Automation of SDTM data derivation with R

Ben Barnaby-Pass (ben.pass@phastar.com)

1. Abstract

Converting raw data to SDTM format is a crucial stage in any clinical trial. Currently, SAS is the predominant language employed for this process, requiring considerable human intervention. Although automation has already been used in PHASTAR to generate SAS code, we have devised an alternative approach that generates automated R code, which substantially reduces human involvement in routine coding tasks.

Our tool utilizes curated metadata, containing vital information essential for executing the RAW to SDTM derivation process. Subsequently, the tool generates a set of automated functions, facilitating the creation of SDTM datasets with minimal post-processing requirements. This approach not only streamlines a significant portion of coding tasks but also establishes a standardized data derivation process across various trials.

2. The Motivation

We are amid creating a **SDTM automation tool** which will let the user create **SDTMs from RAW datasets with ease in R**. A large part of SDTM data derivation is automated using R. This package provides an open-source **alternate to the mapit-SAS tool available within PHASTAR**.

3. The Value

This tool is designed **to automate manual statistical programming tasks** and could expand to other data analytics functions, **saving programmers' time** and **reducing delivery timelines while maintaining quality & consistency of codes.** This tool sets the foundation for **future AI-driven data analysis**, **using metadata to guide Large Language Models** in creating complex code for insightful data analysis while **preserving source data confidentiality.**

4. The Process

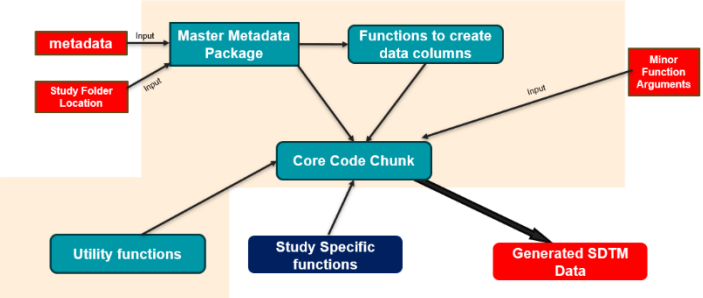
The code automation process is involved. First, we specify a metadata with variable mapping, this is helped with the use of the **CDISC 360 standards**.

A	B	G	H	I	J	K	L
RAW_TAB	RAW_COLUM	RAW_CO	RAW_WHR	SDTM_TAB	SDTM_COLUN	SDTM_WHRCLAU	SDTM_ASSIGNED_VALUE
AE	AEACN	0		AE	AEACN		DOSE NOT CHANGED
AE	AEACN	3		AE	AEACN		DRUG INTERRUPTED
AE	AEACN	4		AE	AEACN		DRUG WITHDRAWN
AE	AEACN	98		AE	AEACN		NOT APPLICABLE
AE	AECONTRT	C49487		AE	AECONTRT	AEYN == 'C49488'	N
AE	AECONTRT	C49488		AE	AECONTRT	AEYN == 'C49488'	Y

Based on this metadata, **`mapitR` generates R functions to create each column** of pre-mentioned SDTM domains. Finally, all such functions are automatically called that *almost* gives us the final data domains subject to elementary post-processing.

5. Workflow

Here's an example of the workflow that we expect from the R SDTM work:



With correct metadata, as the process is executed, it generates functions like:

```
add_AEACN_AE <- function(x) {
  x |>
    mutate(S__AEACN =
             case_when((AEACN == '0' ) ~ 'DOSE NOT CHANGED',
                       (AEACN == '3' ) ~ 'DRUG INTERRUPTED',
                       (AEACN == '4' ) ~ 'DRUG WITHDRAWN',
                       (AEACN == '98' ) ~ 'NOT APPLICABLE'))
}

add_AECONTRT_AE <- function(x) {
  x |>
    mutate(S__AECONTRT =
             case_when((AECONTRT == 'C49487'  & AEYN == 'C49488') ~ 'N',
                       (AECONTRT == 'C49488'  & AEYN == 'C49488') ~ 'Y'))
}
```

Once functions are created, relevant datasets are **programmatically read from RAW domain** & these functions are combined in an automated way. **However, a user doesn't need to know the intricacy of the process** – they need to run a couple of R functions.

6. Current Status & Future

`mapitR` is deployed in PHASTAR & **used in studies on trial basis**. With vignettes, a package website & internal training in place, onboarding with the package is seamless**. The package has a dedicated maintenance team** and sees continuous improvement as new needs emerge.