Technology Changes with Time, Have you?

Python for Implementing & Automating **CDISC** Compliant Data

Abstract

This project demonstrates the design and implementation of an automated pipeline to map heterogeneous clinical trial source datasets into the CDISC-compliant Study Data Tabulation Model (SDTM) DM (Demographics) domain. Using a specification-driven approach, raw data from multiple files (DM_IN, Disposition, Informed Consent, Randomization, Trial Arms) are ingested, harmonized, and transformed according to predefined rules for variable naming, derivations, and controlled terminology. Key processes include automated generation of USUBJID keys, standardized ISO 8601 date conversion, demographic and treatment-arm assignment, and metadata application using SDTM-compliant labels, lengths, and formats. The resulting DM dataset is exportable in XPT, CSV, Excel, and SQL formats, ready for regulatory submission or integration with other domains. In practice, this automated framework significantly reduces manual programming effort and QC time compared to traditional study-by-study SDTM programming, enabling consistent, auditable, and rapid DM domain preparation.

Background

Clinical trial data from multiple systems must be standardized to CDISC's Study Data Tabulation Model (SDTM). This project implemented a repeatable, automated pipeline that ingests heterogeneous raw datasets and produces a validated, submission-ready SDTM DM (Demographics) domain.

Objective

- Ingest multiple source files (DM_IN, Disposition, Informed Consent, Randomization, Trial Arms).
- Map raw variables to SDTM-compliant attributes (STUDYID, USUBJID, SUBJID, RFSTDTC, ARMCD, etc.).
- Apply transformation rules from a central specification.
- Output standardized datasets in CSV, Excel, and SQL for downstream use.

1. Execution Approach

Step	Action
1. Create USUBJID	Derived in each dataset: `'ABC-400'
2. Sort Datasets	All datasets sorted by **USUBJID** for consistent merging.
3. Merge Core Datasets	**DM_IN**, **DS**, **INCO**, and **RAND** merged on **USUBJID**.
4. Date Handling	Converted **FIRSTDT** to numeric then formatted to ISO 8601 (YYYY-MM-DD) to derive **RFSTDTC**.
5. Arm Assignment Merge	In **RD**, renamed R_ARM to ARMCD. Sorted **RD** and **TA** by ARMCD, merged to pull ARMCD and ARM. Then merged arm data into DM by **USUBJID**.
6. Variable Retention & Labeling	Rearrange variables as per DM domain specification
7. Race Recoding	Map numeric race codes to SDTM controlled terms and rename back to **RACE**
8. Output Dataset	Created final DM with all required and expected variables with specified attributes.

2. Transformations Applied

Transformation	Description
Dates	All date variables converted to **ISO 8601 (YYYY-MM-DD).**
Age	Derived from **BRTHDTC** and **RFSTDTC**; **AGEU** set to **"YEARS"**
Sex & Race	Mapped numeric codes to controlled SDTM text values.
Treatment Arms	Extract **ARMCD** and **ARM** from **RD + TA** merge; mirrored into **ACTARMCD / ACTARM**
Country & Defaults	**COUNTRY** set to **"USA"**; **DTHDTC** and **DTHFL** left null; **ETHNIC** set to **"NOT REPORTED"**

3. Challenges Faced

Challenge	Resolution
Different source file formats / column names	Wrote generic loader and header-cleaning functions.
Non-standard date formats	Built **parse_date_safe**() function with robust **ISO 8601** conversion.
Missing or inconsistent race/sex codes	Created explicit mapping function to controlled SDTM terms.
Arm assignment split across RD and TA	Automated **RD+TA** merge by **ARMCD** before merging with DM.
Ordering and metadata compliance	Used retain-order lists and applied labels/lengths via Python metadata mapping

Technology Changes with Time, Have you?

Python for Implementing & Automating **CDISC** Compliant Data

Abstract

This project demonstrates the design and implementation of an automated pipeline to map heterogeneous clinical trial source datasets into the CDISC-compliant Study Data Tabulation Model (SDTM) DM (Demographics) domain. Using a specification-driven approach, raw data from multiple files (DM_IN, Disposition, Informed Consent, Randomization, Trial Arms) are ingested, harmonized, and transformed according to predefined rules for variable naming, derivations, and controlled terminology. Key processes include automated generation of USUBJID keys, standardized ISO 8601 date conversion, demographic and treatment-arm assignment, and metadata application using SDTM-compliant labels, lengths, and formats. The resulting DM dataset is exportable in XPT, CSV, Excel, and SQL formats, ready for regulatory submission or integration with other domains. In practice, this automated framework significantly reduces manual programming effort and QC time compared to traditional study-by-study SDTM programming, enabling consistent, auditable, and rapid DM domain preparation.

4. Benefits

- **Consistency** – One standardized derivation pipeline for DM domain.
- **Traceability** – Full audit trail from source variable to SDTM variable.
- **Efficiency** – Automation of code lists and date conversions reduces QC time.
- **Regulatory-Ready** – Output immediately ready for submission or integration with other SDTM domains.
- **Flexibility** – Adding new variables or sources requires only spec updates, not code rewrites.

5. Time Strategy / Time Saved

Aspect	Manual Approach	Automated Approach
Variable mapping & derivations	Several days per study to write SAS code manually	Central JSON/spec + Python/SAS engine executes in minutes
QC of date formats & controlled terms	Manual review across datasets	Built-in conversion & recoding functions ensure uniformity automatically
Combining arm data & demographics	Manual joins with multiple intermediate datasets	Automated RD+TA merge and single pass integration

6. Deliverables

Deliverable	Description
Python Scripts	**Implement data ingestion, transformation, mapping, and export**
SDTM DM Dataset	**Standardized DM dataset exportable to CSV, Excel, SQL**
Process Documentation	**Structured, auditable workflow reusable across studies**

Background and Overall Results

The SDTM Project 1 successfully demonstrates the automation of mapping heterogeneous clinical trial datasets into a CDISC-compliant SDTM DM (Demographics) domain. Using a specification-driven pipeline, raw data from multiple sources were ingested, harmonized, and transformed with standardized variable naming, controlled terminology, and ISO 8601 date formats. Automated processes, including USUBJID generation, arm assignment merges, and demographic derivations, ensured consistency, traceability, and regulatory compliance across the resulting DM datasets.

Compared to traditional manual approaches, the automated framework significantly reduced programming and quality control time, improved data accuracy, and provided reusable, flexible workflows that can easily accommodate new variables or data sources. The final deliverables, exportable in CSV, Excel, SQL, and XPT formats, are immediately submission-ready and integrable with other SDTM domains, demonstrating both practical efficiency and adherence to industry standards