Project 2
Adverse Events (AE) Data Cleaning & Derivation

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

1. Project Overview

- This project demonstrates end-to-end cleaning and derivation of an Adverse Events (AE) dataset in a clinical trial setting, using Base SAS.
- The workflow aligns with industry practices: data traceability, controlled terminology mapping, derived variables, and SDTM-aligned variable names.

2. Datasets

- adverse_ae.csv → Raw adverse event dataset.
- ae_final.csv → Cleaned and standardized dataset.
- ae_final_deriv.csv → Final dataset with derived variables.

3. Steps Performed

Data Import

- PROC IMPORT used to load raw CSV into SAS.
- Cleaning & Standardization
- Patient IDs standardized to format P### → USUBJID.
- Adverse Event Terms (ae_term) corrected for typos (e.g., “Vomting” → “Vomiting”), proper case applied → AEDECOD.
- Dates (ae_start, ae_end) standardized with ANYDTDTE., DDMMYY10., MMDDYY10., and missing codes handled (NA/N/A/blank) → AESDTC, AEENDTC.
- Seriousness mapped to YES/NO/UNK → AESER.
- Outcome grouped into standard categories (RECOVERED, ONGOING, NOT RECOVERED, FATAL) → AEOUT.
- Action Taken harmonized (NONE, DOSE REDUCED, DRUG WITHDRAWN, UNKNOWN) → AEACN.
- Relation to Study Drug mapped to (RELATED, NOT RELATED, POSSIBLE, UNLIKELY, UNKNOWN) → AEREL.

Variable Mapping

- patient_id → Cleaned to clean_id → Final variable USUBJID
- ae_term → Cleaned to clean_ae_term → Final variable AEDECOD
- ae_start → Cleaned to clean_start → Final variable AESDTC
- ae_end → Cleaned to clean_end → Final variable AEENDTC
- seriousness → Cleaned to clean_serious → Final variable AESER
- outcome → Cleaned to outcome_clean → Final variable AEOUT
- action_taken → Cleaned to action_clean → Final variable AEACN
- relation → Cleaned to relation_clean → Final variable AEREL

Renaming for SDTM Alignment

- Variables renamed according to CDISC SDTM AE domain conventions (e.g., clean_id → USUBJID, clean_ae_term → AEDECOD).

Derivations

- Duration = Number of days between AE start and end (inclusive).
- This avoids confusion if dates are missing..
- AESERFL = 1 if serious (AESER=YES), else 0.
- AEONGOFL = 1 if AEENDTC missing (event ongoing), else 0.

4. Outputs

- ae_final.csv → Cleaned dataset (SDTM-aligned).
- ae_final_deriv.csv → Dataset with derived variables.
- SAS Outputs: Frequency checks for seriousness, relation, action taken, and duration distributions.

Example of Cleaning Impact

- Raw → Cleaned
- p-002 → P002
- DIZZINES → Dizziness
- 10/03/23 → 10MAR2023
- n → NO

5. Tools Used

- SAS 9.4 (DATA step, PROC IMPORT, PROC EXPORT, PROC FREQ, PROC PRINT, PROC CONTENTS).

Conclusion

The Adverse Events (AE) dataset was successfully cleaned and standardized, with all variables mapped to CDISC-SDTM AE domain structure. Derived variables such as Duration, AESERFL, and AEONGOFL were created, ensuring the dataset is both traceable and analysis-ready for clinical reporting.