PhUSE 2017

Paper DH11

SDTM-ADaM Hybrid: Supermodel or Frankenstein?

Ken Stoltzfus, Accenture, Berwyn, USA

ABSTRACT

The CDISC standards SDTM and ADaM have been adopted by a majority of sponsors, particularly since the formal enactment of the FDA 2014 binding guidance. Considering this, the industry trend now is to focus on efficiencies in the operational processing of SDTM & ADaM data. In that spirit, this paper intends to explore the proposition for utilizing pieces of two distinct CDISC models as a single operational structure and the potential benefits and challenges that it might produce. This will consist of a background description of operational hardships, followed by an exploration into some aspects surrounding such a hybrid model, and concluded by benefits and challenges looking forward.

BACKGROUND REASONING

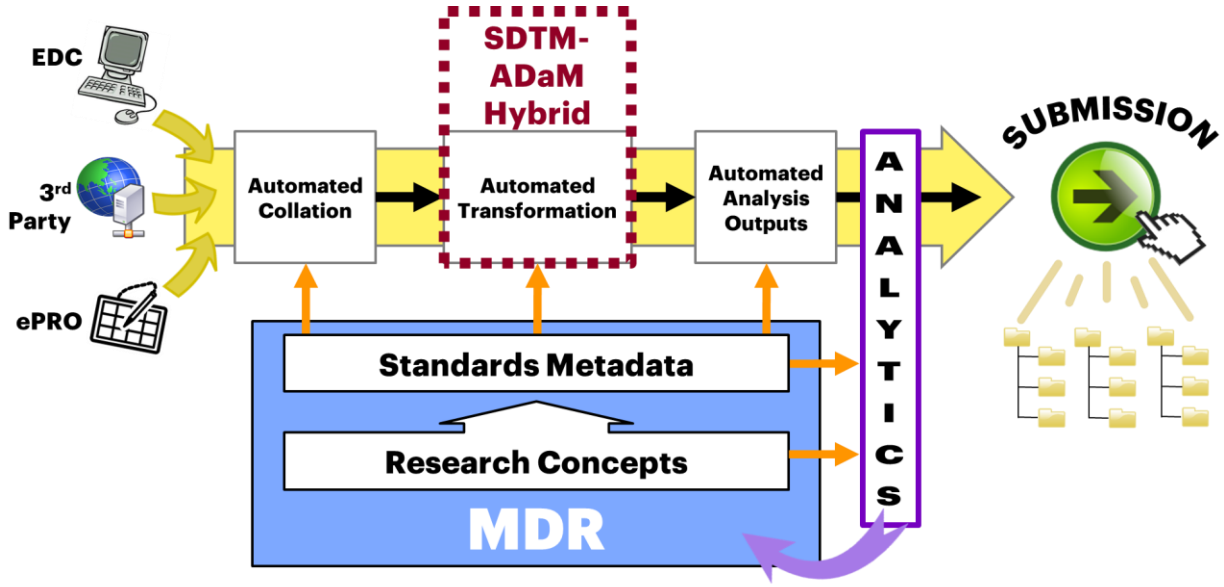
The final packaging of SDTM and ADaM datasets for submission is the relatively uncomplicated part of the process. The real questions arise from the difficulties in implementing the standards further upstream in the clinical data flow without negatively impacting them. In response, a number of sponsors have investigated and even employed hybrid approaches to SDTM & ADaM structures individually within the clinical data flow to address those operational realities in their organization, but what about a combined SDTM-ADaM hybrid structure that incorporates the necessary pieces of each? This has the potential be a structure with the flexibility to serve multiple masters throughout the clinical data flow before being appropriately refined for compliance prior to submission.

The intent of this paper is **not** to recommend that sponsors provide SDTM-ADaM hybrid datasets to any regulatory authorities as part of a regulatory submission. That would **not** be a compliant practice. The intent here is to explore potential uses of existing CDISC standard models in novel ways that could bring efficiencies in the clinical data flow, could put validated clinical results on the desk of internal sponsor evaluators sooner, and perhaps could eventually decrease time to market for treatments that patients need.

The ecosystem of conducting clinical trials is more challenging than ever. The rising costs of clinical development coupled with an increased adherence to regulatory data standards apply ever-growing pressure upon organizations to manage their clinical data flow with more efficient and more valuable methods. Opposing this progress are entrenched practices in the clinical data flow that prevent metadata from being leveraged to increase data consistency and decrease development times, that create inefficient handoff of the data between functional silos, and that reduce flexibility to adapt to partially outsourced models.

A new approach to the clinical data flow is needed to break free of these deep-rooted operational models, a path forward that will rely on automation and business intelligence brought about by metadata and standards. This metadata-driven route provides a nimble, more consistent clinical data flow that enables organizations to uncover and then apply the right business information back into their subsequent clinical trials. It is envisioned to tie together the disparate pieces of the clinical data flow in the following manner:

Intelligent Data Flow



- Data flow originates from multiple sources and proceeds through a sequenced set of automated steps that produce data and output deliverables
- Automated Collation, Automated Transformation, and Automated Analysis Outputs are driven by standards metadata that is centrally managed within an MDR as part of a set of research concepts that link the metadata into logical, medical clusters
- An Analytics layer utilizes both metadata from the MDR and clinical data and output from the clinical data flow to perform analytics designed to inform study design, assess standards, and ultimately refine the clinical data processes
- The Automated Transformation step encompasses both the generation of SDTM and ADaM datasets, and the sub-processes that support them
- Perhaps a single source operational model exists that could support the transformations needed for SDTM datasets, while also facilitating production of a segment of ADaM datasets upon which the corresponding analysis output could be quickly generated

An SDTM-ADaM hybrid model could potentially be a valuable tool within this approach to become more efficient, standardized, metadata-driven, and automated.

EXPLORATION OF THE SDTM-ADAM HYBRID CONCEPT

In the first SDTM-ADaM Hybrid meeting, it was rapidly evident that there were two mindsets on exactly how to proceed: one was to start from the ground up by putting pieces of the 2 CDISC models together in novel ways to construct examples, while the other relied upon a top-down approach that would think through the strategic objectives and obstacles in order to better define the details as part of a practical plan forward. Each has its own merits, but with time and money having the most influential, it was decided that the strategic approach was the most appropriate and cost-effective way to proceed. This paper stays true to that approach by focusing on process and design rather than detailed data definitions.

Subsequently, the first question that arose as the initial SDTM-ADaM Hybrid discussions took place was very basic and straight-forward:

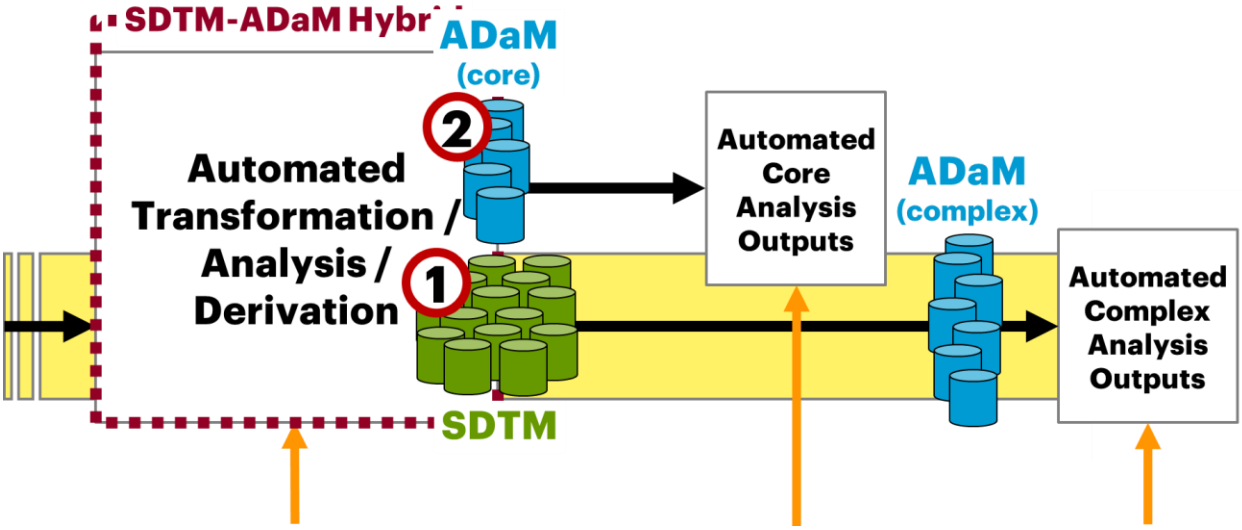
> *Would the hybrid model just be a massive domain-agnostic superset of all SDTM and ADaM variables from which the desired variables could be pulled as appropriate to create individual SDTM and ADaM datasets?*

It was a reasonable opening salvo, but was quickly dismissed as the negatives mounted rapidly. First, the danger of circularity was quickly identified if ADaM variables were to be added onto SDTM structures, whereby ADaM variables

created from SDTM variables would then be associated in reverse onto their variables of origin. Secondly, a superset containing hundreds of variables would likely result in data tables of unwieldy size that would scare off most data reviewers, require a high amount of additional processing to support the structure, and necessitate a multitude of dependencies between variables within the superset. The sheer number of relationships required within the metadata for this would likely trigger a metadata management nightmare and thus a real probability of a dangerous bottleneck in the clinical data flow. These negatives proved to be too cumbersome to warrant the idea of a superset.

One of the primary drivers for an SDTM-ADaM Hybrid was the potential to produce a core subset of analysis outputs much earlier in the clinical data flow to facilitate earlier decision-making. In the vast majority of sponsors, there are lagtimes between the production of SDTM, the production of ADaM, and the generation of TLFs, often due to functional handoffs, compartmentalized validation, and process inefficiencies. It is projected that an SDTM-ADaM Hybrid could greatly reduce those handoffs and inefficiencies for a subset of core analysis outputs by including some analysis and derivation into the hybrid model and therefore further upstream into the clinical data flow.

To deliver those reduced lagtimes and maximized process efficiencies, it was recognized that any operational hybrid model would need to produce both SDTM and ADaM validated datasets expeditiously upon command. Even though most of the analysis and derivation for these core ADaM datasets would already be “baked” into the hybrid model, it is acknowledged that there will be a few SDTM nuances that are needed for an ADaM dataset (e.g.: the --SEQ variable in SDTM). This would necessitate a quick 1-2 sequence of SDTM-ADaM production whenever the hybrid model is called upon to produce the datasets. The core ADaM datasets would immediately be leveraged for the core analysis outputs while the necessary SDTM datasets would serve as the basis for the remaining, more complex analysis outputs, as represented in the following diagram:

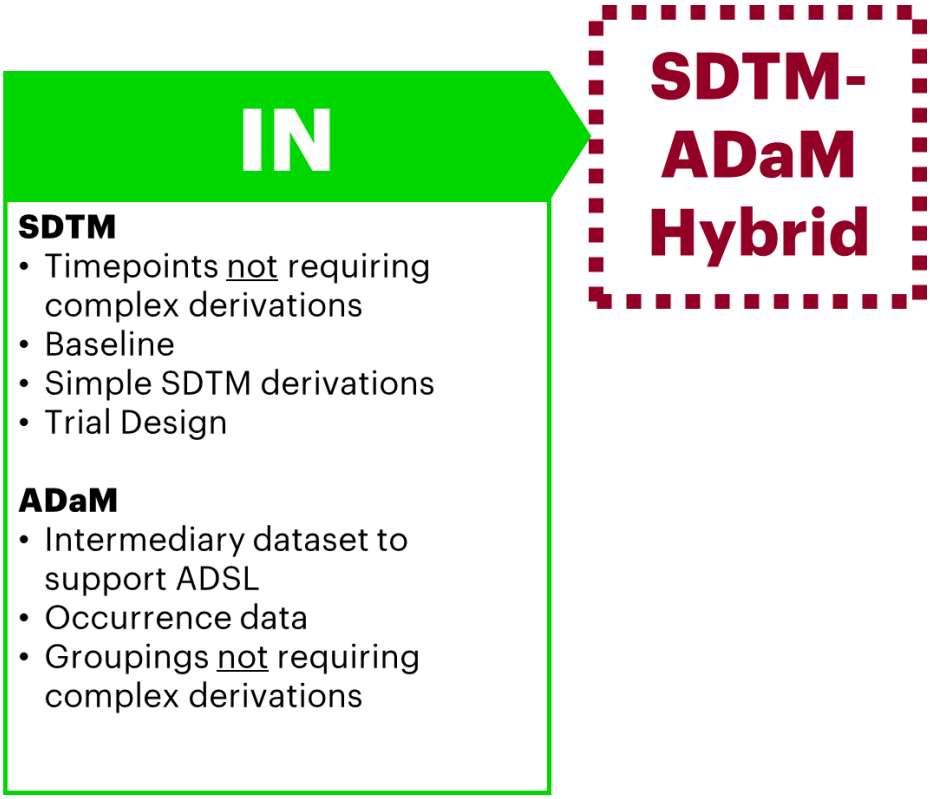


Because of the aforementioned issue of circularity of adding ADaM variables onto an SDTM structure, the exploration of the hybrid continued with the generalization of only adding SDTM to ADaM structures. Consequently, the discussion immediately turned towards differences in ADaM dataset structures as part of investigating the most valuable and most practical candidates for inclusion in the core analysis outputs. Due to the varied structures within the ADaM model, the viability of this hybrid would likely be based upon the nature of individual ADaM structures. The following table describes a high-level analysis of these differences.

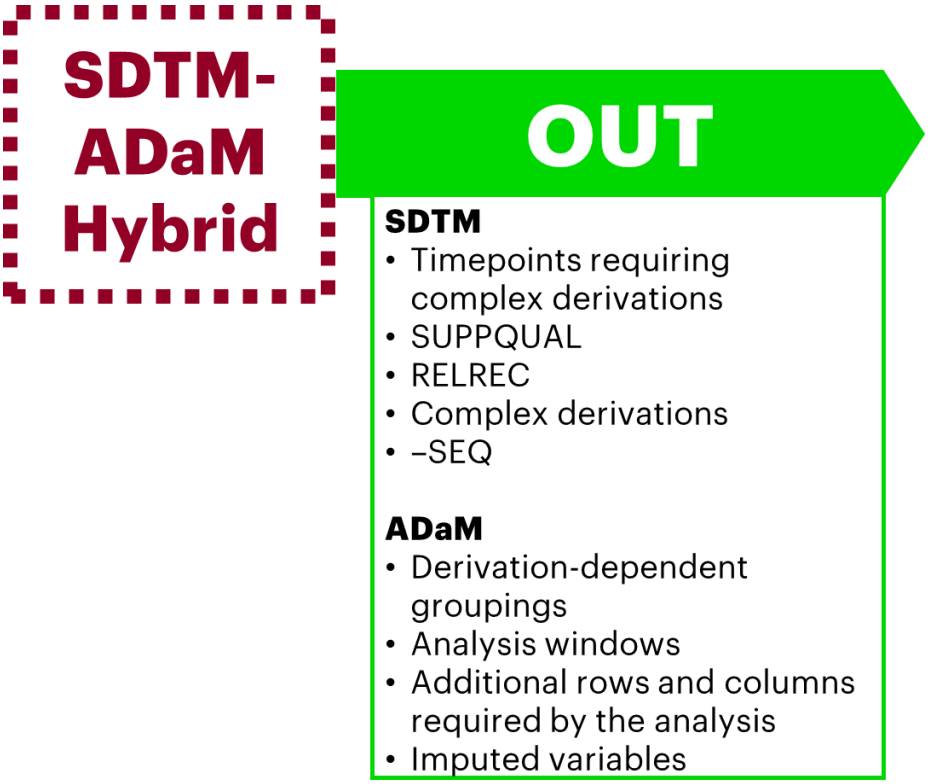
Structure	Example	Description	Evaluation
Subject-Level Analysis	ADSL	Subject-level variables provide key facts about the subject that are analysis-enabling or facilitate interpretation of analysis.	• There would be a need for the use of an intermediary dataset in the hybrid that would contain the necessary SDTM data behind derived flags, factors, and groupings • Some variables would be highly study-specific
Basic Data Structure (BDS)	Vital Signs, Laboratory Results	A BDS structure contains one or more records per subject, per analysis parameter, per analysis timepoint.	• Hybrid is probably suitable only for simple analysis structures where minimal additional columns and rows are necessary • Any SDTM qualifiers that are incorporated into PARAM are good candidates for the hybrid structure • Imputed variables would likely need to be computed after the hybrid just before the creation of the ADaM dataset
Occurrence Data Structure (OCCDS)	Adverse Events, Concomitant Medications, and Medical History	Occurrence analysis is the counting of subjects with a given record or term, and often includes a structured hierarchy of dictionary coding categories.	• Best candidate for the hybrid model, as it is already most like SDTM+ structures • Occurrence counts are manageable in the hybrid • Not only for Events and Interventions, but also for Findings results such as Labs where the result can be converted to a text-based event and analyzed as an Event
Time-to-Event (TTE)	Time to Death	The basis of Time-to-Event analysis is the time from a defined starting point to the time of occurrence of the event of interest.	• TTE computations are likely too complex to facilitate use within a hybrid model, whether they support single or composite endpoints

Moving from SDTM to ADaM, the calculations and derivations necessary to populate variables begin to increase both in number and complexity. Well-reasoned, thoughtful decisions will need to be made to determine whether each calculation or derivation can be executed as input to the SDTM-ADaM Hybrid, or must wait until the production of the ADaM dataset is performed. The following diagram describes a high-level, first cut of processing that would be needed to support this hybrid. Acknowledging that the exact order of processing would still need to be determined, this preliminary list includes, but is not limited to:

PRE-PROCESSING FOR CREATION OF SDTM-ADAM HYBRID



POST-PROCESSING FOR CREATION OF INDIVIDUAL SDTM AND ADAM DATASETS



Any operational model used within the clinical data flow would need a sound, efficient set of processes around it to provide support and consistency. The SDTM-ADaM hybrid would be no different. Furthermore, because of the unique nature of the hybrid model, these supportive processes could potentially include data management activities already performed as part of the clinical data flow. While a detailed analysis of these processes would still be needed to evaluate the absorption of them into the hybrid pre-processing, the possibility does exist for activities such as complex data checks, medical coding, and SAE reconciliation.

To ensure the vision of Intelligent Data Flow in the future, well-crafted and well-managed metadata is a key component that can deliver consistency, standardization, and automation to clinical deliverables. This cannot be underestimated. It is not enough to just have standard definition metadata straight from the CDISC SHARE repository, which only ensures that standard CDISC domains are compliant. There is a second level of metadata that is linked to the CDISC metadata that defines how the clinical processes will leverage and utilize the data throughout the data lifecycle, such as transposition metadata and results metadata. All of the metadata must be centrally managed by CDISC-knowledgeable resources within a metadata repository, preferably within a tool that provides workflows to make governance as fast and efficient as possible.

One final caveat to the role of metadata: Intelligent Data Flow relies upon the metadata to deliver speed and automation, particularly the re-useable standards that are integrated with corresponding efficiencies throughout the lifecycle. If the clinical teams do not sufficiently and appropriately use the standards, the benefits of an SDTM-ADaM Hybrid will barely be noticeable. A low rate of standards usage by the clinical team in their trial (i.e.: <60%) results in copious amounts of re-work and “starting from scratch.” If the clinical team does not remain diligent in providing the necessary content for the metadata in a timely manner, incorrect and/or inconsistent deliverables will follow. There must be an organizational commitment to the importance of updating the metadata as a critical component of study set-up.

BENEFITS

This exploration of the SDTM-ADaM Hybrid model revealed, at a high-level, the potential for benefits. While tangible numbers still need to be produced via a more in-depth analysis of the hybrid model, there are clear, early indications that warrant additional examination.

- Enables a subset of TLFs to be produced earlier in the clinical data flow, potentially focused on Safety purposes:
    - Demography
    - Adverse Events
    - Medical History
    - Concomitant Medications
    - Vital Signs
    - Laboratory Findings
- Opens up the capability to do more complex data checks earlier in the clinical data flow, thereby identifying problem areas in the data that historically aren't discovered until much later in the analysis process
- Promotes more effective review of analysis data by allowing internal reviewers to see more SDTM data next to the analysis data
- Allows decision-makers to see core analysis results much earlier in the clinical data flow, thereby enabling earlier decision-making and more judicious use of time, money, and resources
- Reduces the amount of siloed processing between functional areas that is typical when preparing the CDISC standards SDTM and ADaM
- Minimizes the number of handoffs required between these existing functional area silos
- Encourages flexibility with the clinical data flow by enabling multiple on/off ramps for data into and out of a fully metadata-driven clinical data flow

CHALLENGES

Also revealed by the exploration of the SDTM-ADaM Hybrid model were challenges to be faced and addressed as part of an implementation.

- Demonstrates a real need for a well-defined, restrictive process for production, storage, and utilization of snapshots of the SDTM and ADaM datasets, including a detailed, thorough audit trail that provides a clear record of transactions out of the hybrid model
- Reinforces the reality that a significant number of ADSL variables would be study-specific in both algorithm and content, regardless of whether the hybrid model is used or not
- Introduces potential factors that will need to be evaluated to determine where in the clinical data process that derivations are computed, including but not limited to, SDTM class and complexity of the derivation
- Sheds light on where this hybrid would probably alter the interfaces between the functional areas responsible for the production of SDTM and ADaM in the past, re-defining them as more collaboration than handoffs
- Highlights the importance of traceability from ADaM back through SDTM, especially the role of metadata in addition to the CDISC attributes to ensure traceability
    - In addition to the normal metadata link from the ADaM data back to its SDTM predecessor, an additional layer of metadata will be needed to provide the link from both the SDTM and ADaM data directly back to the SDTM-ADaM Hybrid
- Shows that a strict sequencing of processes would need to be established to eliminate circularity within the clinical data flow both before and after the hybrid model
- Confirms that the complexity of implementing such a hybrid would require an initial investment of time and effort in order to reap the benefits in the future

CONCLUSION

The industry now stands ready to drive efficiencies into their usage of SDTM and ADaM well within their clinical data flow. Late-stage conversions before submission are almost relics of the past. With potential benefits of flexibility, early reporting, and early decision-making, it is worth a continued exploration to uncover how to best utilize this new operational hybrid model from SDTM and ADaM.

REFERENCES

Study Data Tabulation Model Implementation Guide: Human Clinical Trials, Version 3.2

Analysis Data Model Implementation Guide, Version 1.1

ADaM Structure for Occurrence Data (OCCDS), Version 1.0

The ADaM Basic Data Structure for Time-to-Event Analyses, Version 1.0

ACKNOWLEDGMENTS

George Stuts, for his innovative thinking and inquisitive nature within the data integration and standardization space

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Ken Stoltzfus
Accenture
Accelerated R&D Life Sciences - Clinical Data Strategies
1160 W. Swedesford Rd.
Berwyn, PA 19312 USA
Email: kenneth.m.stoltzfus@accenture.com