ETL for a Hypothetical Clinical Trial with Power Query

Project Description

This project implements an ETL (Extract, Transform, Load) process for processing SAS files from a hypothetical clinical trial using Power Query in Power BI Desktop. The goal is to feed an automated payment system.

The system calculates payments for completed visits of patients participating in the trial. The ETL rules consider:

- Patient ID

- Patient status (Screened, Enrolled, Early Terminated, etc.)

- Visit name

- Visit date

- Whether the Screening visit was combined or not

- Patient age group (as payments vary by age group)

Power Query Objectives

- Confirm each patient's age group

- Confirm whether the Screening visit was combined

- Confirm the patient's final status

- Prepare the final visits table for payment calculation

Reports to Use (Fictitious for Repository)

- DM.csv – Patient demographics

- DOV.csv – Site visits

- DOV_PH.csv – Phone visits

- IE.csv – Screening failures

- EOS.csv – Patient final status

Note: The CSVs in this repository contain fictitious data with only the variables necessary to replicate the Power Query flow. They do not contain real patient information.

Usage Instructions

1. Open Power BI Desktop.

2. Go to **Get Data → Text/CSV** → load DM.csv. Select **Transform Data** (NOT Load).

3. In Power Query Editor, load DOV.csv the same way.

Clean DM table to have only one record per patient:

- Filter the AGE column to remove nulls (.) and keep only valid rows.

- Select the SUBNUM column.

- Go to **Home → Remove Rows → Remove duplicates**.

Clean DOV table (expand VISCOMB values):

- Right-click the DOV table, select **Duplicate,** and rename it VISCOMB_BY_SUBJECT.

- Keep only SUBNUM and VISCOMB columns.

- Filter VISCOMB to include only rows with values 1 or 2.

- Select SUBNUM → **Home → Group By**:

    o Group By: SUBNUM

    o New Column Name: VISCOMB_FINAL

    o Operation: Max

    o Column: VISCOMB

Merge DOV and VISCOMB_BY_SUBJECT:

- Select DOV → **Home → Combine → Merge Queries**

- Join SUBNUM from DOV with SUBNUM from VISCOMB_BY_SUBJECT

- Join Kind: Left Outer

- Expand VISCOMB_FINAL

Remove visits without dates:

- Filter VISDAT in DOV to keep only rows with dates.

Add subject statuses (Screened, Enrolled, Early Terminated):

- Duplicate DOV → rename DOV_FLAGS

- Keep only SUBNUM, STUDYEVENTTITLE, and VISDAT

- Ensure VISDAT has no nulls

- Add Custom Columns:

  - **SCREENED:** if [STUDYEVENTTITLE] = "Screening" then 1 else 0

  - **ENROLLED:** if [STUDYEVENTTITLE] = "Day 1" then 1 else 0

  - **EARLY_TERMINATED:** if [STUDYEVENTTITLE] = "Early Termination" then 1 else 0

- Group by SUBNUM to get max values for each status

Screen Failures (IE.csv):

- Load IE.csv

- Filter SVFYN = "Yes"

- Keep SUBNUM and IENODAT

- Remove duplicates on SUBNUM

- Rename IENODAT → SCREEN_FAIL_DATE

- Add column HAS_SCREEN_FAILED = if [SCREEN_FAIL_DATE] <> null then 1 else 0

Final Statuses (EOS.csv):

- Load EOS.csv

- Keep SUBNUM, EOSDECOD, and EOSDAT

- Add column SUBJECT_STATUS with logic:

if [EOSDECOD] = "Complete" then "Complete"

else if [EOSDECOD] = "Lost to follow up" then "Lost to Follow-up"

else if [EOSDECOD] = "Adverse event" then "Treatment Discontinued"

else if [EOSDECOD] = "Death" then "Treatment Discontinued"

else if [EOSDECOD] = "Physician decision" then "Treatment Discontinued"

else if [EOSDECOD] = "Protocol deviation" then "Treatment Discontinued"

else if [EOSDECOD] = "Study terminated by Sponsor" then "Treatment Discontinued"

else null

- Remove duplicates on SUBNUM

Create SUBJECT_STATUS table:

- Duplicate DM → rename SUBJECT_BASE

- Remove AGE

- Merge with DOV_FLAGS → expand SCREENED, ENROLLED, EARLY_TERMINATED

- Merge with IE.csv → expand HAS_SCREEN_FAILED

- Merge with EOS.csv → expand SUBJECT_STATUS

- Add column FINAL_SUBJECT_STATUS with logic:

if [EOS.SUBJECT_STATUS] = "Complete" then "Complete"

else if [EOS.SUBJECT_STATUS] = "Lost to Follow-up" then "Lost to follow up"

else if [EOS.SUBJECT_STATUS] = "Treatment Discontinued" then "Treatment Discontinued"

else if [DOV_FLAGS.EARLY_TERMINATED] = 1 then "Early Terminated"

else if [IE.HAS_SCREEN_FAILED] = 1 then "Screen Failure"

else if [DOV_FLAGS.ENROLLED] = 1 then "Enrolled"

else if [DOV_FLAGS.SCREENED] = 1 then "Screened"

else null

Add status to DOV table:

- Merge DOV with SUBJECT_BASE → expand FINAL_SUBJECT_STATUS

Derive age groups:

- Merge DOV with DM → expand DM.AGE → convert to Whole Number

- Add Custom Column AGE_GROUP:

if [DM.AGE] >= 18 and [DM.AGE] <= 64 then "18–64"

else if [DM.AGE] >= 65 and [DM.AGE] <= 74 then "65–74"

else if [DM.AGE] >= 75 then "75+"

else null

Add DOV_PH visits:

- Load DOV_PH.csv

- Keep SUBNUM, STUDYEVENTTITLE, VISDAT

- Merge VISCOMB_FINAL, FINAL_SUBJECT_STATUS, AGE and derive AGE_GROUP as above

- Duplicate DOV → rename DOV_final

- Ensure DOV_PH and DOV_final have the same columns → remove extras

- Append DOV_final + DOV_PH → export as CSV