# Project 3:
# Laboratory Data Cleaning and TLF Generation

Prepared By:
**Muzakkir Ansari**
B.Sc Statistics, 2024
Mumbai, India
muzakkiransari001@gmail.com

## 1. Project Overview

This project focuses on cleaning laboratory (LB) test data and generating essential Tables, Listings, and Figures (TLFs) in alignment with CDISC SDTM conventions. The dataset included raw laboratory results (e.g., patient ID, test names, values, units, reference ranges, and result indicators). The goal was to standardize these values, derive key variables, and prepare the dataset for reporting and analysis.

## 2. Datasets

- lab_raw.csv → Raw lab data including patient IDs, visit dates, test names, numeric results, measurement units, and low/high reference ranges.
- lbtest_mapping.csv → Mapping file created to standardize inconsistent spellings of lab test names (e.g., “Glucose Fstng”, “Glucose fasting” → “Glucose (Fasting)”).

## 3. Steps Performed

### Data Import

- PROC IMPORT was used to load the raw lab dataset. The mapping file was also imported for later use in harmonizing lab test names.

### Patient ID (USUBJID)

- Raw IDs had inconsistent formats (e.g., p-001, P001, 002).
- Cleaned by stripping spaces/symbols, ensuring uniform prefix "P", and zero-padding to 3 digits (e.g., P001).
- Final variable: USUBJID.

### Visit Date (LBDTC)

- Raw visit dates had multiple formats (e.g., 03/03/2023, 01MAR2023).
- Standardized using ANYDTDTE10. informat and formatted to DATE9..
- Final variable: LBDTC.

### Laboratory Test Name (LBTEST)

- Raw values contained different spellings/notations (e.g., Glucose fasting, GLUCOSE FSTNG).
- Created a mapping file (labtest_mapping.csv) to map all variations to standard names.
- A deduplication step was performed to ensure no duplicate records per patient, visit date, and laboratory test using PROC SORT NODUPRECS.
- Joined back with the raw data to assign a clean test name.
- Final variable: LBTEST.

### Lab Result (LBORRES, LBSTRESN)

- Raw results were numeric but sometimes treated as character.
- LBORRES: original result.
- LBSTRESN: standardized numeric value derived using INPUT after cleaning.

### Units (LBORRESU, LBSTRESU)

- Variations like mg/dl, MG/DL, mg/dL were standardized.

- In cases where the laboratory unit could not be mapped to a valid category, the value was standardized to "UNK" (Unknown).
- Final standardized unit variable: LBSTRESU (e.g., mg/dL, g/dL, U/L, /uL).

**Reference Ranges (LBSTNRLO, LBSTNRHI)**

- LBSTNRLO derived from cleaned low values.
- LBSTNRHI carried from raw high values.

**Range Indicator (LBSTNRIND)**

- Raw indicators were inconsistent (e.g., Low, NORMAL, blank).
- Standardized to LOW, HIGH, NORMAL, UNK.
- Additionally, recalculated programmatically using LBSTRESN vs. reference ranges to validate correctness.

**Derivations**

- Derived variables are crucial in clinical trials. For example, LBSTNRIND (Low, Normal, High, Unknown) provides quick interpretation of results. A validation step using PROC COMPARE ensured consistency between collected and derived values.

**Final Dataset**

- The cleaned dataset contains the following SDTM-aligned variables:
- USUBJID, LBDTC, LBTEST, LBORRES, LBSTRESN, LBORRESU, LBSTRESU, LBSTNRLO, LBSTNRHI, LBSTNRIND.
- The final dataset (labtest) was created in the SAS WORK library. It can optionally be exported to a permanent location (e.g., project3\out\labtest.sas7bdat or .csv) for further use.

**TLFs**

- Listing – Detailed listing of all cleaned lab results by subject, test, and ranges.
- Summary Table – Frequency of lab results (LOW, NORMAL, HIGH, UNK) by test.
- Figure – Bar chart showing distribution of results by test and range category.

## 4. Outputs

- labtest.sas7bdat → final cleaned dataset.
- Frequency tables and bar charts generated within SAS Results Viewer.

## 5. Tools Used

- SAS 9.4 (Base SAS, PROC SQL, PROC FREQ, PROC SGPLOT).

## 6. Conclusion

This project successfully demonstrated the cleaning of raw laboratory data and the creation of SDTM-aligned variables. By mapping, standardizing, and validating critical fields such as test names, results, units, and reference ranges, a high-quality analysis dataset was prepared. The generated TLFs provide clear insights into laboratory abnormalities and data trends, which are essential for clinical reporting.