# End-to-End SDTM Mapping for an Oncology Clinical Trial (Phase I)

**Study Overview**

* **Study Title**: Phase I, Open-Label, Multi-Center Oncology Trial
* **Therapeutic Area**: Oncology
* **Population**: Patients with solid tumors
* **Design**: Non-randomized, early-phase safety and tumor response study
* **Objective**: Evaluate safety and tumor response to an investigational product in solid tumor subjects

# Table of contents

1. Study Overview  
2. Objectives  
3. Tools and Standards  
4. Folder Structure  
5. Input Datasets  
6. Mapping Specifications  
7. SDTM Domains Created  
8. Variable Derivations  
9. Key Assumptions  
10. Output Overview  
11. Challenges and Resolutions  
12. Summary and Learnings  
13. Author Info

## 1. Study Overview

Therapeutic Area: Oncology  
Study Phase: Phase I  
Study Type: Open-label, multi-center, non-randomized  
**🏥 Multi-center**

**Definition**:  
An open-label study is a type of clinical trial where **both the researchers and the participants know** which treatment is being given.

**Why it matters**:

* No blinding is used.
* Common in early-phase studies (e.g., Phase I).
* Useful when blinding is not ethical or practical.

Population: Patients with solid tumors  
Objective: Evaluate safety and tumor response to an investigational product.

**🔁 Non-randomized**

**Definition**:  
Participants are **not randomly assigned** to treatment groups. Assignment is done by study design or clinician judgment.

**Why it matters**:

* Less control over bias compared to randomized trials.
* Often used in **Phase I** or early exploratory studies.
* Suitable when all patients receive the same investigational treatment.

## 2. Objectives

* Transform raw clinical trial data into **CDISC SDTM-compliant** datasets using Base SAS.
* Implement **oncology-specific tumor domains** including TU (Tumor Identification), TR (Tumor Results), and RS (Response Assessment).
* Apply and validate **controlled terminology and CDISC compliance rules** throughout the mapping process.
* Generate and organize **submission-ready .sas7bdat datasets** suitable for regulatory review and downstream analysis.

## 3. Tools and Standards

- SAS 9.4: Used for data cleaning, mapping, and derivation.  
- Microsoft Excel: Used for mock data and mapping specifications.  
- CDISC SDTMIG v3.2: Followed for tumor domain compliance.

## 4. Folder Structure

- Raw\_Datasets/ – Contains raw input files (Excel/CSV)  
- Mapping\_Specifications/ – Mapping files per domain  
- SAS\_Code/ – SAS scripts for data transformation  
- Output\_SDTM/ – Final SDTM datasets in .sas7bdat format  
- Documentation/ – Project summary, assumptions, and derivation logs

## 5. Input Datasets

* **DM.xlsx** – Demographics (e.g., SUBJECT, AGE, SEX, RACE)
* **EX.xlsx** – Exposure data (e.g., EXSTDAT, EXENDAT, dosage details)
* **LB.xlsx** – Laboratory test results (e.g., test codes, values, dates)
* **TU.xlsx** – Tumor Identification (e.g., tumor location, ID, sequence)
* **TR.xlsx** – Tumor Measurements (e.g., diameter, date, location)
* **RS.xlsx** – Tumor Response (e.g., response category, result date)

## 6. Mapping Specifications

Each SDTM domain was developed based on a dedicated **mapping specification sheet** outlining source-to-target variable mappings. Variables such as USUBJID, TUSEQ, TRSEQ, and RSSEQ were derived in accordance with **CDISC SDTM Implementation Guide (SDTMIG v3.2)** standards.

Oncology-specific **linkage variables** like TULINKID, TRLINKID, and RSLINKID were generated to maintain traceability between tumor location, measurement, and response across TU, TR, and RS domains.

## 7. SDTM Domains Created

* **DM.xlsx** – Demographics (e.g., SUBJECT, AGE, SEX, RACE)
* **EX.xlsx** – Exposure data (e.g., EXSTDAT, EXENDAT, dosage details)
* **LB.xlsx** – Laboratory test results (e.g., test codes, values, dates)
* **TU.xlsx** – Tumor Identification (e.g., tumor location, ID, sequence)
* **TR.xlsx** – Tumor Measurements (e.g., diameter, date, location)
* **RS.xlsx** – Tumor Response (e.g., response category, result date)

## 8. Variable Derivations

* **USUBJID**: Constructed by concatenating STUDYID and SUBJECT (e.g., USUBJID = STUDYID || "-" || SUBJECT) to uniquely identify each subject across domains.
* **TUSEQ, TRSEQ, RSSEQ**: Derived using SAS BY USUBJID logic along with RETAIN to assign sequence numbers within each subject for TU, TR, and RS records respectively.
* **RFXSTDTC**: Represents the first study treatment date/time, derived from SPCPKB1 (IPFD1DAT + IPFD1TIM) or from EXSTDAT if dosing details are captured in EX.
* **RFENDTC**: Represents the end of study reference period; derived from EXENDAT when available, otherwise falls back to EXSTDAT.

## 9. Key Assumptions

* All subjects included in the TU domain have **at least one valid tumor location** available in the source data.
* Tumor measurements in the TR domain are assumed to be **recorded consistently per subject, visit, and location**, enabling accurate derivation of TRSEQ.
* **Controlled terminology** for variables such as SEX, RACE, and ETHNIC was applied manually within SAS code using conditional logic (e.g., IF-ELSE) based on CDISC codelists.

## 10. Output Overview

* Final **SDTM-compliant datasets** generated in .sas7bdat format include:
* DM.sas7bdat – Demographics
* TU.sas7bdat – Tumor Identification
* TR.sas7bdat – Tumor Measurements
* RS.sas7bdat – Tumor Response
* Each dataset was validated using SAS procedures:
* **PROC SORT** – Ensured proper ordering and removal of duplicates.
* **PROC COMPARE** – Compared intermediate and final datasets to verify accuracy of transformations.

## 11. Challenges and Resolutions

* **Missing or incomplete date values** (e.g., RFENDTC) were handled by applying **fallback logic** using alternate variables such as EXSTDAT when EXENDAT was unavailable.
* **Controlled terminology mappings** for variables like SEX, RACE, and ETHNIC were implemented using conditional IF-ELSE logic in SAS, based on CDISC codelists.
* **Sequencing variables** (TUSEQ, TRSEQ, RSSEQ) required careful handling; issues related to incorrect or duplicate sequencing were resolved using RETAIN, BY USUBJID, and FIRST. logic in data step processing.

## 12. Summary and Learnings

* Gained a comprehensive understanding of the **end-to-end SDTM mapping workflow** using realistic mock clinical data.
* Successfully applied derivation logic for **complex oncology-specific variables** such as TUSEQ, TRSEQ, and RSSEQ.
* Enhanced **SAS programming proficiency** through the use of macros, conditional logic, data step processing, and dataset merging techniques.
* Developed confidence in implementing **CDISC-compliant domains** and preparing submission-ready datasets.

## 13. Author Info

**Author**: Prasanthi Kata  
**Role**: Clinical SAS Programmer (Intern)  
**Profile**: [GitHub – Prasanthi-sas](https://github.com/Prasanthi-sas)