# ETL Discussion

**ETL Team:** Amelia Taylor, Gonjetso Chinyama, Moses Gwaza   
Document created by: Amelia Taylor

## **CASE 1:** OBSERVATION: Observation Table

The first question we had was whether one should use the **observation** tables only? The Pentaho ETL for WHO populated the visit table too.   
  
Rational would be that the survey responses and tests are not visits but observations. In theory, the observation period domain contains records which uniquely define the spans of time for which a person is at-risk to have clinical events recorded within the source systems. The visit domain contains the spans of time a person continuously receives medical services from one or

more providers at a care site in a given setting within the health care system. Visits are classified into 4 settings: outpatient care, inpatient confinement, emergency room, and long-term care.

In the current ETL the Observation table contains 7 observations per person. These correspond to two observation\_concept\_ids:

* 40285617 (stands for observation procedure, non-standard concept ) is of type Procedure
* 3579792 (this is deprecated, non-standard concept) and is of type Admin

For each survey the following are recorded as Observations:

1. Current residence (observation\_type\_concept\_id=1032160, LOINC) value\_as\_string=Hospital
2. Location (observation\_type\_concept\_id= 4109688 SNOMED representing city) value\_as\_string = city code or name
3. Occupation (observation\_type\_concept\_id=4033543, SNOMED) representing occupation
4. Rural Road (environment ( observation\_type\_concept\_id= 4119867 SNOMED) this being either Urban or Rural
5. Age (observation\_type\_concept\_id = 4265453 SNOMED)
6. Name (observation\_type\_concept\_id=45766362 SNOMED)
7. Census district ( observation\_type\_concept\_id = 44803989 SNOMED)

2-7 are recorded as Procedures and 1 is recorded as Admin.

**For the MW data we also need to map the remaining observations which refer to pregnancy for example, including all the medical history. These do not appear in the WHO.**

## **CASE 2:** VISIT- Mapping MW IDSR into **visit\_occurrence** and **visit\_detail** tables.

We are not convinced that our data from the IDSR qualifies for visit\_occurrences but the Line List does contain visits (hospitalisations and treatment data).

**The ETL for WHO synthetic data**

We have followed the ETL for these fields from the WHO Pentaho ETL. All visits were assigned the same constants:

* **visit\_concept\_id** = “4139501” (representing “Health Center”) and
* **visit\_type\_concept\_id** = “32809” (representing “Case Report Form”).

These were applied as constants in the Pentaho ETL transformations.

| 32809 | visit\_type\_concept\_id | 005 populate visit occurrence/001 populate condition occurrence/001 populate observation | Case Report Form | <https://athena.ohdsi.org/search-terms/terms/32809> |
| --- | --- | --- | --- | --- |
| 4139501 | visit\_concept\_id |  |  | <https://athena.ohdsi.org/search-terms/terms/4139501> |

Of course in our data we have different types of visits including POE. Some visits are not taking place at a health facility but (in the case of mass testing) on the site of organisations such as schools/ prisons/ banks etc.

According to the [documentation:cdm:visit\_occurrence [Observational Health Data Sciences and Informatics] (ohdsi.org)](https://www.ohdsi.org/web/wiki/doku.php?id=documentation:cdm:visit_occurrence) the following apply:

| visit\_source\_value | No | string(50) | The source code for the visit as it appears in the source data. |
| --- | --- | --- | --- |
| visit\_source\_concept\_id | No | Integer | A foreign key to a Concept that refers to the code used in the source. |

| visit\_concept\_id | Yes | integer | A foreign key that refers to a visit Concept identifier in the Standardized Vocabularies. |
| --- | --- | --- | --- |

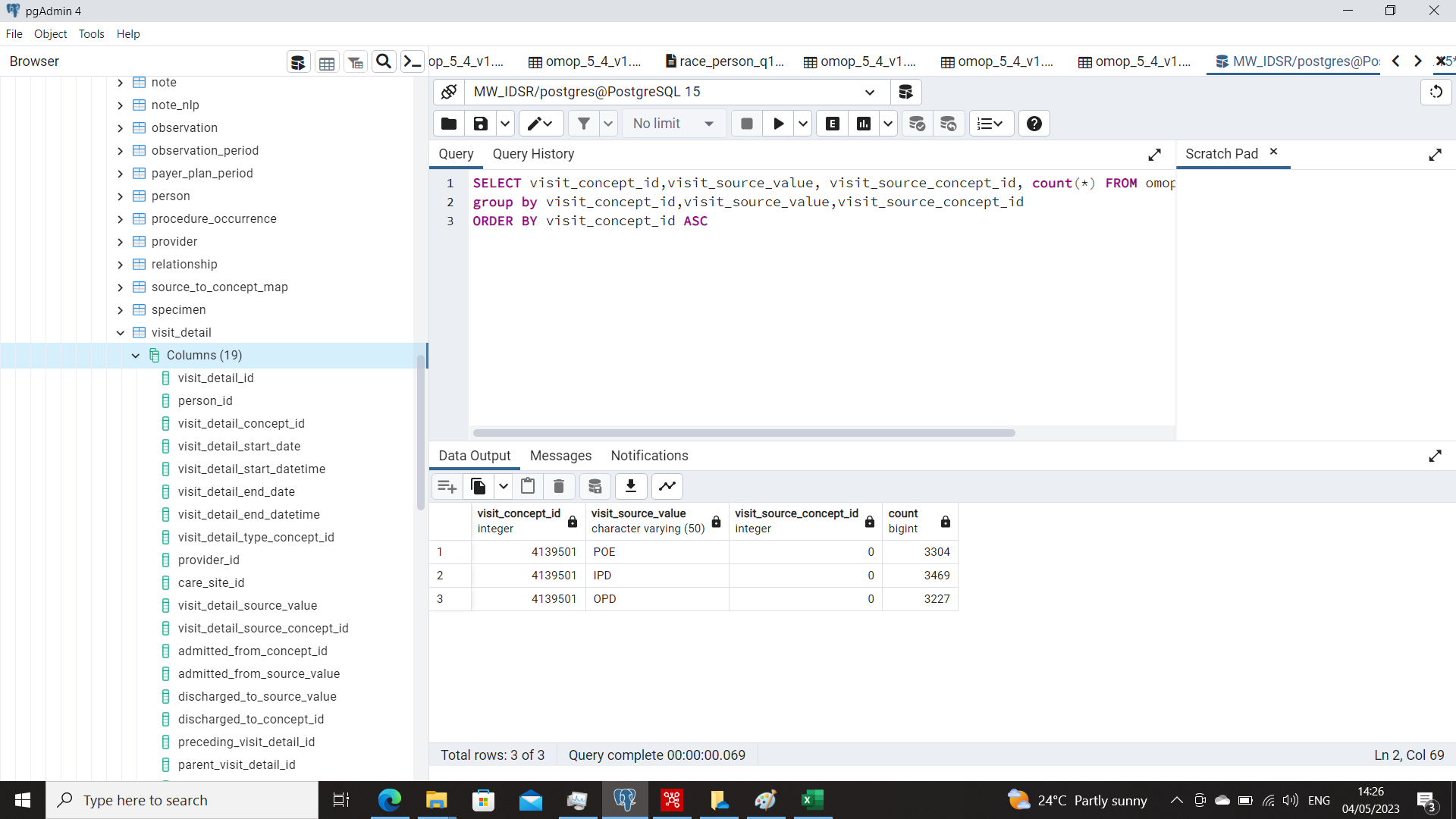
| visit\_type\_concept\_id | Yes | Integer | A foreign key to the predefined Concept identifier in the Standardized Vocabularies reflecting the type of source data from which the visit record is derived. |
| --- | --- | --- | --- |

However, as the Pentaho ETL did not use vocabularies, visit\_source\_concept\_id and visit\_type\_concept\_id and visit\_concept\_id are not applied yet foreign keys constraints. These values may likely need changing and checking once vocabularies are used.

*SELECT visit\_concept\_id,visit\_source\_value, count(\*) FROM omop\_5\_4\_v1.visit\_occurrence*

*group by visit\_concept\_id,visit\_source\_value*

*ORDER BY visit\_concept\_id ASC*



**Connection to the Table visit\_detail**

*SELECT A.person\_id, visit\_detail\_concept\_id, visit\_concept\_id,visit\_source\_value, visit\_source\_concept\_id*

*FROM omop\_5\_4\_v1.visit\_occurrence as A, omop\_5\_4\_v1.visit\_detail as B*

*where A.visit\_occurrence\_id = B.visit\_detail\_id;*

The tables visit\_occurrence and visit\_detail are linked via the keys visit\_occurrence and visit\_detail\_id.

## CASE 3: PERSON - Check the mapping for ethnicity. OMOP Table: **person**

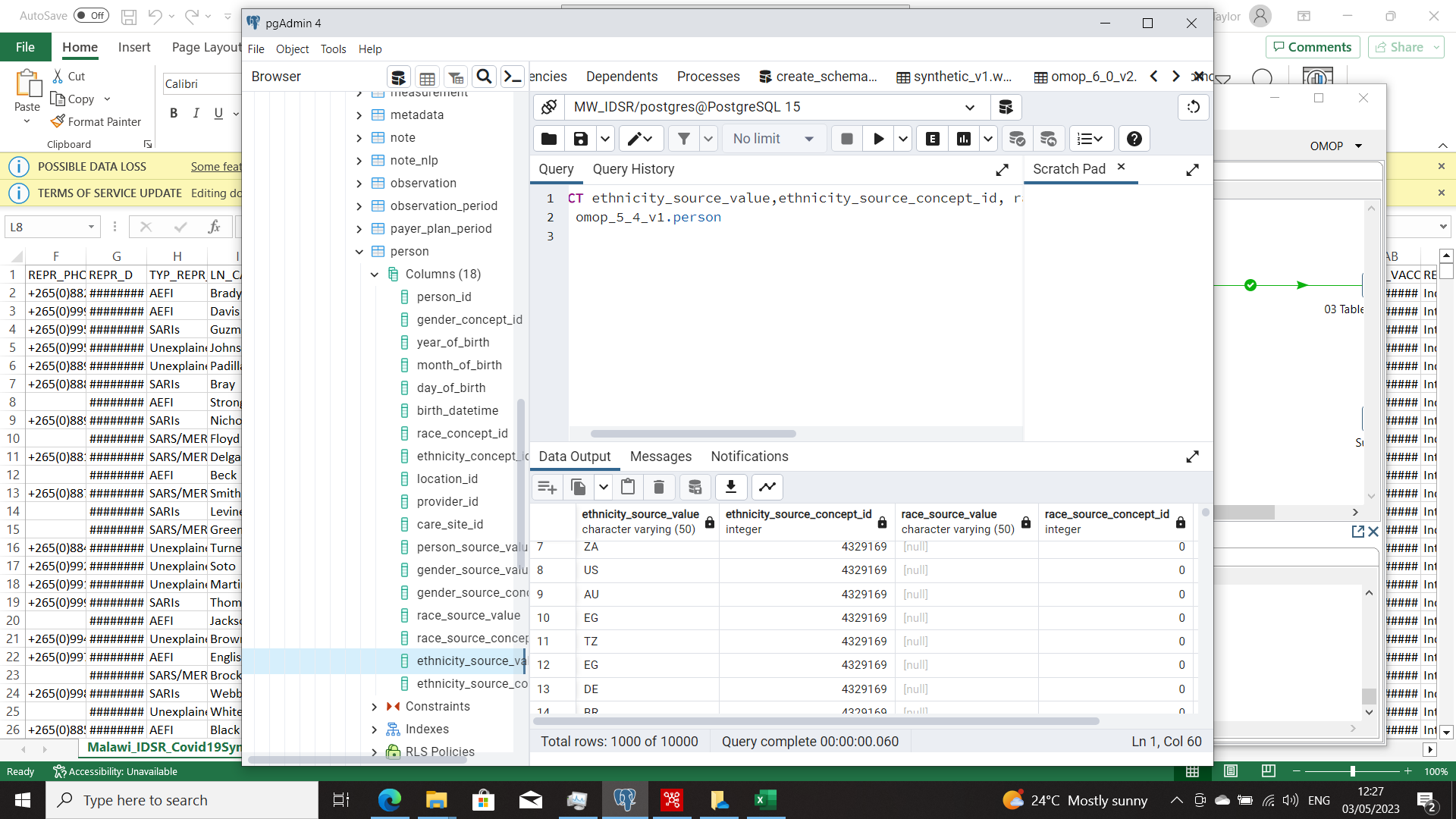
The Malawi IDSR data does not record ethnicity[[1]](#footnote-0), only nationality.

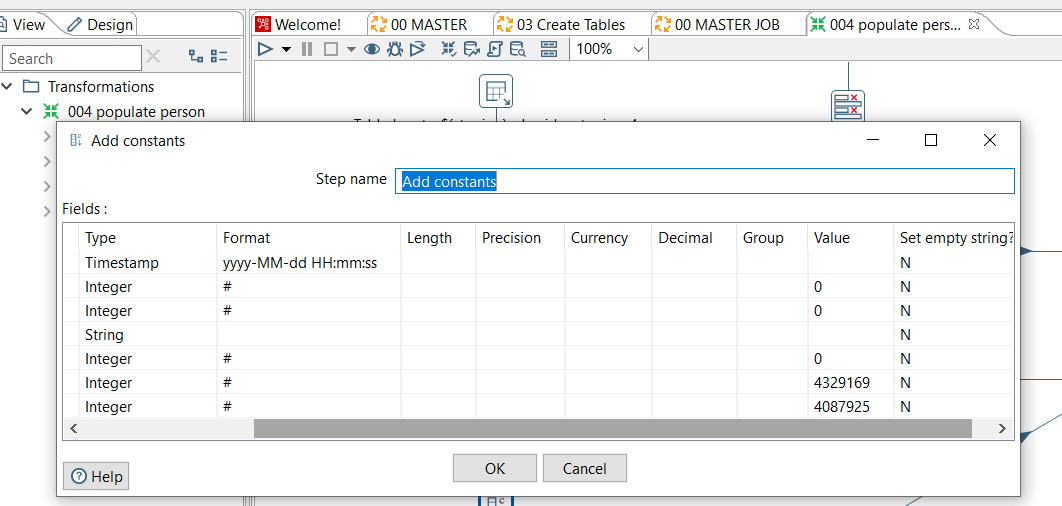
**The way the PENTAHO ETL treat this**

The fields ethinicity\_source\_value and ethnicity\_Source\_concept\_id have the same concept id. For ethinicity\_source\_value the concept is 4329169 (representing “country”). Ethnicity\_concept\_id is always equal to 4087925 (representing “ethnicity”). These two were defined in the original ETL from Tathagata to be constant/unchanging from the first record to the last, hence why they are the same for all records.

0 - default for null / missing data, e.g race

Ethnicity in the OMOP CDM follows the OMB Standards for Data on Race and Ethnicity: Only distinctions between Hispanics and Non-Hispanics[[2]](#footnote-1): “Hispanic or Latino” (concept\_id=38003563) and “Not Hispanic or Latino” (concept\_id=38003564). This means, the two categories are orthogonal to each other, and both Latinos and non-Latinos can have any racial or ethnic background. A list of concepts for Ethnicity are listed here: [ethnicity\_to\_concept.csv](https://drive.google.com/file/d/1rKAwqfnPFOZZuWHRY6MEAIwvUZJ5KU65/view?usp=share_link).





SELECT distinct ethnicity\_concept\_id FROM omop\_5\_4\_v1.person

SELECT ethnicity\_source\_value,race\_source\_value, race\_source\_concept\_id, race\_concept\_id FROM omop\_5\_4\_v1.person

where race\_source\_value IS NOT NULL;

## 

## 

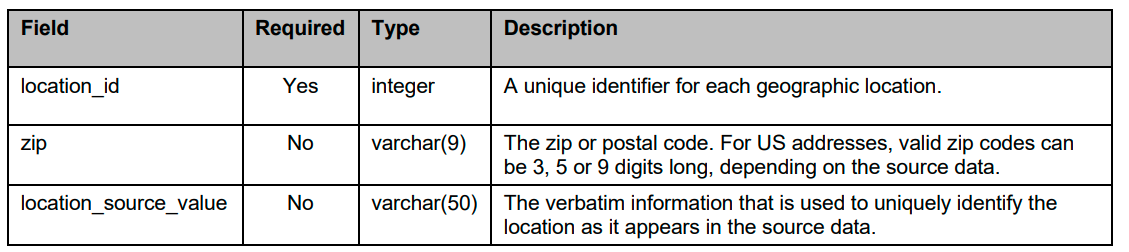
## CASE 4: LOCATION - Check the mapping for location. OMOP Table: **location**.

**TO INVESTIGATE/CHECK:**

1. Malawi IDSR data has the following fields for Location: NEAR\_LMK and PHYS\_ADDR. These have not been mapped as yet - either in the MW or WHO transformations.
2. LANDMARK which is NEAR\_LMK is mapped to patient\_residence which gets changes in the SQL in the Master when converting the tables to be similar to that of WHO. This patient residence then changes to location\_source\_value in OMOP in location table
3. The location table gets populated twice. First in Populate care site where address\_1 is the report\_site. And then in populate location where address\_1 is patient\_address
4. In populate care site, location source value is report\_district. In population location, the location source value is patient\_residence

**TO NOTE:**

The Location table represents a generic way to capture physical location or address information. Locations are used to define the addresses for Persons and Care Sites.



**CONVENTIONS**

* Each address or Location is unique and is present only once in the table.
* Locations do not contain names. In order to construct a full address that can be used on the Postal Service, the address information from the Location needs to be combined with information from the Care Site. The Person table does not contain name information.
* All fields in the Location tables contain the verbatim data in the Source. None of them are mandatory, but a valid Location record should at least contain either a Location Name or Location Zip.
* Zip codes are handled as strings of up to 9 characters length. For US addresses, these represent either a 3-digit abbreviated Zip code as provided by many Sources for Patient protection reasons, or the full 5-digit Zip code or the 9-digit (ZIP + 4) codes are recorded. Unless for specific reasons, analytical methods should expect and utilize only the first 3 digits. For international addresses, different rules apply.

## 

## CASE 5: PROVIDER- Check the mapping for the provider. OMOP Table: **provider**.

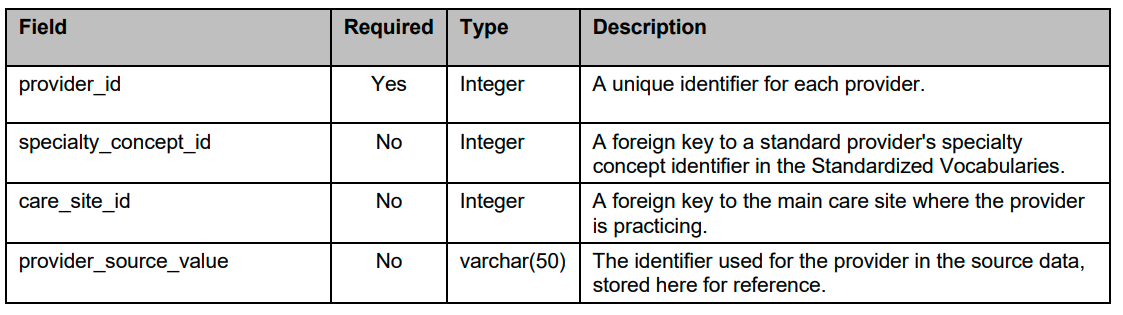
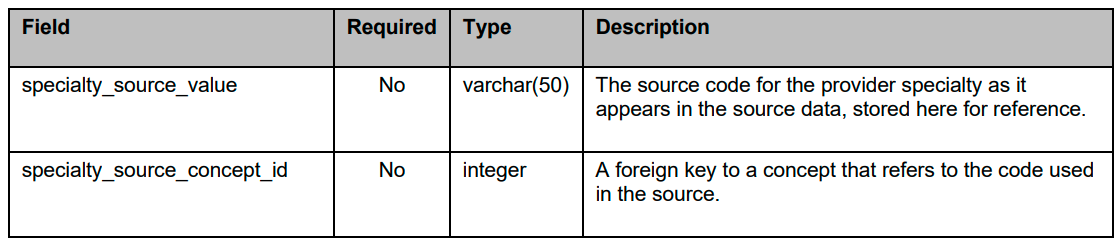
**TO INVESTIGATE/CHECK/CORRECT:**

* The MW IDSR data we have reporters and form completers (usually they are the same person) P\_COMPL\_FRM\_NM
* How are these looked at in the current transformations? P\_COMPL\_FRM\_NM which is reporters and form completers is in Malawi only and therefore was not dealt with
* Select count for Table: provider gives us 272 records for the MW IDSR dataset with 10,000 records which contains 578 distinct reporters. But there are 272 unique First names. This means that the mapping was actually done on the field REPR\_NM not using the **REPR\_ID**
* **Speciality\_Source\_Value:** contains no value but the corresponding speciality\_concept\_id is set to 45877976
* The link to the Table: care\_site needs also to be fixed.

**TO NOTE**

The Provider table contains a list of uniquely identified health care providers. These are typically

physicians, nurses, etc

**CONVENTIONS:**

* Providers are not duplicated in the table.
* Valid Specialty Concepts for both the test (measurement\_concept\_id) belong to the "Provider" domain. The Specialty Concepts are based on the CDC specialty classification.
* This table is used to represent the fixed relationship between Providers and Care Sites. Providers are also linked to Care Sites through Condition, Procedure and Visit records

## 

## **IN PROGRESS**

1. All cases above
2. Investigate the use of the OMOP SURVEY\_TABLE <https://medical-data-models.org/35349> And <https://github.com/OHDSI/CommonDataModel/issues/137>
3. Investigate the use of the Observation table instead of the Visit\_Occurrence for the MW IDSR data.

1. Interesting discussion on ethnicity: <https://forums.ohdsi.org/t/race-and-ethnicity-in-the-omop-cdm/8700/9> [↑](#footnote-ref-0)
2. <https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:ethnicity> [↑](#footnote-ref-1)