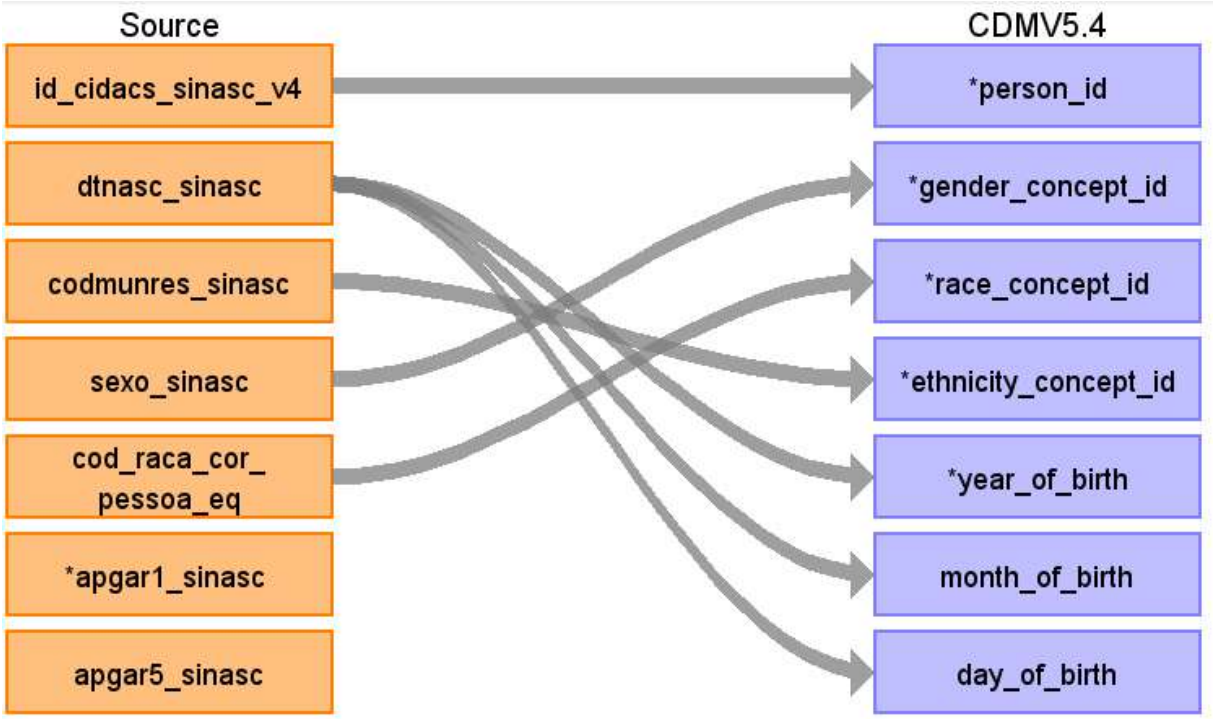
		then visit_end_date == dtnasc + 2 dias	
visit_type_concept_id			Apply 32879 - Registry
care_site_id	id_cidacs_mae_sinasc	See care_site table	

5.0 Source Data Mapping of Infant

Below, the design of the mapping process related with the ETL of infant data is presented under the format of images and tables extracted from the OHDSI tool named *Rabbit in a Hat.* The infant mapping included 09 template tables.



5.1 Table Name: PERSON

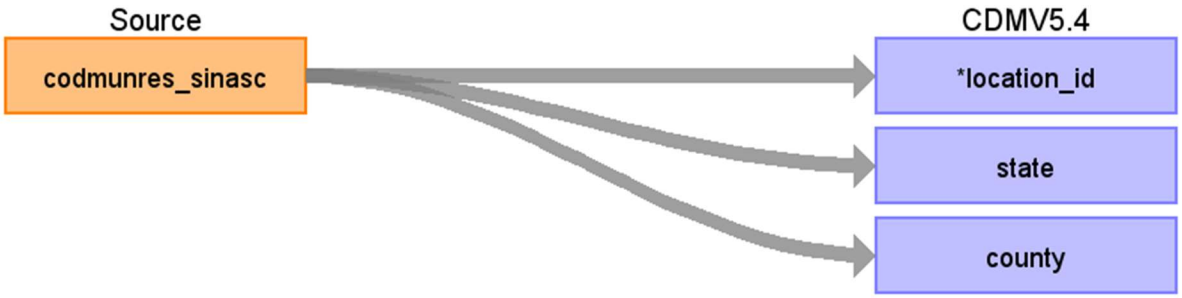


Destination Field	Source Field	Logic	Comment
person_id	id_cidacs_sinasc_v4	Ceate person_id based of id_cidacs_sinasc_v4	Unique sequential identifier of the Infant
gender_concept_id	sexo_sinasc	if sexo_sinasc == 0 then gender_conceptid = 0 if sexo_sinasc == 1 then gender_concept_id = 8507 if sexo_sinasc == 2 then gender_concept_id = 8532 if sexo_sinasc == 88 then gender_concept_id = 4214687 if sexo_sinasc == 99 then gender_concept_id = 0	Concepts CDM 8507 == Male 8532 == Female 4214687 == Gender unknown Concepts SOURCE 0 – Null 1 – Male 2 – Female 88 – ignored 99- Inconsistence

race_concept_id	Cod_raca_cor_pessoa_eq	if cod_raca_cor_pessoa_eq == 0 then race_concept_id == 0 if cod_raca_cor_pessoa_eq == 1 then race_concept_id == 8527 if cod_raca_cor_pessoa_eq == 2 then race_concept_id == 38003598 if racacorn_sinasc == 3 then race_concept_id == 8515 if racacorn_sinasc == 4 then race_concept_id == 4212311 if racacorn_sinasc == 5 then race_concept_id == 19387526 if racacorn_sinasc == 99 then race_concept_id == 0	
ethnicity_concept_id	codmunres_sinasc	create variable 'X' --> extract two first digits from the variable 'codmunres_sinasc' if 'X' varies from 11 to 53 then ethnicity_concept_id = 38003563 (Hispanic or Latino) else ethnicity_concept_id = 0	
year_of_birth	dtnasc_sinasc	To build this data, is required the remotion of day and month from the variable dtnasc_sinasc	
month_of_birth	dtnasc_sinasc	To build this data, is required the remotion of day and month from the variable dtnasc_sinasc	

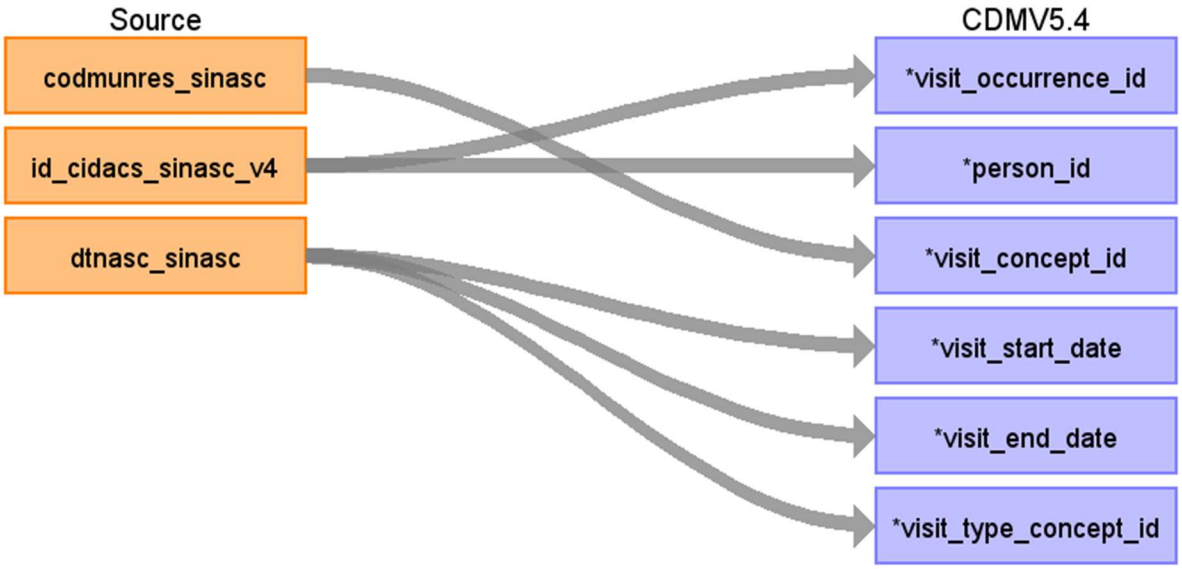
day_of_birth	dtnasc_sinasc	To build this data, is required the remotion of month and year from the variable dtnasc_sinasc	

5.2 Table Name: LOCATION



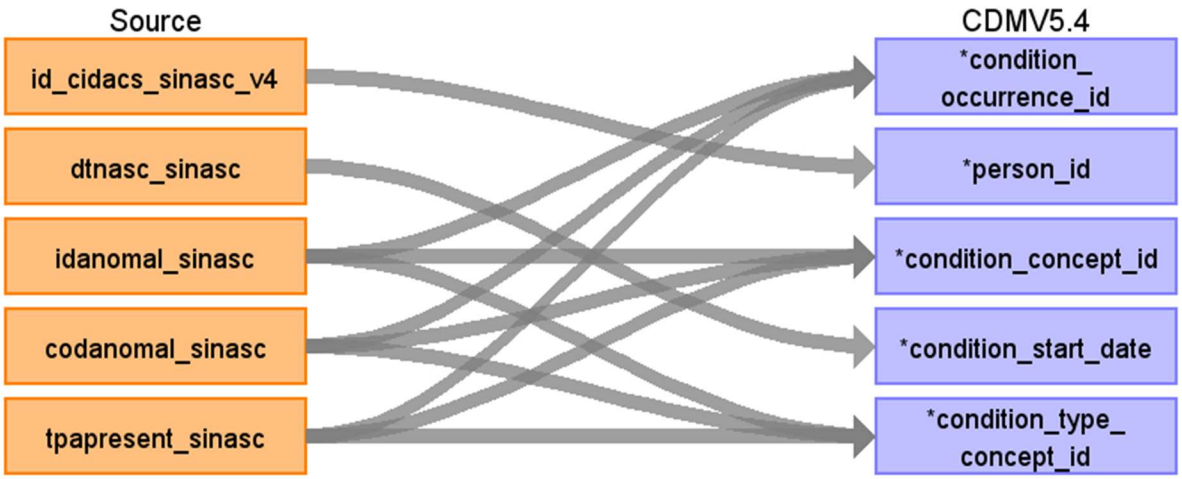
Destination Field	Source Field	Logic	Comment
location_id	codmunres_sinasc	Create ID	Each instance of a location in the source data must be assigned this unique key. Cada instância de um Local nos dados de origem deve receber esta chave exclusiva.
city			
state	codmunres_sinasc	Extract two first digits from the variable 'codmunres_sinasc'.	Unique identifier for brazilian municipalities. In this case the two first digits correponds to the state.
county	codmunres_sinasc	codmunres_sinasc = county	

5.3 Table Name: VISIT_OCCURRENCE



Destination Field	Source Field	Logic	Comment
visit_occurrence_id	id_cidacs_sinasc_v4	Create ID	
person_id	id_cidacs_sinasc_v4	See Person table	
visit_concept_id	Codmunres_sinasc	visit_concept_id == 8650 (Birthing Center)	
visit_start_date	dtnasc_sinasc	dtnasc_sinasc = visit_end_date	
visit_end_date	dtnasc_sinasc	Assign date of birth 'dtnasc_sinasc' + 2 days as visit_end_date	
visit_type_concept_id			Apply 32879

5.4 Table Name: CONDITION_OCCURRENCE



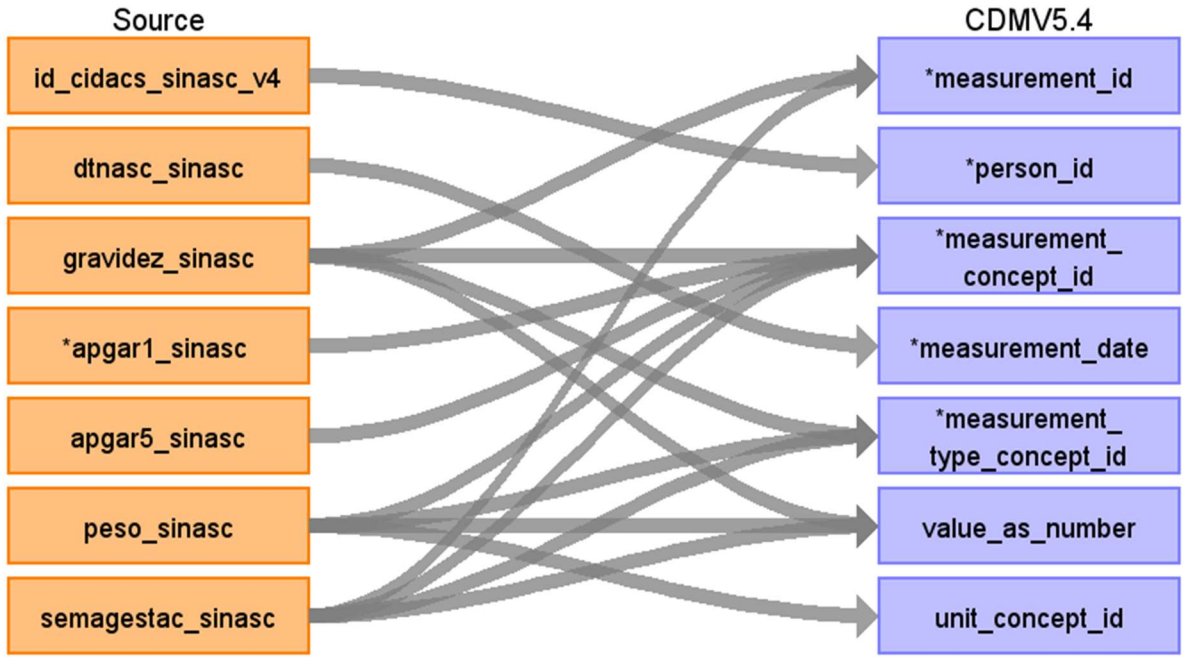
Destination Field	Source Field	Logic	Comment
condition_occurrence_id	codanomal_sinasc tpapresent_sinasc idanomal_sinasc	Create ID Create ID Create ID	
person_id	id_cidacs_sinasc_v4	See Person table	

condition_concept_id	codanomal_sinasc
	tpapresent_sinasc
	idanomal_sinasc
		Código da anomalia CID-10
		Mapear as ocorrências para o vocabulário nativo do CDM ICD-10
		
		Caso haja combinações de CIDs na mesma linha, quebrar em N registros/ocorrências
		if tpapresent_sinasc == 1 then:
		condition_concept_id = 4312344
		
		if tpapresent_sinasc == 2 then:
		condition_concept_id = 4195566
		
		if tpapresent_sinasc == 3 then:
		condition_concept_id = 81358	

		it tpapresent_sinasc == 9 OR 88 OR NA then:
		condition_concept_id = 0
		if idanomal_sinasc == 0 OR 2 OR 88:
		then condition_concept_id = 0
		if idanomal_sinasc == 1:
		then condition_concept_id = codanomal_sinasc	

condition_start_date	dtnasc_sinasc	condition_start_date == dtnasc_sinasc	
condition_type_concept_id	codanomal_sinasc tpapresent_sinasc idanomal_sinasc	Apply 32879 Apply 32879 Apply 32879	Apply 32879

5.5 Table Name: MEASUREMENT



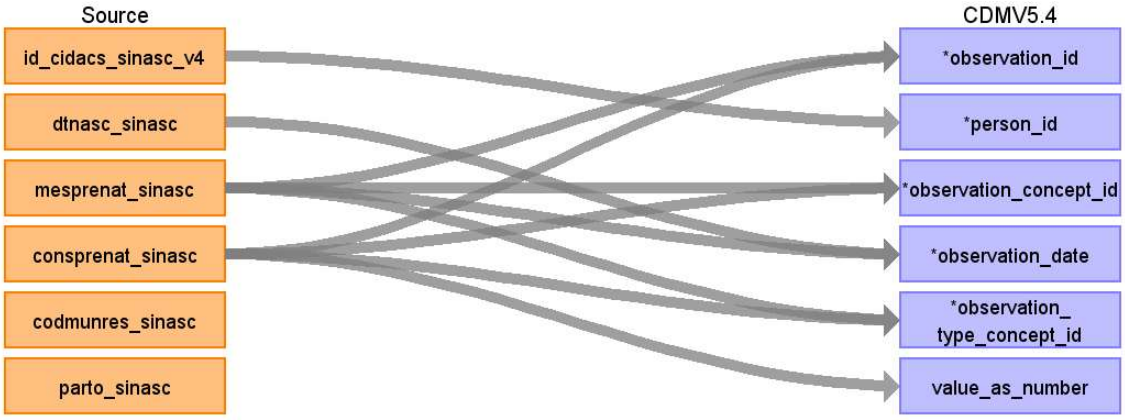
Destination Field	Source Field	Logic	Comment
measurement_id	gravidez_sinasc semagestac_sinasc	Create ID Create ID	
person_id	id_cidacs_sinasc_v4	See Person table	
measurement_concept_id	gravidez_sinasc apgar1_sinasc apgar5_sinasc semagestac_sinasc peso_sinasc	Apply 4077859 - Number of fetuses if apgar1_sinasc = 0 then measurement_concept_id = 4014468 (Apgar at 1 minute = 0) if apgar1_sinasc = 1 then measurement_concept_id = 4015430 if apgar1_sinasc = 2 then measurement_concept_id = 4015289 if apgar1_sinasc = 3 then measurement_concept_id = 4014305 if apgar1_sinasc = 4 then measurement_concept_id = 4015291 if apgar1_sinasc = 5 then measurement_concept_id = 4014469 if apgar1_sinasc = 6 then measurement_concept_id = 4014306 if apgar1_sinasc = 7 then measurement_concept_id = 4015292 if apgar1_sinasc = 8 then measurement_concept_id = 4016052 if apgar1_sinasc = 9 then measurement_concept_id = 4015431	Distincts no Datasource 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 19, 89, 99 e None Distincts no Datasource 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 99 e None

		if apgar1_sinasc = 10 then measurement_concept_id = 4014470 if apgar1_sinasc = 19 OR 89 OR 99 then measurement_concept_id = 4127662 (Inconsistent) if apgar1_sinasc = None then measurement_concept_id = 40482639 (Not detected by measurement) if apgar5_sinasc = 0 then measurement_concept_id = 4014307 (Apgar at 5 minutes = 0) if apgar5_sinasc = 1 then measurement_concept_id = 4016054 if apgar5_sinasc = 2 then measurement_concept_id = 4014308 if apgar5_sinasc = 3 then measurement_concept_id = 4016055 if apgar5_sinasc = 4 then measurement_concept_id = 4014471 if apgar5_sinasc = 5 then measurement_concept_id = 4014309 if apgar5_sinasc = 6 then measurement_concept_id = 4014310 if apgar5_sinasc = 7 then measurement_concept_id = 4016056 if apgar5_sinasc = 8 then measurement_concept_id = 4016465	

		if apgar5_sinasc = 9 then measurement_concept_id = 4016466 if apgar5_sinasc = 10 then measurement_concept_id = 4016467 if apgar5_sinasc = 99 then measurement_concept_id = 4127662 (Inconsistent) if apgar5_sinasc = None then measurement_concept_id = 40482639 (Not detected by measurement) Apply 3012266 Aaply 4264825 (Birth weight)	
measurement_date	dtnasc_sinasc	dtnasc_sinasc == measuremente_date	
measurement_type_concept_id	gravidez_sinasc semagestac_sinasc peso_sinasc	Apply 32879 Apply 32879 Apply 32879	Apply 32879
value_as_number	gravidez_sinasc semagestac_sinasc peso_sinasc	if gravidez_sinasc == 0 then value_as_number == 0 if gravidez_sinasc == 1 then value_as_number == 1 if gravidez_sinasc == 2 then value_as_number == 2 if gravidez_sinasc == 3 then value_as_number == 3 if gravidez_sinasc == 88 OR 99 then value_as_number == Null	

		semagestac_sinasc == value_as_number value_as_number = peso_sinasc	
unit_concept_id	peso_sinasc	unit_concept_id = 8504 (gram)	

5.6 Table Name: OBSERVATION

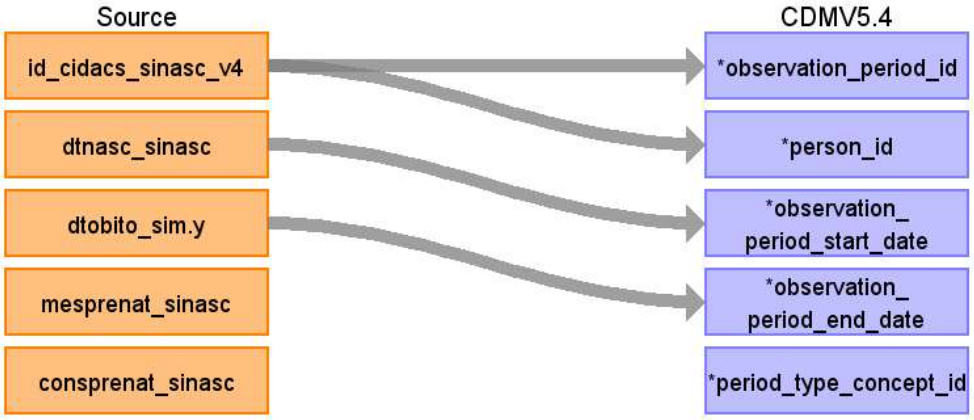


Destination Field	Source Field	Logic	Comment
observation_id	mesprenat_sinasc consprenat_sinasc	Create ID Create ID	
person_id	id_cidacs_sinasc_v4	See Person table	
observation_concept_id	mesprenat_sinasc consprenat_sinasc		
observation_date	dtnasc_sinasc mesprenat_sinasc	dtnasc_sinasc = observation_date observation_date == mesprenat_sinasc	
observation_type_concept_id	mesprenat_sinasc consprenat_sinasc	32879	
value_as_number	consprenat_sinasc	Value_as_number == consprenat_sinasc	

5.7 Table Name: DEATH

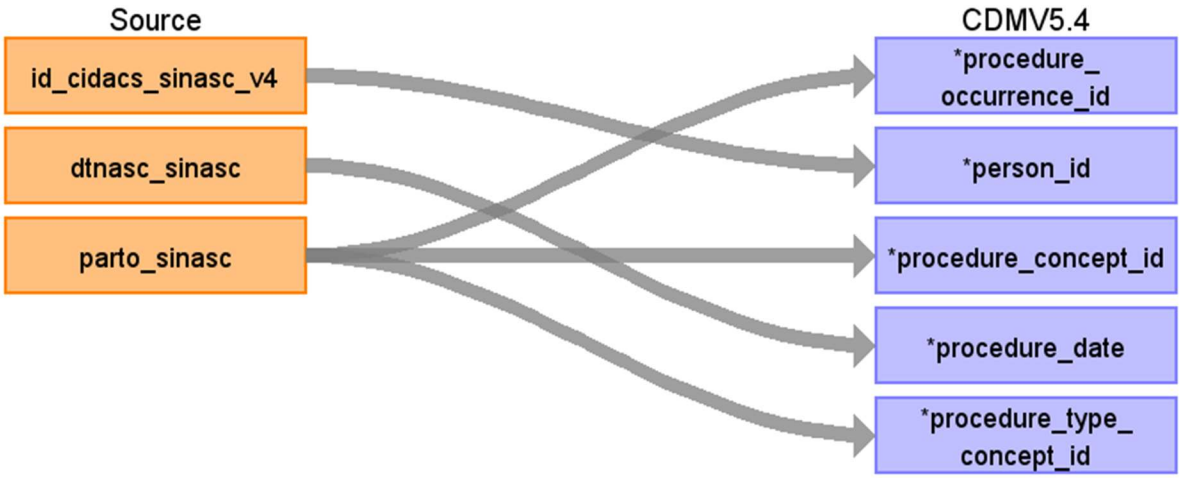
Destination Field	Source Field	Logic	Comment
person_id	id_cidacs_sinasc_v4	See Person table	
death_date	dtobito_sim.y	death_date = dtobito_sim.y	
death_type_concept_id			Apply 32879

5.8 Table Name: OBSERVATION_PERIOD



Destination Field	Source Field	Logic	Comment
observation_period_id	id_cidacs_sinasc_v4	Create ID	
person_id	id_cidacs_sinasc_v4	See Person table	
observation_period_start_date	dtnasc_sinasc	observation_period_start_date = dtnasc_sinasc	
observation_period_end_date		observation_period_end_date = 31/12/2020	
period_type_concept_id			Apply 32879

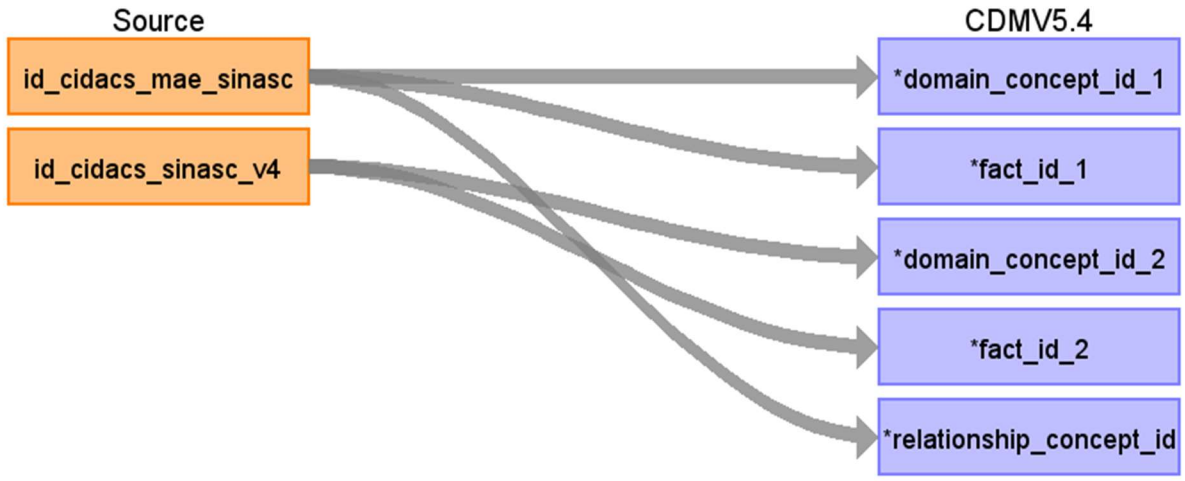
5.9 Table Name: PROCEDURE_OCCURRENCE



Destination Field	Source Field	Logic	Comment
procedure_occurrence_id	parto_sinasc	Create ID	
person_id	id_cidacs_sinasc_v4	See Person table	
procedure_concept_id	parto_sinasc	if parto_sinasc == 0 OR 88 OR 99 then procedure_concept_id == 0 if parto_sinasc == 1 then procedure_concept_id == 44784097 (Vaginal delivery of fetus) if parto_sinasc == 2	

		then procedure_concept_id== 4015701 (Cesarean section)	
procedure_date	dtnasc_sinasc	For concept parto_sinasc use dtnasc_sinasc == procedure_data	
procedure_type_concept_id	parto_sinasc	Apply 32879	

6.0 Source Data Mapping Fact_Relationship



Destination Field	Source Field	Logic	Comment
domain_concept_id_1	id_cidacs_mae_sinasc	4248584 (Mother)	
fact_id_1	id_cidacs_mae_sinasc	Create ID mother of [ID virá da tabela Person]	
domain_concept_id_2	id_cidacs_sinasc_v4	4285883 (Child)	
fact_id_2	id_cidacs_sinasc_v4	Create ID child of [ID virá da tabela Person Infant]	
relationship_concept_id	id_cidacs_mae_sinasc	4277283 (Natural Mother)	

7.0 Challenges and Perspectives in Implementing the OMOP Model

The main challenges faced in implementing the OMOP v5.4 data model in the context of CIDACS were: maintaining the quality and heterogeneity of the data in the mapping process, compatibility with terminologies and technical training and resources.

The OMOP model offers a structure that facilitates multicentre studies, enhances analyses of large volumes of data and promotes a broader view of public health issues. With the adoption of more recent versions and continuous improvements, the use of OMOP in collaborative research at CIDACS is expected to increase.

OMOP is increasingly being integrated with predictive analysis and artificial intelligence (AI) technologies. This paves the way for faster and more accurate analyses, with possibilities for automated cohort studies, trend identification, among other innovations, and this can be seen as an opportunity for the Centre. Continuously developing the technical capacity of the team and sharing good practice between health professionals and data scientists and engineers is a priority. This will help overcome barriers to entry and ensure that the team can conduct further implementations of the OMOP model in other projects and initiatives, strengthening interoperability and collaborating more broadly and meaningfully with CIDACS' mission.

7.1 Data Quality and Heterogeneity

There was a joint effort between working groups 1 and 2 to avoid losing information or introducing errors into the OMOP database. The model's ETL conventions restrict or amplify the multidisciplinary team's challenges insofar as record exclusion definitions are necessary and can impact on the search results. These scenarios need to be identified before implementation to avoid rework. The joint action of a multidisciplinary team is essential at this stage to control changes in the implementation documentation, losses of data granularity and biases that can significantly affect the research results.

Exploring the White Rabbit tool is crucial in the run-up to ETL implementation to subsidise the multidisciplinary team's discussions on the necessary and pressing decisions for the ETL execution phase.

The documentation must be kept up to date so that there is no loss of historicity and decisions for possible doubts that may arise in future stages.

Controlling changes in mapping decisions, concepts, vocabularies, etc. is crucial when the modelling work is carried out by an interdisciplinary group and the dynamism of the ETL

implementation phase itself. The Data Engineer needs the attention of this group so as not to create obstacles in the implementation phase when the model needs to be revised.

7.2 Compatibility of Terminologies

Adopting CDM OMOP, which focuses predominantly on modelling clinical data, represented a significant challenge for the team. CIDACS, which has a robust focus on large linked databases, is designed to analyse the impact of social, economic and environmental factors on health inequalities. Unlike environments where OMOP is largely applied to standardised clinical data, the complexity of CIDACS lies in the use of administrative and population data, which requires adaptations to the model to accommodate information from non-clinical sources.

This perspective challenges the multidisciplinary team to adapt OMOP to effectively exploit heterogeneous and interconnected data that goes beyond the clinical context, encompassing social determinants of health and economic data. The need to adequately represent socioeconomic variables required the development of customised mappings and strategies to ensure that the nuances of this information were captured without losing analytical rigour.

Therefore, the implementation of OMOP in the context of CIDACS illustrates both the opportunities and limitations of the model when dealing with large-scale public health and inequality issues. Integrating these adaptations into OMOP can broaden the model's applicability to other population health research projects, contributing to more comprehensive and holistic analyses of health inequalities.

The aim is to structure an inclusive table adapted to the context of the Global South in order to consistently incorporate socioeconomic variables

OMOP v5.4 relies heavily on standardised terminology, such as SNOMED CT, LOINC and RxNorm. Not being familiar with these terminologies requires efforts to translate and adapt them to local terminology, without losing the semantics of the original data.

7.3 Technical expertise and resources

The implementation process requires specialised technical knowledge in ETL and programming, as well as access to robust mapping and storage tools. The presence of a Data Engineer in a face-to-face and immersive activity with the dedicated multidisciplinary team at CIDACS was a differentiator during the ETL implementation stage. This can be a challenge, especially with shared human resources and people undergoing training and knowledge development in health data modelling.

The formation of a dedicated multidisciplinary team is crucial to the quality of the mapping, which requires in-depth knowledge of the source and target data.

In the context of CIDACS, preparing an analysis environment with OHDSI tools facilitated working together. The configuration of the OHDSI environment on AWS was not finalised and, due to changes during the course of the project, this had no impact on the delivery of the WG1 product.

The design of a solution and the creation of a specific computer infrastructure to install the OHDSI tools and carry out the activities inherent to the dedicated team and the construction of the database was essential to the completion of the work.

Given that there is no record of any other institution carrying out similar work in an ON-PREMISE environment, i.e. in a local infrastructure where the data is stored, processed and analysed within the centre's physical servers, without relying on cloud services, this stage was challenging and carried out by the CIDACS Data Platform teams.

At CIDACS, this model is used to guarantee security, privacy and total control over the sensitive data that is processed. Compliance with ethical and regulatory standards for the protection of health data is one of the essential guidelines at the centre.

8.0 Toolbox

The OHDSI tools used in this project are described below:

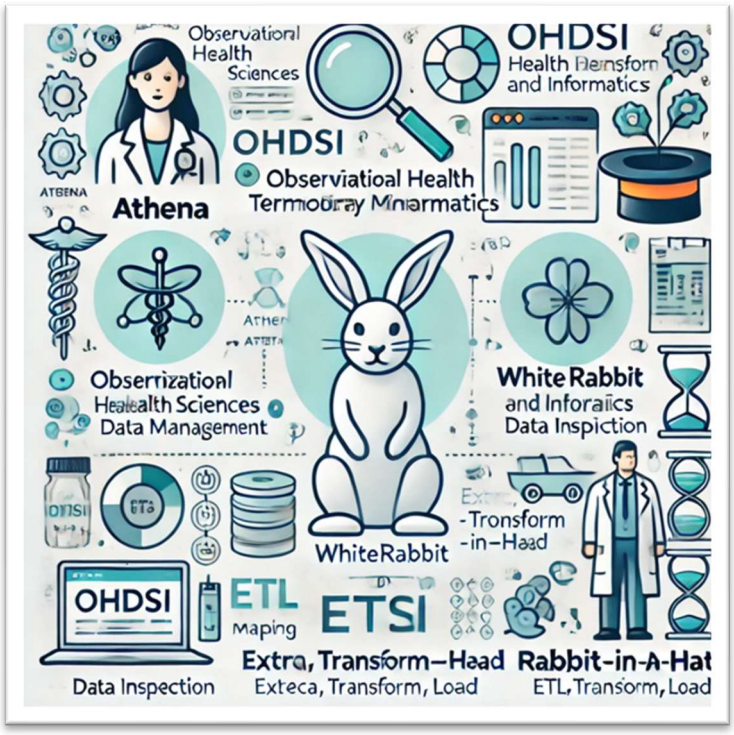
ATHENA: used to manage vocabularies and terminologies. It allows you to search for concepts and standardised vocabularies.

WHITERABBIT: used for data inspection, helps with ETL preparation.

RABBIT-IN-A-HAT: used for ETL mapping and documentation.

DATA QUALITY DASHBOARD: used to assess and monitor the quality of the data that has been transformed for OMOP CDM.

USAGI: used to help in the process of mapping codes from a source system to OMOP standard terminologies.



EXCEL: It was not possible to carry out this work together using the Rabbit in a Hat tool. So Excel was the additional tool chosen and used for mapping source and destination data, especially for inter-institutional alignment between CIDACS and the PHDC (The Provincial Health Data Centre) team. The concepts were transformed into Natural Language and the OMOP model was replicated in the Excel spreadsheet, thus enabling joint work, alignment in the use of concepts for federated analysis, exchange of experiences, active discussions in all phases for the construction of the OMOP database.

14 Mapping (PHDC) | 7 Mapping (CIDACS) | 6 Natural fields comparison | Target + template | Source (CIDACS) | Source (PHDC) | Vocabularies

ETL 100%	Natural language target field CIDACS	OMOP target field	Mother / infant	OMOP target field in sync	Core analysis	Candidate number	Type	Concept name	Target Table	Required?	Vocabulary ID	Concept ID
☑	Mother ID	person_id	Mother	☑	☑	1	int		person	*		
☑	Mother Observation Period	observation_period_id	Mother	☑	☑	1	int		observation_period	*		
		period_type_concept_id	Mother	☑	☑	1	int	Registry	observation_period	*	Type Concept	32879
		person_id	Mother	☑	☑	1	int		observation period person			
		observation_period_start_date	Mother	☑	☑	1	date		observation_period	*		
		observation_period_end_date	Mother	☑	☑	1	date		observation_period	*		
☑	Mother gender	gender_concept_id	Mother	☑	☑	1	int	MALE FEMALE GENDER UNKNOWN	person	*	SNOMED	8507 8532 4214687
☑	Mother's race	race_concept_id	Mother	☑	☑	1	int	WHITE BLACK ASIAN MIXED RACIAL GROUP INDIGENOUS PEOPLE	person	*	RACE SNOMED MESH	8527 38003598 8515 4212311 19387526
☑	Mother's ethinicity	ethnicity_concept_id	Mother	☑	☑	1	int	HISPANIC OR LATINO	person	*	ETHNICITY	38003563
☑	Mother's residence (state / municipality	location_id	Mother	☐	☑	1	int		location | person	*		
		state	Mother	☐	☑	1	varchar		location			
		county	Mother	☐	☑	1	varchar		location			

Mapping (CIDACS) - Mother

ETL 100%	Natural language target field CIDACS	OMOP target field	Mother / infant	OMOP target field in sync	Core analysis	Candidate number	Type	Concept name	Target Table	Required?	Vocabulary ID	Concept ID
☑	Fetal presentation (occiput anterior, posterior, or breech, etc.)	condition_occurrence_id	Infant	☐	☐	1			condition_occurrence	*		
		condition_concept_id	Infant	☐	☐	1			condition_occurrence	*		
		condition_type_concept_id	Infant	☐	☐	1		Registry	condition_occurrence	*	Type Concept	32879
		condition_concept_id	Infant	☐	☐	1			condition_occurrence	*	SNOMED	4312344
		condition_start_date	Infant	☐	☐	1			condition_occurrence	*		
☑	Gestational age at birth	measurement_concept_id	Infant	☐	☐	1	int	Gestational age	measurement	*	LOINC	3012266
		measurement_date	Infant	☐	☐	1	date		measurement	*		
		value_as_number	Infant	☐	☐	1	int		measurement			
		measurement_type_concept_id	Infant	☐	☐	1	int	Registry	measurement	*	Type Concept	32879
☑	Prenatal care (at least 4)	observation_concept_id	Infant	☐	☑	1	int	Prenatal visit	observation	*	SNOMED	4313474
		observation_date	Infant	☐	☑	1	date		observation	*		
		observation_type_concept_id	Infant	☐	☑	1	int	Registry	observation	*	Type Concept	32879
		value_as_number	Infant	☐	☑	1	float		observation			
☑	First ANC visit (within 20 weeks")	observation_concept_id	Infant	☐	☑	1	int	Prenatal initial visit	observation	*	SNOWMED	4311447
		observation_date	Infant	☐	☑	1	date		observation	*		
		observation_type_concept_id	Infant	☐	☑	1	int	Registry	observation	*	Type Concept	32879
☑	Type of delivery	procedure_concept_id	Infant	☐	☑	1	int	Vaginal delivery of fetus Cesarean section	procedure_occurrence	*	SNOMED	44784097 4015701

Mapping (CIDACS) - Child

Considerations on using the USAGI tool

The team chose to use Athena as the tool for mapping the terminologies in the original database to OMOP concepts. However, the query is made one at a time and not in batches, as the Usagi tool allows.

Nevertheless, for the implementation of this ETL, the Usagi tool was recommended by the consultancy.

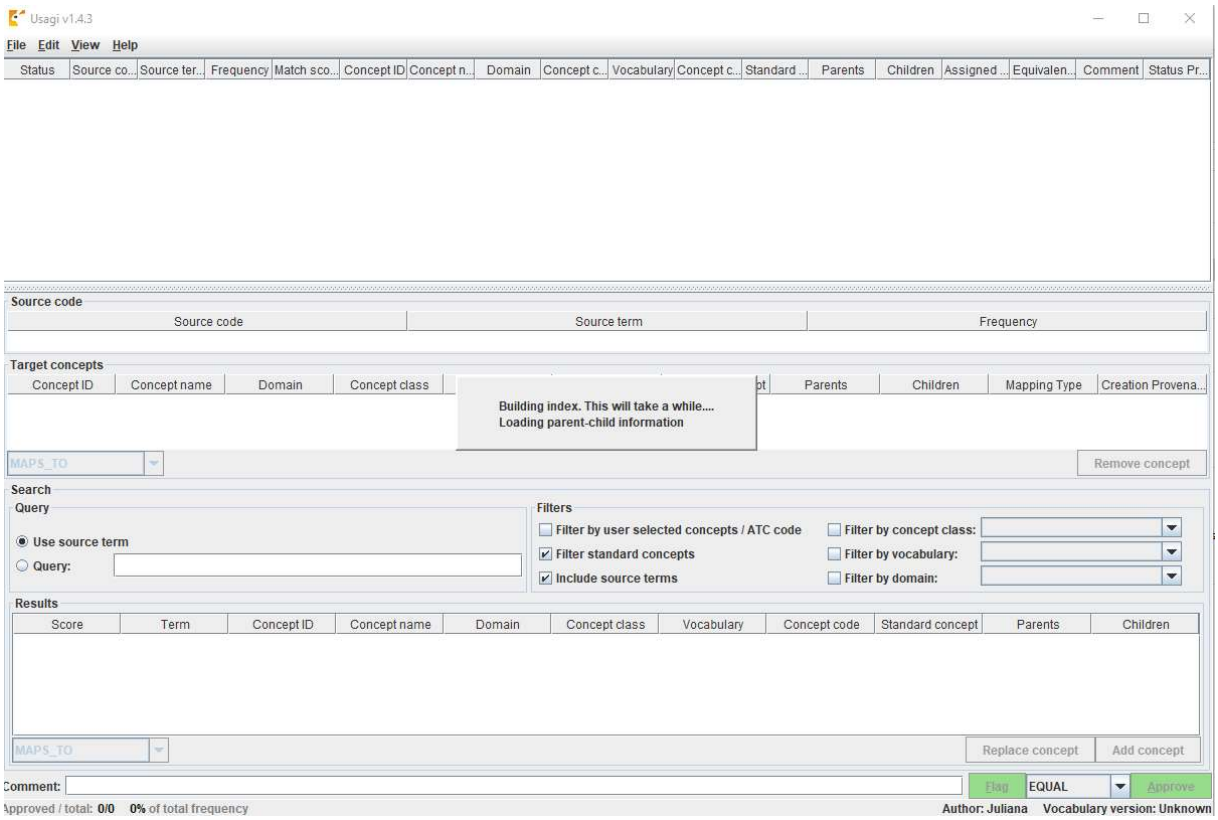
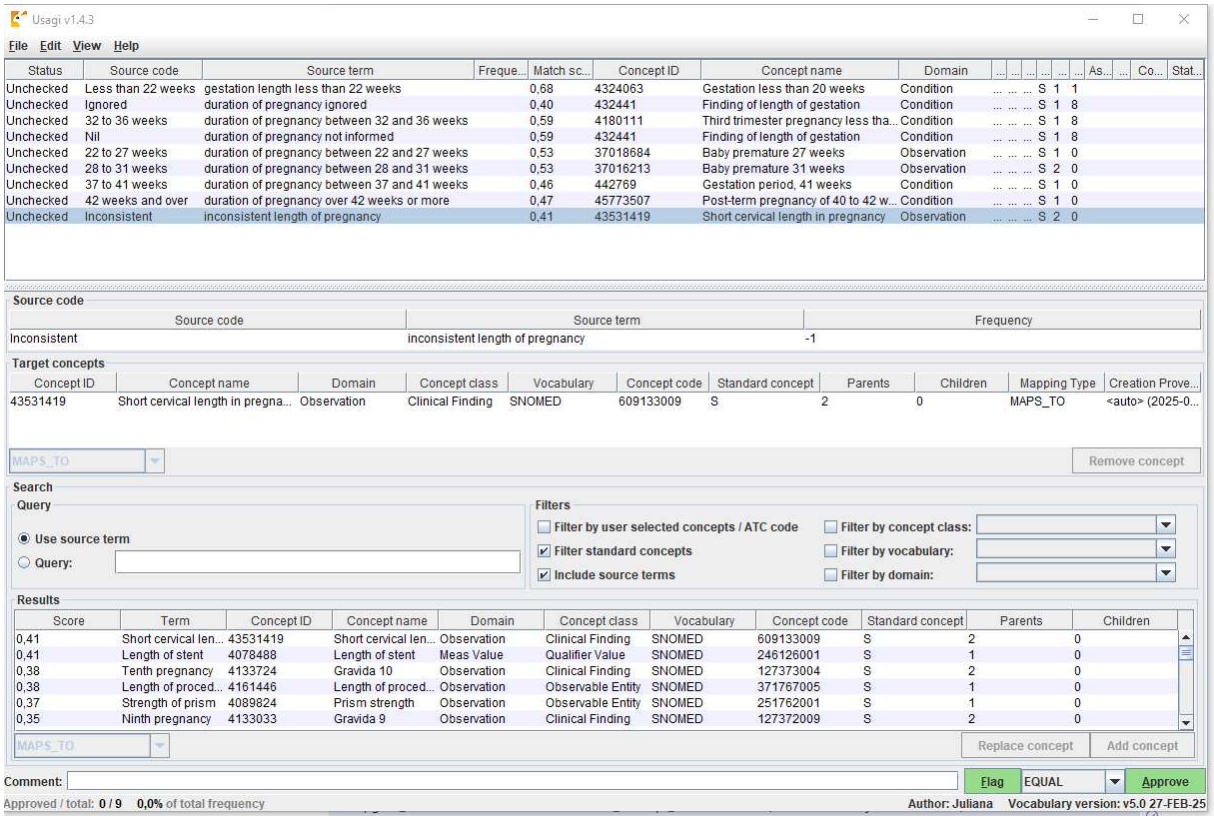


Image shows the vocabulary indexing stage in Usagi.

Flags: Equal; equivalent; wide; narrower; inexact; unmatched; unreviwed



The ETL process was developed using Python and PySpark, ensuring scalability and efficiency in data processing. The data was stored in a PostgreSQL database, providing structured and secure access to the converted data.

9.0 Computational Infrastructure

In the context of observational studies carried out by the OHDSI community, cloud computing infrastructure is the most widely used and widely adopted for transforming data into OMOP CDM and using analytical tools.

CIDACS follows a strict protocol for accessing and using sensitive data through a controlled analysis environment to guarantee security, confidentiality and traceability. Given this scenario, the use of a cloud computing environment was ruled out for manipulating, transforming and analysing the data.

In order to adopt the OMOP Common Data Model (CDM) and the OHDSI toolset at CIDACS, an in-house solution had to be built to enable the implementation of ETL and the use of the tools in a controlled ecosystem, without the possibility of exporting individually identifiable or pseudo-anonymised data, in accordance with best data protection practices.

The infrastructure of the analysis environment formatted for the OHDSI tools represented a breakthrough in the standardisation and interoperability of the data analysed at CIDACS.

In this way, we obtain from this environment

- Strict authentication and access control policies.

- Guarantees computational isolation between projects.

- Allows complete auditing of actions taken.