A Harmonized, Report-Friendly SDTM and ADaM Data Flow

Aileen L. Yam, sanofi-aventis, Bridgewater, NJ
Marie-Rose Peltier, sanofi-aventis, Croix-de-Berny, France

ABSTRACT

This paper shows an in-house implementation strategy that integrates the requirements of the SDTM SDS model and the ADaM model with the needs for generating tables, listings and graphs to create a harmonized, report-friendly data flow.

In designing an optimal data flow, the creative task is to understand all the requirements and viewpoints and to arrive at decisions that are a balanced, coherent and consistent representation of the CDISC requirements and the users' world. In addition to the fulfillment of functions and logic, the visual appearance of the flow is important: the design has to be elegantly clear and lean to convey the message. This paper also summarizes our experience in structuring the design of the data flow.

I. REDUNDANCY AND "WHICH COMES FIRST" ARE MATTERS OF PERSPECTIVES

Here is a couple of examples of user viewpoints:

1. From an SDTM perspective, some variables in the ADaM domains are redundant. From an ADaM perspective, some variables in the SDTM domains are redundant.

2. There is also a chicken-and-egg question on where derived variables should be created first – the SDTM domains or the ADaM domains.

Each of these perspectives brings along its own context of understanding about what matters, what can be designed, and what methods or processes are appropriate.

Beyond these perspectives is a fundamental question: What approach and process would be compliant with regulatory standards, and at the same time, fit our business requirements the most? To answer this question, the goals and focuses of SDTM SDS and ADaM domains are analyzed in relation to how these domains are used for report-writing and analyses, and how the domains and reports are validated.

II. GOALS AND FOCUS OF SDTM SDS

For the SDTM SDS domains:

- The variables are mostly CRF variables. For report-writing and graphs, more derived variables are needed.

- The variables are character variables. Numeric values are needed for sorting and analysis.

- Date variables are ISO dates. Numeric date values are needed for the calculation of relative days, duration, etc.

- Population flags, drug variables and other common variables are not readily available for use in analysis.

- Some data structures are designed to facilitate standardization, and are not one PROC away.

III. GOALS AND FOCUS OF ADAM

For the ADaM domains:

- CRF variables are further processed. It can be difficult to trace the source of discrepancies or to validate derived variables.

- Variable attributes are inconsistent across domains.

- The lack of standardization makes it difficult to develop macros or programs for variables that are common to or similar in root construction for both SDTM and ADaM.

IV. CRITERIA OF AN OPTIMAL DATA FLOW

The design of a data flow is a heuristic process. No tool is right for everything. Our objectives are generality, flexibility, maintainability, efficiency and simplicity. It is natural to leave some issues unresolved until there is more information for fine-tuning.

After some trial and error and much discussion, our in-house process has evolved into a data flow with a combination of the essential elements in SDTM SDS, ADaM and report-writing. The combined domains maximize the merits of SDTM and ADaM to address the requirements in tables, listings and graphs as well as to facilitate validation.

The combination is intended to make each domain complete and can be used for:

1. producing tables, listings and graphs,

2. creating SDTM SDS domains and adding SUPPQUAL and RELREC at submission,

3. creating additional domains which are dependent on the variables from the combined domains. The additional domains are also used for producing tables, listings and graphs.

V. DETAILS OF THE DATA FLOW

First, raw data from data entry and data management are downloaded into SAS® data sets. From the SAS raw data, variables are mapped where applicable and derived, and report-writing variables are created.

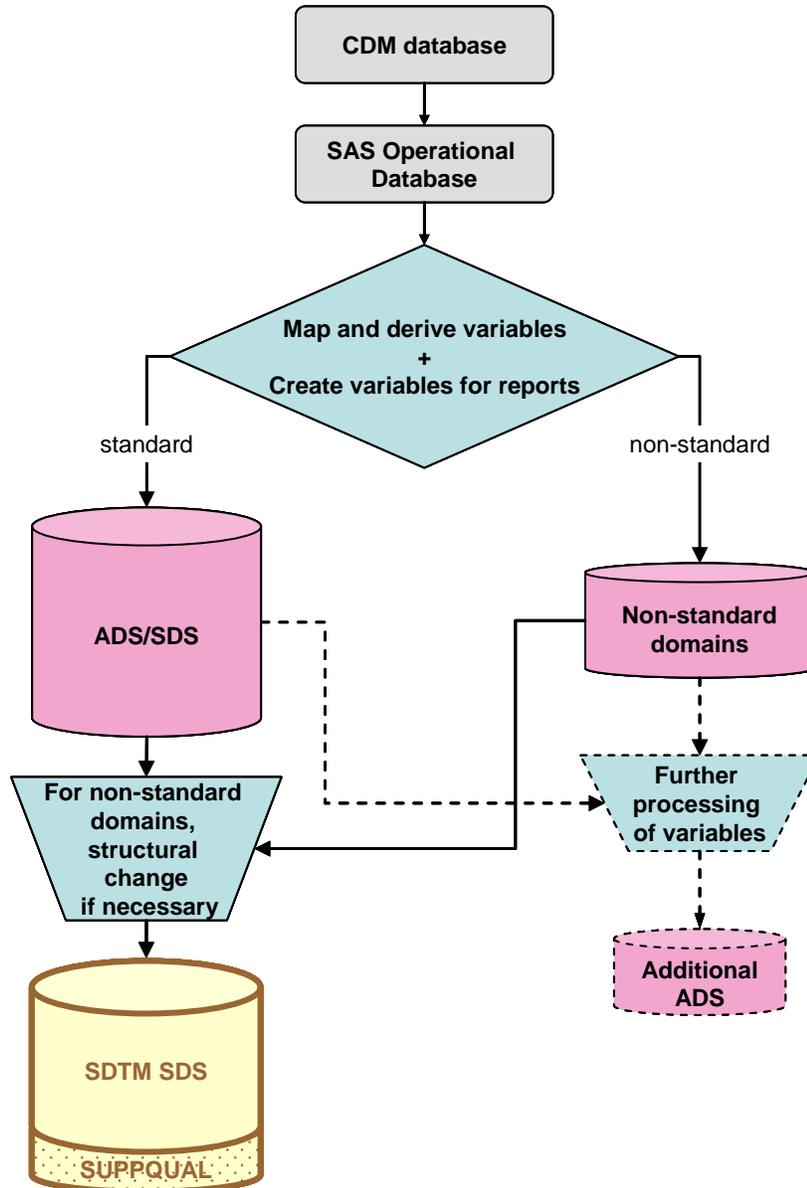In terms of data structures, there are two paths: standard and non-standard.

Data that can fit into the SDTM SDS structures and can be report-friendly at the same time will go through the standard path. Mapping to SDTM SDS structures is performed and additional variables for analysis and report-writing are created into what we call "ADS/SDS domains". Most of the data, especially safety data, can go through the standard path.

Data that cannot fit into the SDTM SDS structures and be report-friendly at the same time or be amenable for creating additional analysis data will go through the non-standard path. Data that go through non-standard path usually keep their original structures from raw data. Like the standard domains, the "Non-standard domains" contain raw data variables, derived variables and report-writing variables.

Both the "ADS/SDS domains" and the "Non-standard domains" represent a fusion of SDTM, ADaM and report variables. They are analysis-ready, report-ready, SDTM-ready and ADaM-ready data. They are submitted, along with the SDTM SDS domains. Data in "Non-standard domains" are sometimes in horizontal structures and need to be transposed to vertical structures to create the SDTM SDS domains.

Occasionally, second-level analysis data are needed. The second-level analysis data contain variables that are further derived from the "ADS/SDS domains" or the "Non-standard domains". Example of the second-level analysis data that we call "Additional ADS" are: *exposure summary* domain created from *exposure* domain, *collapsed AE* domain from *AE* domain, *time to analysis* from *efficacy* data. If the second-level analysis data are created, they are also submitted.

The process flow is presented visually below. CDM stands for Clinical Data Management raw data.

```
                        ┌─────────────────┐
                        │  CDM database   │
                        └────────┬────────┘
                                 │
                        ┌────────▼────────┐
                        │ SAS Operational │
                        │    Database     │
                        └────────┬────────┘
                                 │
                         ◇───────▼────────◇
                        ╱  Map and derive  ╲
                       ╱     variables      ╲
                       ╲         +          ╱
                        ╲ Create variables ╱
                         ◇  for reports  ◇
               standard   │            │  non-standard
                          ▼            ▼
                    ┌──────────┐   ┌──────────────┐
                    │ ADS/SDS  │   │ Non-standard │
                    │          │   │   domains    │
                    └────┬─────┘   └──────┬───────┘
                         │                │
                ┌────────▼────────┐  ┌────▼────────┐
                │ For non-standard│  │   Further   │
                │    domains,     │  │ processing  │
                │   structural    │  │ of variables│
                │    change       │  └────┬────────┘
                │  if necessary   │       │
                └────────┬────────┘  ┌────▼────────┐
                         │           │ Additional  │
                    ┌────▼─────┐     │    ADS      │
                    │ SDTM SDS │     └─────────────┘
                    │          │
                    │ SUPPQUAL │
                    └──────────┘
```

Note: The color pink stands for analysis-ready/report-ready data and are used for reporting.
       Both the pink-colored data and the SDTM SDS/SUPPQUAL data are submitted.
       The dotted line means when needed.

VI. RATIONALE

The reasons that led to the process described above are:

- It is more cost-efficient to have one set of data specifications that combines the essence of SDTM, ADaM and report-writing than two separate sets of data specifications (one for SDTM SDS and one for ADaM) that have redundant variables.

- It is easier to create, maintain and update one set of data specifications and domains.

- There is no need to program, validate and crosscheck two sets of domains on the same topic.

- The two models are integrated to complement each other to provide variables for analysis and reports.

- The integrated data structures tend to fit reporting objectives better.

- There is no need to break variables in domains apart into SUPPQUAL with RELREC for SDTM and to put them back again for report writing.

- The process encourages the standardization of ADaM domain variables. The same set of rules in nomenclature and variable attributes for both SDTM and ADaM allows for the use of some common macros and programs.

- The validation of derived variables is facilitated, since all the variables are in the same data set and the same structure.

VII. CONCLUSION

Although the data flow design presented here may serve as a how-to article on process and rules that can be conveniently applied, we are constantly exploring on how lessons learned from all aspects of the data flow design can be applied for overall improvement.

For additional information, please contact:

Aileen L. Yam
sanofi-aventis
200 Crossing Blvd.
Bridgewater, NJ 08807
work phone: 908-231-3492
email: aileen.yam@sanofi-aventis.com