Paper CD05

Considerations for CSR output production from ADaM datasets

Andrew Noller, Quintiles, Reading, England

ABSTRACT

As defined by the ADaM IG, one fundamental principle of an ADaM dataset is that it is "analysis-ready" or requires minimal programming from the dataset to produce the statistical analysis, which typically takes tabular form. However, listings of original and derived data which are also required for a CSR might require more than minimal programming from the same domain ADaM dataset as the tabular analysis due to structure or content.

In this presentation I will look at some of the particular difficulties in listing production from ADaM compliant datasets and different approaches to efficient creation. I will also share some of the efficiency gains that I have found in table production.

Introduction

When we are asked as programmers to implement CDISC standards on particular study data what are our immediate concerns? As examples we wish to check the suitability of the CRF and database to provide SDTM friendly data, are there any unusual study designs or data domains that require special attention of the SDTM guide or just how many ADaM datasets will I need to create to allow output production to be a neat step from dataset to table, listing or figure. There are many considerations to be taken on board to ensure that SDTM datasets, ADaM datasets and the define.xmls that may accompany them to meet the requirements laid out by CDISC to be compliant.

One area that may get undervalued as part of the whole process is the need for effective and efficient CSR output production. By this I mean can the outputs be produced with little or no manipulation from the dataset, is the structure of the dataset of help or hindrance to production and are there any consistencies or constraints in the ADaM standards that will enhance or otherwise production techniques that were in place before ADaM became such a well used standard.

The ADaM IG v1.0 is clear in its purpose to 'support efficient generation, replication and review of analysis results' and I believe it does have beneficial standardisation of analysis variables and a basic data structure that encourages a repeatable approach to output programming. There is a prioritization of the statistical analysis element of outputs that support the key endpoints of a study and rightfully so, best illustrated by the supporting document released this year 'ADaM Examples in Commonly Used Statistical Analysis Methods'. However, it must be considered that no matter how key these outputs and their clear metadata are to a study, the CSR is not complete without all types of summary tables and a complete set of listings that are typically required.

My task as a programmer is to ensure that the ADaM dataset for a particular domain contains all the necessary information for tables, listings and figures without the need to manipulate the dataset much or derive any new variables whilst keeping to standards set by the implementation guide. Below I will run through various examples that allow us to investigate specific issues that we may be presented with by our data or the structures and rules and also highlight where our job may be made easier by consistent use of ADaM.

BDS STRUCTURE INFLUENCE

The Basic Data Structure available to us in the ADaM IG is a structure that is well suited to Findings class SDTM datasets focussing on the parameter and results nature of these types of dataset. Depending on the CSR output shells given for a particular domain we may find that aspects such as units, categorisation and an absence of analysis provide situations that mean BDS needs to be adapted or may not even be the ideal structure for the simplest case.

SUBHEADING – ADLB Unit Example

Let's take a study that requires the lab data to be reported in both Conventional and SI units with each of these result types reported in different tables but both required to be in the same listing in 2 separate columns. Firstly, we can see that the PARAM variable must include the units after the test details. If we are to analyse the same parameter but with two unit options then two records will be required in the dataset to describe each subject, parameter, visit but these will be unique due to the units within PARAM. Here is some example data to illustrate this:

USUBJID	PARAM	PARAMCD
ABC-123-456	ALBUMIN (g/L), SI	SIALB
ABC-123-456	BLOOD UREA NITROGEN (mmol/L), SI	SIBUN
ABC-123-456	CALCIUM (mmol/L), SI	SICA
ABC-123-456	SERUM CREATININE (umol/L), SI	SICREAT
ABC-123-456	GAMMA-GT (U/L), SI	SIGGT
ABC-123-456	GLUCOSE (mmol/L), SI	SIGLUC
ABC-123-456	POTASSIUM (mmol/L), SI	SIK
ABC-123-456	MAGNESIUM (mmol/L), SI	SIMG
ABC-123-456	SODIUM (mmol/L), SI	SINA
ABC-123-456	PHOSPHATE/ PHOSPHORUS (mmol/L), SI	SIPHOS
ABC-123-456	TOTAL BILIRUBIN (umol/L), SI	SITBILI
ABC-123-456	TOTAL PROTEIN (g/L), SI	SITPROT
ABC-123-456	URIC ACID (umol/L), SI	SIUACID
ABC-123-456	ALBUMIN (g/dL), US	USALB
ABC-123-456	BLOOD UREA NITROGEN (mg/dL), US	USBUN
ABC-123-456	CALCIUM (mg/dL), US	USCA
ABC-123-456	SERUM CREATININE (mg/dL), US	USCREAT
ABC-123-456	GAMMA-GT (U/L), US	USGGT
ABC-123-456	GLUCOSE (mg/dL), US	USGLUC
ABC-123-456	POTASSIUM (mmol/L), US	USK
ABC-123-456	MAGNESIUM (mg/dL), US	USMG
ABC-123-456	SODIUM (mmol/L), US	USNA
ABC-123-456	PHOSPHATE/ PHOSPHORUS (mg/dL), US	USPHOS
ABC-123-456	TOTAL BILIRUBIN (mg/dL), US	USTBILI
ABC-123-456	TOTAL PROTEIN (g/dL), US	USTPROT
ABC-123-456	URIC ACID (mg/dL), US	USUACID

This is handled easily enough by the BDS structure as any transformation of AVAL should be added as a new parameter. In this case the unit conversion already exists in SDTM and on this occasion we need more than one unit type for lab analysis even though a single chosen unit type is the norm.

The issue really comes when we have to produce the listing from this dataset as we are asked to put both sets of results in the same listing. Typically when creating a listing from a BDS dataset it could require a simple transpose of the data using variables already given in the guide for the BDS such as lab parameters listed in separate columns where we can use PARAM/PARAMCD/PARAMN as the ID in a PROC TRANSPOSE. In this case we are asked to create a lab listing with column headers such as this:

Subject Number | Lab Test | Visit | Result in SI Units | Unit | Result in Conventional Units | Unit

Here the lab parameters are to be presented vertically following the structure of the BDS and this would be ideal if we were just reporting the one unit per parameter. Here both unit results are on the same row so we require a transpose of the ADLB to get the data in the right structure but we cannot use a PARAM variable as it contains different units. Thus, in this situation we have to keep some SDTM variables in ADLB to prevent having to derive any variables in the listing that would have been undesirable in a BDS dataset. The SDTM variables for lab test, standard units and original units (LBTEST, LBSTRESU and LBORRESU) are added to ADLB to allow the dataset to be analysis ready for the listing. LBTEST instrumental in a successful transpose and all three variables independently required to create the columns in the listing as annotated below:

[SUBJID]	*[LBTEST] [VISIT] [AVAL, transposed]*	*[LBSTRESU] [AVAL, transposed]*	*[LBORRESU]*
Subject Number	Lab Test / Visit	Result in SI Units / Unit	Result in Conventional Units / Unit

Could this situation have been dealt with by listing directly from the SDTM dataset? Perhaps but there are always going to be variables from ADSL that you would need to merge with SDTM LB before you could produce your listing and how would you display a derived variable (such as LDL/HDL ratio) that has not come from the database but has been derived as a new row in your ADLB? This would necessitate the use of ADLB for your listing and if derived parameters need to be listed in CSR outputs then this needs to be considered when planning your ADaM datasets.

SUBHEADING – Listing only Example

There have been cases where an ADaM dataset needs to be produced to create only a single listing where the SDTM would have been sufficient, also adding a level of complexity that does not add value in development.

If we have data that has not only been requested to be listed and not summarised then this could be awkward if we have a system setup to produce ADaM BDS datasets from certain domains or have just been asked to provide ADaM datasets for these domains.

Take some physical examination data, which is findings dataset PE in SDTM. If we create ADPE as we are supposed to on a BDS structure then there are a number of extra steps that we go through in getting the right variables mapped and showing a clear path from SDTM to ADaM when really this time could be saved by taking the SDTM dataset and creating a listing direct from the SDTM dataset. However, this does require merging of treatment and populations from your ADSL dataset in your listing program. So potentially there is worth in creating the ADPE for this reason.

You may also have information in your PE domain that does not fit neatly into the variables given by the SDTM Implementation guide meaning a SUPPPE is in place with data such as a free text field for 'Other' body system. If supplemental data does need to be shown in the listing then an ADaM dataset makes sense to be created.

A final thought in this area should be can we create analysis datasets for the purpose of only listing data that are not ADaM datasets and do not comply with their rules. In our PE example this might be when the listing we are given wants the body systems to be presented as columns in a horizontal data format. Do we really want to create a dataset, whose sole purpose is to do this, as a vertical BDS one? If we ask questions from a different perspective, should the listing shells where a choice is possible generally be designed with a vertical mindset to follow the flow of CDISC SDTM and ADaM structures? Or going further, does this data need to be listed at all in the future if the purpose of the listing in CSR appendices is to prove that the table summaries can be justified when there may be tools to search the data easily and replace the need for the CSR listing?

LISTING DATA MANIPULATION

Although it is rare it is possible that the data that is entered into some listings may require some manipulation from the form that it arrives from CRF through SDTM and into the ADaM dataset.

A good example of this is Questionnaire data where quite often the sheer quantity of data for a questionnaire may require some creativity in fitting all the relevant information easily on a listing. If we take Norfolk Quality of Life for Diabetic Neuropathy, one part of this ask patients if they have had symptoms in any of the following areas; feet, legs, hands, arms or none. There are seven different symptom types (e.g. Numbness) and they can select one or more of these options. The fact that they can choose one or many areas of the body causes us issues in simple listing production when you view the shell:

Listing X.X.X
Norfolk Quality of Life - Diabetic Neuropathy – Part 1
(Randomized Subjects)

Treatment Group: <insert treatment group>

Subject ID	Visit Date of Assessment	Numbness	Tingling, Pins and Needles	Electric Shocks	Other Unusual Sensations	Superficial Pain	Deep Pain	Weakness
xxxxxxxx	xxxx/ YYYY-MM-DD	Feet/Legs	None	Feet/Hands	Feet/Legs/Hands/Arms	Arms	Feet/Legs	Feet/Legs
	xxxx/ YYYY-MM-DD	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs
	xxxx/ YYYY-MM-DD	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs
	xxxx/ YYYY-MM-DD	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs
	xxxx/ YYYY-MM-DD	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs
	xxxx/ YYYY-MM-DD	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs
	xxxx/ YYYY-MM-DD	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs
	xxxx/ YYYY-MM-DD	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs
	xxxx/ YYYY-MM-DD	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs	Feet/Legs

If you take the ADaM dataset, ADQO, for this each individual entry of Feet, Legs, Hands, Arms, None for each of the seven symptoms is given its own parameter entry. This is because there is separate tabulated statistical analysis done on a total derived from this and another part of the questionnaire. The data in ADQO simply mapped from the SDTM is in the table below:

PARAMCD	PARAM	PARAMN	AVAL	AVALC
SYNNUMBF	NUMBNESS - FEET	1.1	1	Y
SYNNUMBL	NUMBNESS - LEGS	1.2		
SYNNUMBH	NUMBNESS - HANDS	1.3	1	Y
SYNNUMBA	NUMBNESS - ARMS	1.4		
SYNNUMBN	NUMBNESS - NONE	1.5		
SYNTINGF	TINGLING, PINS AND NEEDLES - FEET	2.1		
SYNTINGL	TINGLING, PINS AND NEEDLES - LEGS	2.2	1	Y
SYNTINGH	TINGLING, PINS AND NEEDLES - HANDS	2.3		
SYNTINGA	TINGLING, PINS AND NEEDLES - ARMS	2.4		
SYNTINGN	TINGLING, PINS AND NEEDLES - NONE	2.5		
SYNSHOCF	ELECTRIC SHOCKS - FEET	3.1		
SYNSHOCL	ELECTRIC SHOCKS - LEGS	3.2	1	Y
SYNSHOCH	ELECTRIC SHOCKS - HANDS	3.3		
SYNSHOCA	ELECTRIC SHOCKS - ARMS	3.4	1	Y
SYNSHOCN	ELECTRIC SHOCKS - NONE	3.5		
SYNSENSF	OTHER UNUSUAL SESATIONS - FEET	4.1		
SYNSENSL	OTHER UNUSUAL SESATIONS - LEGS	4.2		
SYNSENSH	OTHER UNUSUAL SESATIONS - HANDS	4.3		
SYNSENSA	OTHER UNUSUAL SESATIONS - ARMS	4.4		
SYNSENSN	OTHER UNUSUAL SESATIONS - NONE	4.5	1	Y

So, how best to contain the concatenated information required above within the ADaM BDS structure so that the listing will be a straightforward transpose away? We cannot use AVALC for the existing parameters as they will only show one area at a time so as to be one-to-one with the parameter they represent. We cannot create a new variable called something like LVAL(C) as this would be flagged as incorrect when we had our ADaM dataset checked using an appropriate tool. Thus, we must create a series of parameters, one per symptom, as new rows in our dataset to meet the ADaM requirements and allow us to create our listing without too much trouble. See the table below for this:

PARAMCD	PARAM	PARAMN	AVAL	AVALC
SYNNUMBF	NUMBNESS - FEET	1.1	1	Y
SYNNUMBL	NUMBNESS - LEGS	1.2		
SYNNUMBH	NUMBNESS - HANDS	1.3	1	Y
SYNNUMBA	NUMBNESS - ARMS	1.4		
SYNNUMBN	NUMBNESS - NONE	1.5		
SYNNUMB	NUMBNESS	1.6		Feet/ Hands
SYNTINGF	TINGLING, PINS AND NEEDLES - FEET	2.1		
SYNTINGL	TINGLING, PINS AND NEEDLES - LEGS	2.2	1	Y
SYNTINGH	TINGLING, PINS AND NEEDLES - HANDS	2.3		
SYNTINGA	TINGLING, PINS AND NEEDLES - ARMS	2.4		
SYNTINGN	TINGLING, PINS AND NEEDLES - NONE	2.5		
SYNTING	TINGLING, PINS AND NEEDLES	2.6		Legs
SYNSHOCF	ELECTRIC SHOCKS - FEET	3.1		
SYNSHOCL	ELECTRIC SHOCKS - LEGS	3.2	1	Y
SYNSHOCH	ELECTRIC SHOCKS - HANDS	3.3		
SYNSHOCA	ELECTRIC SHOCKS - ARMS	3.4	1	Y
SYNSHOCN	ELECTRIC SHOCKS - NONE	3.5		
SYNSHOC	ELECTRIC SHOCKS	3.6		Legs/ Arms
SYNSENSF	OTHER UNUSUAL SESATIONS - FEET	4.1		
SYNSENSL	OTHER UNUSUAL SESATIONS - LEGS	4.2		
SYNSENSH	OTHER UNUSUAL SESATIONS - HANDS	4.3		
SYNSENSA	OTHER UNUSUAL SESATIONS - ARMS	4.4		
SYNSENSN	OTHER UNUSUAL SESATIONS - NONE	4.5	1	Y
SYNSENS	OTHER UNUSUAL SESATIONS	4.6		None

Again, potentially if we viewed the listing as being sufficiently different in content and structure to that required for the table then we could have created a non-ADaM listing specific dataset. However, I believe that putting all we require in the one ADaM dataset is a better option as producing a fresh dataset specification for the purpose of 1 listing is not necessarily the most efficient especially if your organisation independently QC's the dataset. Also, this extra parameter row approach is good practice for genuine derived summary analysis that can often be done in efficacy ADaM datasets.

PRECISION DIFFERENCE BETWEEN TABLE AND LISTING

Typically in some analysis datasets there are times when a new statistic must be derived from an algorithm to create a result that was not one that could be captured on the CRF. These derivations vary from a simple average to something much more detailed and complex and the result could end up being one whose precision is captured to many decimal places. The number that we use within our summary or statistical table would usually be the most accurate version held by SAS but if this derived number is presented in a listing it is usually illustrated as a more rounded set of digits to a specific decimal place selection.

The data below shows part of a bleed information analysis dataset:

PARAM	PARAMCD	PARAMN	AVAL	AVALC
Number of Bleeds	NBLEED	1	1	1
Number of Bleeds Category	NBLEEDC	2	1	1
Number of Spontaneous Bleeds	NBLSPON	17	0	0
Number of Spontaneous Bleeds Category	NBLSPONC	18	0	0
Number of Traumatic Bleeds	NBLTRAU	19	1	1
Number of Traumatic Bleeds Category	NBLTRAUC	20	1	1
Number of Unknown Bleeds	NBLUNK	21	0	0
Number of Unknown Bleeds Category	NBLUNKC	22	0	0
Average Time to First Injection	TIMAFFIU	3	2	2.00
Time from Last Injection Any Reason per Subject	TMLIANYS	13	6	6.00
Last 3 months on study	3MONTHS	24	0	0.00
Difference from estimated prior bleeds	DIFF	55	-33.324541284033 7	-33.32
Internal	INTERNAL	31	1.6754587159533	1.68
Joint	JOINT	29	0	0.00
Muscle	MUSCLE	30	0	0.00
Overall	OVERALL	23	1.6754587159533	1.68

These are all derived rows within an ADaM dataset and, without going into any details about how some of these have been created, it is clear to see that AVAL is not always mapped one-to-one with AVALC. This goes against the ADaM Implementation guide and although this does not meet requirements it has been designed with simplicity in mind, AVAL for use in the tables AVALC for use in the listings.

A compliant method would have been to create the listing result as a new row in the data with '(Listing)' text after what exists for the original PARAM with a new PARAMCD and PARAMN. However, it would be preferable not to have to increase the number of observations in our ADaM dataset to assist listing data in being 'output-ready'.
Another option I have seen used is for AVALC to be 1-to-1 mapped but a formatting variable, AVALFMT/PARAMFMT, is added to contain the desired format for the listing to display (e.g. 7.2). This does mean a small derivation step in the listing unfortunately.

Clearly, this precision issue is something that does not frequently occur and only affects listings but could be easily remedied in the ADaM IG by the addition of an extra BDS variable that could contain a listing specific result that comes directly from the AVAL result but is not a 1-to-1 map of it. This would only be included in an ADaM dataset when strictly needed.

IMPROVED TABLE SUMMARY PRODUCTION

Table summary production is much more efficient using ADaM thanks to standard variables such as PARAM, AVAL, CHG. It can also be said that various types of standard safety tables that follow BDS closely can be produced much quicker thanks to the standardized variable naming conventions for BDS structure datasets compared to methods that may have been used before the ADaM standard was with us.

For any study that is not following ADaM to create its analysis datasets or indeed the majority of studies before ADaM arrived may not have had a setup between result based datasets such as labs, vital signs, ecgs, etc. that provided a generic naming convention or possibly not even a generic structure.

As long as the design of summary table shells such as change from baseline or shift types are consistent for your CSR then it is possible to have much more generic and straightforward programs, saving much time. The more consistent the shells are for each domain across studies the more likely standard table programs can become. Also, the more familiar you become with the naming conventions of ADaM and mapped SDTM variables, the quicker output programming can become as you gain a familiarity with variables after having an extended period of time using them.

For statistical analysis tables, the ADaM document 'Examples in Commonly Used Statistical Analysis Methods' gives helpful instruction for the steps taken to construct an 'analysis-ready' dataset that is also clear in its methodology when viewed by others, which should help review of these types of table. Again, a BDS structure with BDS variables means that more common statistical tables should be quicker to produce.

CONCLUSION

It could be said that as more time passes with use of ADaM standards that output production will become more streamlined and that time saved in analysis dataset production and table/figure production will more than offset time spent on listing production and their preparation in datasets. However, thought does need to be given to how we best enable listing production to be efficient as we draw closer to an ADaM standard for our analysis datasets.

Solutions to these listing difficulties that I have seen include bending ADaM standards by including one or two extra variables that have not been defined in the basic data structure, creating listings direct from SDTM by merging ADSL and creating customer version ADaM datasets that are used only for the listings. Not necessarily the quickest but the most compliant way is to add all you need in your ADaM dataset sticking to vertical rules by adding any extra rows if certain derivation is required for listing alone.

One long term improvement to time spent in CSR output production could be the reduction of listings in general and a move towards either automated SDTM generated listings or use of data review tools for report writers, DSMBs and any other people that need to review the data in fine detail. Surely this is better than going through large listings manually? The FDA have said that they would use SDTM to generate 'standard' listings and that SDTM could provide reviewers the ability to create their own listings. However, the CSR listings might be different from those generated from the SDTM data as the SDTM data will have no imputed dates and baseline values might vary from ADaM-created baseline values.

Lastly, having a standard that reduces the number of variables we use and standardises the naming convention for many variables allows for an efficient process to be put in place from ADaM dataset construction through to CSR output production. The more time spent using ADaM the better we will become at anticipating the variables we expect to see at the output production stage even when in the early stages of CRF and database review if output shells are generated early enough.

References

CDISC ADaM Implementation Guide Version 1.0.
ADaM Examples in Commonly Used Statistical Analysis Methods

Contact Information

Your comments and questions are valued and encouraged. Contact the author at:
Andrew Noller
Quintiles Ltd
Station House
Market Street
Bracknell
Berkshire / RG12 1HX
Work Phone: +44 (0)1344 708646
Fax: Not available at this time
Email: Andrew.noller@quintiles.com
Web: www.quintiles.com

Brand and product names are trademarks of their respective companies.