Paper TT05

Digital SAP Model Using Biomedical Concepts

Tom Ratford, Veramed

ABSTRACT

The SAP is a pivotal biometrics document which is the source of truth for the detailed definitions used in clinical trial analytics and reporting. However, traditionally the SAP is created as a Word document meaning the semantic information is not easily accessible and cannot be used to drive downstream automation - resulting in inefficient manual processes existing throughout the programming lifecycle. We need access to the semantic information contained in a SAP in a machine-readable format, this enables automation of statistical analysis and reporting processes, resulting in faster and higher quality analysis projects. This presentation describes a structured data model for SAPs based on Biomedical Concepts with links back to a USDM digital protocol, and evaluates the feasibility of using this model to automate:

- SAP Model creation from legacy SAP documents
- Automation of SAP review
- Creation of a list of deliverables
- Creation and Validation of ADaM specifications
- Mock Shells
- Define.xml references

INTRODUCTION

Our industry is in a process of furthered digitisation. If adoption of CDISC ADaM & SDTM by regulatory authorities can be seen as the start of a “first industrial revolution” of biometrics, there is the case to be made that the concentrated, collective efforts around CDISC 360i and DDF can be seen as the beginning of our second industrial revolution. USDM v3.0 introduced biomedical concepts (BCs), a standard designed to assist in the creation of CRFs and to simplify the subsequent mapping from CDASH to SDTM. SDTM has also seen more recent success in automation efforts, either via BCs or other projects such as OAK SDTM (not discussed in this paper). These efforts have been made attainable due to the digitisation of the CRF, and by extension, the digitisation of the protocol. The limited growth of ADaM and TFL automation have been due to the use of static formats (docx, pdf) for the SAP, resulting in crucial linkage between SDTM & ADaM being difficult to extract or quantify. Furthermore, SAPs are typically non-standard versus a more common structure found in protocols. Automation efforts are limited to the individual efforts of companies versus the collective efforts of a standards authority. This paper argues that, with an extension to the USDM model and a reuse of the BC catalogue, crucial SAP information can be extracted into a model that can connect the elements of ADaM required to help automate common tasks.

ANATOMY OF A SAP

A SAP, whilst less consistent in structure than say: a protocol, generally contains the same information regardless of the overall exact structure. Some effort has been made to standardise the SAP [1], but we are not going to consider the use of a standardised SAP a requirement for this model. Instead, we can consider the SAP to be split into three major groups, with various sections underneath these groups that will exist within the SAP. Many of these sections are typically repeated information from the protocol. This motivates considering our models an extension to USDM, as opposed to the creation of a new standard from scratch. The purpose of this new model is to help automate ADaM and TFL programming, so we are particularly interested in modelling information from sections related to this. Typically, this is information under the “Analysis Sections” group, as well as some information from the “Useful Information” section, such as reporting conventions and technical details.

Introductory Sections	Analysis Sections	Useful Information Sections
• Table of Contents • Abbreviations and definitions • Introduction • Objectives, Endpoints and Estimands • Study Methods • Sample Size • General Considerations	• Summary of Study Data • Efficacy Analyses • Safety Analyses • Pharmacokinetics • Other Analyses	• Reporting Conventions • Technical Details • Summary of Changes to the Protocol • References • Amendments to the SAP • Appendices

Table 1: Example SAP headings split into major groups

WHAT'S IN AN ESTIMAND?

Within USDM V4[2], an estimand is linked to an endpoint (which is linked to an objective). The endpoint represents the "variable of interest" for the estimand. Alongside this we consider the various interventions this estimand is relevant to, any intercurrent events affecting this estimand, and the relevant analysis population. This is all relevant information for a TFL or ADaM programmer, who would likely be creating a similar model in their head to hold the information they need to correctly program the output. Furthermore, USDM links estimands to a biomedical concept, which allows us to reuse the existing biomedical concepts library.

EXTENDING USDM FOR SAP ELEMENTS

The proposed extension to USDM hooks into the Estimand and Objective classes of the model, as seen in Figure 1. The digital SAP elements begin with two assertions about a SAP: a SAP contains a list of planned analyses, a planned analysis is a list of planned displays. This allows the digital SAP model to cover a variety of topics within the planed analysis – provided they can be linked to an appropriate objective as discussed within a protocol. These analyses can then further be linked to a planned display, which may then be linked to appropriate estimands. For example, an objective in the protocol may be to "assess the safety of the compound", which naturally leads to "Adverse Events" & "Laboratory" planned analyses. The "Adverse Event" planned analysis may then be linked to the "Summary of Adverse Events by System Organ Class" planned display. Planned displays are linked to Estimands due to the appropriate connectivity of the Estimand class in USDM. Notably, Estimands link to Intercurrent Events, Analysis Populations, Study Interventions, and Endpoints.

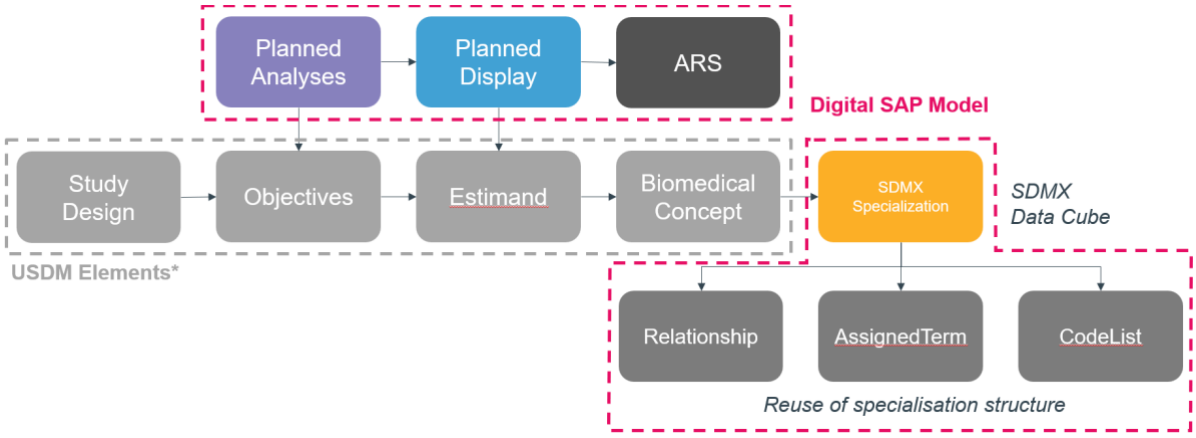

Figure 1: Proposed Digital SAP Model and its connections to USDM

The linkage to the Estimand crucially links to the existing "Biomedical Concept". Which has a new specialisation to the Statistical Data and Metadata eXchange (SDMX) format. This is an established format with simple concepts but crucially a connective hierarchical data model referred to as a "data cube". The reuse of biomedical concepts was chosen due to extensive ontological word performed as

part of this project. Numerous NCIT concepts have already been mapped together as part of this project.

SDMX & DATA CUBE SPECIALISATION

This section briefly explains the SDMX format, before explaining how this format enables the description of CDISC ADaM datasets in a concise format.

SDMX contains (amongst other things) a table representing the data itself, a data scheme description (DSD), and a series of codelists. The data in SDMX is presented in a concise format, with each column required to represent either a dimension (akin to a key variable in CDISC), a measure (a value of interest) or an attribute (an non-unique grouping variable – such as [typically] sex, geographic location, etc…). Codelists enable concise coding of variables within the data itself.

Dimensions			**Dimension**		**Attribute**		**Measures**	
USUBJID	*PARAMCD*	*PARAM*	*AVISITN*	*AVISIT*	*TRT01PN*	*TRT01P*	*AVAL*	*CHG*
101	SYSBP	Systolic BP	4	Week 4	1	Active	120	-5
102	SYSBP	Systolic BP	4	Week 4	0	Control	118	-7
103	SYSBP	Systolic BP	4	Week 4	1	Active	130	+2

Table 2: Simplified Example ADVS with SDMX classifications

Consider a simple subset of an ADVS dataset, as seen in Table 2. Columns have been labelled based on the SDMX component they represent. On particular element of note is that decode columns PARAM, AVISIT, and TRT01P have not been labelled. This is because (whilst SDMX could support these columns) we choose not to include them in the data for simplicity, their values would instead be included in a code list for PARAMCD, AVISITN and TRT01PN respectively. These would be allocated to the appropriate column within the DSD. This simple structure allows us to describe elements of the table in simple sentences – which can be used to form the beginnings of the “Relationship” elements required within the specialisation structure (as seen in Figure 1).

PARAMCD is a dimension of **ADVS**

AVAL is a measure of **ADVS**

ADVS_PARAMCD is the code list for **PARAMCD**

SYSBP is a code ID of **ADVS_PARAMCD**

Systolic BP is the code name of **ADVS_PARAMCD.SYSBP**

A key concept within SDMX is the idea of a “data cube”[3], which is the notion of multi-dimensional data. This is a familiar concept in CDISC, with both ADaM OCCDS and BDS being considered multi-dimensional. This is easily defined within SDMX as the “dimension” elements requires that the resultant columns consist of a unique identifier, therefore if a DSD has multiple dimensions we are dealing with a data cube. Another concept is a “data slice”, which is to identify pre-defined subsets of the cube using the dimensions.

LINKING BIOMEDICAL CONCEPTS TO ADAM

The Biomedical Concepts library contains many conceptual linkages between would-be ADaM parameters or variables and elements of the NCIT ontology. For example: entries for Progression-Free Survival, Overall Survival and Time to Response. We can consider linking these BCs to their appropriate ADaM, as seen in Figure 2. This can be represented within SDMX as data cubes, with dimensions able to be sliced to create additional views.

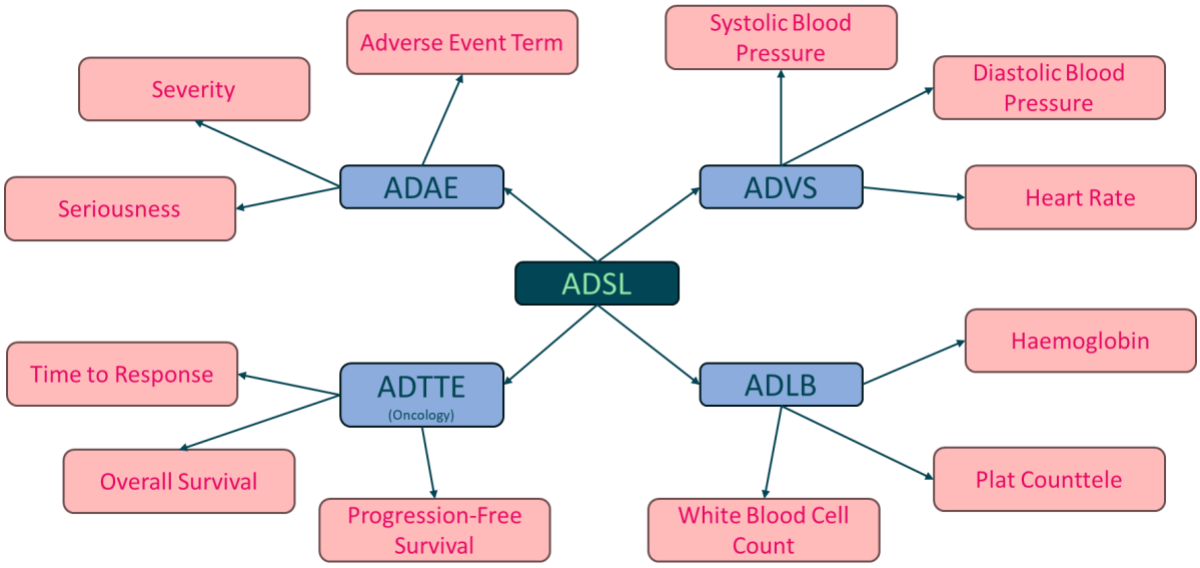

Figure 2: Example Linkage between ADaMs and BCs

Equivalently, you can view all ADaM datasets in a study as a snowflake schema, with ADSL at the centre, a dimension representing any other ADaM datasets and a dimension of Biomedical Concepts relevant to the ADaM. The ADaM dimension has a linkage back to ADSL via the subject identifier, giving access to relevant demographic information or flags. The biomedical concepts link to their relevant ADaM via the ADaMs key variables. We can view the final biomedical concepts information as a smaller table with the relevant measures and additional dimensions to link back to the ADaM. You can then view the resultant ADaM as a join of all of these linkages to the smaller tables for an individual biomedical concept.

This ADaM linkage provides an opportunity to view ADaMs as data slices using their relevant biomedical concepts.

AUTOMATION

Considering a SAP in this way provides many natural ways to help with common automation tasks. The linkage into USDM provides simplified model creation, with only additional analyses and tables required to be specified. This model can then be used within a LLM to help provide a clearer format for review of a traditional SAP document. The creation of a list of deliverables is trivial given the requirement of this data within the model, and the linkage to ARS provides the capability to create shell documents from this as well. As discussed above, viewing the ADaMs in relation to the biomedical concepts relevant to the study can provide the capability to both review ADaM specifications to ensure completeness by navigating through the links and ensuring they are present. This also provides the opportunity for simple specification automation: the linkage can be used to generate appropriate specification documents with correct variables and structures, even if there is missing method information which is still necessary to extract by hand.

CONCLUSION

The digitisation of the Statistical Analysis Plan (SAP) represents a critical step forward in the automation of clinical trial analytics. By extending the Unified Study Design Model (USDM) to incorporate structured SAP elements and leveraging Biomedical Concepts (BCs) and SDMX data structures, we can transform the SAP from a static document into a machine-readable and traversable model. Whilst additional work is required to further justify this approach, the existing work into biomedical concepts provides a untapped opportunity to reuse this structure to drive automation past CRF and SDTM.

REFERENCES

[1] TransCelerate (2024): Clinical Content & Reuse, https://www.transceleratebiopharmainc.com/assets/clinical-content-reuse-solutions/

[2] CDISC (2025): USDM V4.0, https://www.cdisc.org/ddf

[3] Cyganiak R, Reynolds D. (2014): The RDF Data Cube Vocabulary, http://www.w3.org/TR/vocab-data-cube/