Two R scripts to evaluate the quality of the ETL and identify different types of errors.

A reusable method to assess the quality of the ETL process.

**Background:** Creating an extract, transform, and load (ETL) process to get from the source data to the OMOP Common Data Model (CDM) is usually a large task effort. The ETL process is often quite complex and may require changes over time. Hence, it is important to have a method that tests that the ETL does what it is supposed to do and continues to do so.

**Result:** Various types of errors are detected when applying this approach to different CDMs.

- **Missing data**
  - For certain sources, certain data had been mistakenly omitted from the ETL process and were not included in the CDM

- **ETL logic issues**
  - Some information in OMOP originated from wrong elements in the sources (e.g., wrong dates were associated to some records in the CDM)

- **Mapping errors**
  - Some information in the source data had been converted to OMOP using the wrong concepts (e.g., units of some measurements)

**Methods**

1. **Selection of patients**
   - Selection criteria:
     - Number of records in the CDM
     - Number of records in each OMOP domain
     - Length of observation period
     - Number of different sources from which the patient data comes
     - Number of observations in each of the sources

2. **Creation of the gold standard**

3. **ETL assessment**
   - Table-to-table comparison
   - Each row is labelled as True Positive (TP), False Positive (FP) or False Negative (FN)
   - For each table, precision, recall and F1-score are calculated

4. **ETL review**
   - Step 3 and 4 are iterated until an F1-score value of 1 is obtained for all domains.

**Conclusion:** This process has three main advantages:

1. Using a CDM with real data allows you to consider the variety and the complexity of clinical data during the quality control process.
2. Once the gold standard has been created, it can be reused to run the comparison script quickly and effortlessly and re-evaluate the ETL after the review.
3. While leaving maximum flexibility in the choice criteria for patient selection (step 1) and comparisons to be performed (step 3), the R scripts can be used with any database in the OMOP CDM. Furthermore, these scripts will be made available as open-source on GitHub.