Principles and Practicalities in Building ADaM Datasets

PhUSE Single Day Event

North Carolina – September 14, 2011

Cathy Barrows

Director, Quantitative Sciences, Neurosciences

GlaxoSmithKline

Goal:

- Present practical issues / points / considerations in building ADaM datasets

- In the hopes that it will be helpful information that will benefit you…

How?

- Set the stage - Where ADaM has been (a bit of history)
- Where ADaM is now
- Learnings from the development of the ADaM General Examples Document
  - Provide an overview of the document
  - Highlight issues encountered, questions we grappled with, and some of the thinking behind the decisions made

Where ADaM has been

A bit of history…

and an analogy

Statistical Analysis Dataset Model: General Considerations Version 1.0 (Final 2005)

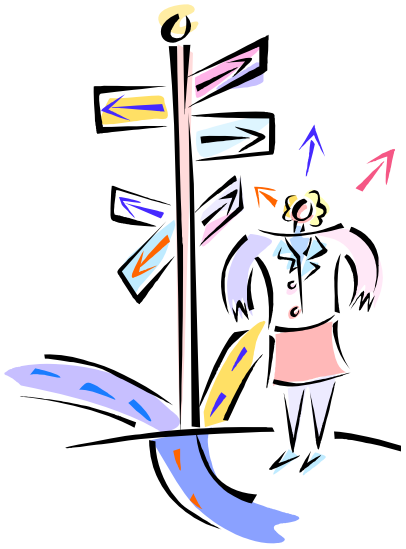
- Key Principles for Analysis Datasets
  - Analysis datasets should:
    - facilitate clear and unambiguous communication
    - be useable by currently available tools
    - be linked to machine-readable metadata
    - be analysis-ready
- Identified categories of analysis variables, defined a few specific variables:
  - --DT --DTM
  - ANLDY, which included Day 0
  - ANLDYT
  - TRTP, TRTPN, TRTA, TRTAN
- Metadata
  - Analysis Dataset Metadata
  - Analysis Variable Metadata
  - Analysis-level Metadata

"Appendix" Documents: Categorical and Change from Baseline (for comment 2005)

- Illustrated structure considered by ADaM team to be most "analysis-ready"
- "However the use of a particular structure in the example is not meant to imply that it is the recommended format."
- For example, change from baseline
  gave 1 recommended +
  2 alternative structures

Analysis Data Model: Version 2.0 (for comment 2006)

- No or little change:
  - Key Principles for Analysis Datasets
  - Metadata
- Added more ADaM variables
- Defined ADSL



- More stringent requirements for ADaM datasets:
  - Analysis datasets must
    - include ADSL
    - consist of the optimum number of analysis datasets
    - maintain SDTM variable attributes if the identical variable also exists in an SDTM dataset.
    - naming convention "ADxxxxxx."
    - consistently follow sponsor-defined naming conventions for datasets and variables

Where we are today -

- ADaM Model Document v2.1
  - Enumerates fundamental principles of ADaM
  - Introduces and defines Traceability
  - Outlines the various types of ADaM metadata
  - General considerations when creating analysis datasets including ADSL and BDS
- ADaM IG v1.0
  - Standard variable naming conventions
  - ADSL variables
  - BDS variables
  - Implementation issues, standard solutions and examples

Published in 2009

In the spirit of continuing to develop the road that is ADaM…

Well along in development:

- Compliance checks
  - Phase 1 available now
  - Phase 2 under development
- ADAE
  - Public review completed (~450 comments)
- ADTTE
  - Public review completed (~150 comments)

In active development, but still early

- Metadata
  - guidance and examples for representing metadata for ADaM
- General Occurrences
  - expand ADAE model to cover similar analyses
  - e.g. conmeds, med history, surgery
- Multiple Endpoints
  - multivariate analyses
  - analysis variables required to be on the same record
- ISS/ISE Integration
  - guidance on standards for data integration
- ADPK
  - guidance for creating PK analysis datasets
- General Examples Document…

Analysis Data Model Examples in Commonly Used Statistical Analysis Methods

- Full examples of applied ADaM implementation
  - sample data, dataset metadata, results, and results level metadata
  - Based on ADaM Model Document V2.1 and ADaMIG V1.0
- Status:
  - Has progressed through the broader CDISC review
  - Post for public review & comments by end of September (???)

Structure of the document

- Section 1 - Introduction
  - Purpose
  - Common statistical analysis methods
    - Mapping to the examples
  - Points to consider when building analysis datasets
    - ADaM concepts and principles applied in example
  - Conventions used in this document
    - Decisions made in developing the examples
  - Caveats and disclaimers

Structure of the document

- Section 2 – Examples
  - ANCOVA
  - Categorical analysis
  - Repeated measures
  - Descriptive statistics
  - Logistic regression
  - Multivariate ANOVA
  - Crossover study
  - Hy's law
- Structure of Examples:
  - Introduction
  - Analysis Metadata (dataset and variable)
  - Analysis dataset illustration
  - Analysis results (sample and results metadata)

DID NOT:

- … implement or advocate new rules or standards
- … attempt to identify specific SDTM domains
  - Focus is on analysis datasets not SDTM
- … attempt to include all possible variables
  - Did try to include those that would be included for the analysis being described
- … attempt to include all variables described in the metadata in the dataset illustrations

Points to consider when building ADs

- Optimum number of analysis datasets
  - Goal is to have the optimum number of analysis datasets needed to perform the various analyses
  - Examples provided of a single dataset that supports multiple analyses (examples 1-4)
  - **Also note** that the same analysis dataset can be used to generate descriptive statistics such as the count and percentages

Points to consider when building ADs

- Ordering of variables
  - Authors of examples each used their own ordering - no specific ordering of variables within the illustrated datasets is applied (ADaM makes no specific recommendation)
  - **Important to note** that within an example the ordering of the variables within the illustrated analysis dataset matches the order of the variables as presented in the associated metadata.

Points to consider when building ADs

- Identification of source dataset
  - When identifying the source dataset for a variable, the immediate predecessor is used, as described in the ADaM, for example:
    - AGE in ADSL - source is identified as DM.AGE
    - AGE in other analysis datasets - source is identified as ADSL.AGE

Points to consider when building ADs

- Parameter value-level metadata
  - Parameter value-level metadata are included for BDS analysis datasets
    - required in variable-level metadata for a BDS analysis dataset
  - ADSL – no parameter value-level metadata
  - Note that parameter value-level metadata is NOT a separately defined set of metadata
    - parameter identifier is “simply” an additional metadata element

Points to consider when building ADs

- Analysis-ready
  - Contain all of the variables needed for the specific analysis
  - No need for first manipulating data
  - Only “simple” manipulations (i.e., “minimal programming”), if any, to prepare for analysis

Analysis-ready – What is meant by "minimal programming"?

- Select? – yes
- Sort? – yes
- Transpose? – no
  - because of the variations in terms of the variable to be transposed, how to define the new variable names, what other fields should be included in the transposed dataset, etc.
- Merge or Join? – sponsor decision
  - difficult to draw the line as to which merges are minimal and which are no longer minimal, so no distinction made by ADaM

Options chosen / Decisions made in the development of the examples

Not intended to imply a requirement or standard!

Parameter Identifier

- Only one PARAM/PARAMCD in the dataset…
  - Parameter Identifier = *ALL* for all variables
  - Parameter Identifier = the PARAMCD for all variables
  - Combination:
    - Parameter Identifier = the PARAMCD for variables that have metadata dependent on the analysis parameter
    - Parameter Identifier = *ALL* for variables expected to be consistent across analysis parameters
    - *Allows concatenation of analysis datasets without having to re-do the metadata*

Illustration of Parameter Identifier:

Parameter Identifier	Variable Name	Variable Label	Variable Type	Display Format	Codelist / Controlled Terms	Source / Derivation
PARAMCD	PARAMCD	Parameter Code	text	$8	BMDLS	Populated with "BMDLS" for records corresponding to Lumbar Spine Bone Mineral Density (based on XX.XXTESTCD)
ALL	AVISIT	Analysis Visit	text	$11	BASELINE, MONTH 6, MONTH 12, MONTH 18, MONTH 24, MONTH 30, MONTH 36,	Refer to Section X.X of the SAP for a detailed description of the windowing and imputation algorithms used to determine the analysis visit based on ADBMD.ADY
BMDLS	AVAL	Analysis Value	float	8.1		AVAL = XX.XXSTRESN or an imputed value if XX.XXSTRESN is missing, apply the LOCF algorithm, i.e. set AVAL equal to the value for the previous post-baseline time point (AVISIT). If the previous timepoint is baseline, leave AVAL missing

Parameter value-level metadata: use of *ALL*, *DEFAULT* ?

- Many decisions about the metadata revolve around its usefulness in the future – machine readable and executable
- Two camps regarding parameter value-level metadata
  - 1) fully itemize so that every variable has metadata for every value of PARAMCD
  - 2) use *ALL* and *DEFAULT* to simplify entry for metadata that does not change across PARAMCDs
- Important to understand that *ALL* and *DEFAULT* are intended as short cuts – how you implement them and/or display them in stylesheets is up to you

Illustrating the approaches to parameter value-level metadata

Parameter Identifier	Variable Name	Variable Type	Display Format	Codelist / Controlled Terms	Source / Derivation
BIL	CRIT1FL	text	$1	Y, N	“Y” if ADLBHY.AVAL>1.5*ADLBHY.ANRHIN, “N” otherwise
ALT	CRIT1FL	text	$1	Y, N	“Y” if ADLBHY.AVAL>1.5*ADLBHY.ANRHIN, “N” otherwise
AST	CRIT1FL	text	$1	Y, N	“Y” if ADLBHY.AVAL>1.5*ADLBHY.ANRHIN, “N” otherwise
HYS1FL	CRIT1FL	text	$1		Blank if ADLBHY.PARAMTYP=”DERIVED”
HYS2FL	CRIT1FL	text	$1		Blank if ADLBHY.PARAMTYP=”DERIVED”
BIL	CRIT1FN	integer	1.0	1=Y, 0=N	From ADLBHY.CRIT1FL
ALT	CRIT1FN	integer	1.0	1=Y, 0=N	From ADLBHY.CRIT1FL
AST	CRIT1FN	integer	1.0	1=Y, 0=N	From ADLBHY.CRIT1FL
HYS1FL	CRIT1FN	integer	1.0	1=Y, 0=N	From ADLBHY.CRIT1FL
HYS2FL	CRIT1FN	integer	1.0	1=Y, 0=N	From ADLBHY.CRIT1FL
DEFAULT	CRIT1FL	text	$1		Blank if ADLBHY.PARAMTYP=”DERIVED”
BIL	CRIT1FL	text	$1	Y, N	“Y” if ADLBHY.AVAL>1.5*ADLBHY.ANRHIN, “N” otherwise
ALT	CRIT1FL	text	$1	Y, N	“Y” if ADLBHY.AVAL>1.5*ADLBHY.ANRHIN, “N” otherwise
AST	CRIT1FL	text	$1	Y, N	“Y” if ADLBHY.AVAL>1.5*ADLBHY.ANRHIN, “N” otherwise
ALL	CRIT1FN	integer	1.0	1=Y, 0=N	From ADLBHY.CRIT1FL

Parameter value-level metadata: use of *ALL*, *DEFAULT* ?

- Metadata for PARAMCDs for which the variable is null?
  - In this example, PARAMTYP="DERIVED" for the HYS1FL and HYS2FL parameters

Parameter Identifier	Variable Name	Variable Label	Variable Type	Display Format	Codelist / Controlled Terms	Source / Derivation
ALL	ANRHIN	Analysis Normal Range Upper Limit (N)	float	7.3		ADLB.ANRHIN if ADLBHY.PARAMTYP=" ", blank otherwise
DEFAULT	ANRHIN	Analysis Normal Range Upper Limit (N)	float	7.3		ADLB.ANRHIN
HYS1FL	ANRHIN	Analysis Normal Range Upper Limit (N)	float			**Not populated** for records with PARAMCD="HYS1FL"
HYS2FL	ANRHIN	Analysis Normal Range Upper Limit (N)	float			**Not populated** for records with PARAMCD="HYS2FL"

1

2

Codelist / Controlled Terminology

- Repeat the codelist metadata (whether it is a list or a link to a list) every time variable is included in a dataset?

Dataset Name	Variable Name	Variable Type	Display Format	Codelist / Controlled Terms	Source / Derivation
ADSL	AGEGR1	text	$6	<25y, 25-50y, >50y	Derived from ADSL.AGE
ADEFF	AGEGR1	text	$6	<25y, 25-50y, >50y (1)	ADSL.AGEGR1
ADEFF	AGEGR1	text	$6	(2)	ADSL.AGEGR1
ADSL	RACE	text	$50	RACE	DM.RACE
ADEFF	RACE	text	$50	RACE (1)	ADSL.RACE
ADEFF	RACE	text	$50	(2)	ADSL.RACE

Codelist – include values that do not appear in the dataset?

- Example:
  - Males and females both eligible for study
  - Only males enrolled
  - Should SEX have codelist of "M,F" or "M"?
- Decision is yes, include all possible values
  - because it could be important to know that value was an option and not used (Example: severity levels of AEs)

Codelist for PARAMCD?

- For PARAMCD, only the values actually used in the specified analysis dataset should be included in the codelist within the variable metadata for PARAMCD
- Similarly, there should be no value used as a parameter identifier for that analysis dataset that is not a PARAMCD within the dataset

Intentional blanks

- Result identifier:
  - Can be left blank ➔ the results being described are not just one specific portion of the display.
- Programming statements can be omitted :
  - Can be left blank ➔ the information provided in the other metadata elements is sufficient to describe the analysis performed.
  - In the examples document:
  - <intentionally left blank>
- Alternative:
  - leave the metadata element empty

Illustration of <intentionally left blank>

Metadata Field	Metadata
DISPLAY IDENTIFIER	Summary E.2
DISPLAY NAME	Subjects with >3% Change from Baseline in Lumbar Spine Bone Mineral Density at Month 36 (ITT Population, OC Data)
RESULT IDENTIFIER	**<intentionally left blank>**
PARAM	DXA BMD at Lumbar Spine (g/cm^2)
PARAMCD	BMDLS
ANALYSIS VARIABLE	CRIT1FL
REASON	Pre-specified in SAP
DATASET	ADBMD
SELECTION CRITERIA	ITTFL="Y" and PARAMCD="BMDLS" and AVISIT="MONTH 36" and ANL01FL="Y" and DTYPE=" " and PCHG not missing
DOCUMENTATION	See SAP Section XX for details. Percentage in each treatment group of the number of subjects with non-missing percent change data at Visit 8 (i.e., AVISIT="MONTH 36") who had >3% change in BMD from Baseline. Subjects with missing change from baseline BMD data at Visit 8 are excluded from the analysis. Number of subjects at MONTH 36 with CRIT1FL="Y" divided by the number of subjects at MONTH 36 with non-missing PCHG. Fisher's exact test used for treatment comparison.
PROGRAMMING STATEMENTS	**<Intentionally left blank>**

Where is imputation defined – AVAL or DTYPE?

- AVAL – include details of the imputation, since is part of how to derive AVAL
- DTYPE – indicates whether or not the imputation was performed for the record

Example of AVAL and DTYPE when imputation is involved

Parameter Identifier	Variable Name	Variable Label	Variable Type	Display Format	Codelist / Controlled Terms	Source / Derivation
BMDLS	AVAL	Analysis Value	float	8.1		AVAL = XX.XXSTRESN or an imputed value if XX.XXSTRESN is missing, apply the LOCF algorithm, i.e. set AVAL equal to the value for the previous post-baseline time point (AVISIT). If the previous timepoint is baseline, leave AVAL missing
BMDLS	DTYPE	Derivation Type	text	$4	LOCF	Populated with “LOCF” if XX.XXSTRESN is missing, to indicate that on that record ADBMD.AVAL is populated using Last Observation Carried Forward method

Another example of AVAL and DTYPE when imputation is involved

Variable Name	Codelist / Controlled Terms	Source / Derivation
AVAL		numeric version of XX.XXSTRESN or an imputed value **Imputation methods:** If there are non-missing data before and after the missing data, the missing data will be imputed using linear interpolation taking time of the measurement into account ("**INTERPOL**": Linear interpolation) If there are no observed data after the missing data and it is the first visit of a period the missing data will be imputed using last observation carried forward. ("**LOCF**": Last observation carried forward)
DTYPE	INTERPOL, LOCF	Populated with imputation method used when the value of AVAL is imputed

The examples illustrate various concepts, as well as providing an example of a dataset to support a specific analysis

Analysis of Covariance and more

- Analysis dataset that supports multiple analyses:
  - Analysis of covariance
  - Categorical analysis
  - Repeated measures
  - Descriptive statistics
- Included are identification of baseline values, change from baseline analysis, and handling of missing data

Examples 1 - 4

Logistic regression analysis

- Analysis dataset that supports a logistic regression including covariates
- Included is one way to use CRITy and CRITyFL in supporting a categorical analysis.

Example 5

Multivariate Analysis of Variance

- Analysis dataset that supports
  - estimation of treatment effect for multiple variables (subscale scores) in the dataset
  - an assessment of overall treatment effect (i.e., a test of the main effect of study drug on the combined subscales)
- Included are analysis results metadata for specific items on a summary table

Example 6

Multivariate Analysis of Variance

- Illustrated analysis dataset is not analysis-ready for the analysis of overall treatment effect
  - a transpose of the dataset is needed
- Included are metadata to support the transpose
  - Alternative: provide the transposed dataset as an ADaM dataset that is not compliant with BDS but fulfills the other requirements of an ADaM dataset

Metadata Field	Metadata
DISPLAY NAME	Multivariate Analysis of Variance Testing the Hypothesis of No Overall Treatment Effect at Week 6 (ITT Population)
RESULT IDENTIFIER	Test for Overall Treatment Effect Considering All Subscales
PARAMCD	ANXIETY, DPRESS, ANGER, VIGOR, FATIGUE, CONFUS
ANALYSIS VARIABLE	AVAL
DOCUMENTATION	Wilks' Lambda multivariate test of treatment effect. See SAP Section XX for details. Program: t-mood-effect.sas The MANOVA statement in PROC GLM is used to generate the result after first transposing ADMOOD. The six mood subscale scores are the dependent variables in the model, with treatment being the only independent variable.
PROGRAMMING STATEMENTS	PROC TRANSPOSE DATA=ADMOOD OUT=ADMOODHZ; VAR AVAL; ID PARAMCD; BY USUBJID TRTPN; RUN; PROC GLM DATA=ADMOODHZ; CLASS TRTPN; MODEL ANXIETY DPRESS ANGER VIGOR FATIGUE CONFUS = TRTPN / NOUNI; MANOVA H=TRTPN; RUN;

Also note the multiple PARAMCDs…

Repeated Measures Analysis of a Crossover Study

- Analysis datasets to support a crossover design study using a mixed effect model
- Included are multiple baseline types, multiple imputation methods, an analysis dataset created from another analysis dataset

Example 7

Illustrates 3 analysis datasets

- ADSL
  - the required subject-level analysis dataset
  - illustrates how the treatment and period variables are used for this study design
- ADFEV
  - includes the individual responses that are collected during the study and imputed records
- ADFEVAUC
  - includes derived response data based on the ADFEV dataset
- *The producer of the datasets elected to not combine ADFEV and ADFEVAUC into one dataset*

Categorical Analysis of Subjects Meeting Hy's Law Criteria

- Analysis dataset that supports an analysis of lab data based on Hy's Law criteria (liver function)
- Included are creation of new rows to contain new analysis parameters, the use of PARAMTYP, the use of the CRITy and SHIFTy variables

Example 8

Other interesting points to observe in the document:

- Difference between DTYPE and PARAMTYP is illustrated
- AVAL and AVALC do not both need to be populated on each row - illustrated in Example 8
- Use of different contents in the same CRIT variables as long as there is consistency within a parameter – illustrated in Example 8

Other interesting points to observe in the document:

- Variable types used in the document are those from CRT-DDS
- No “Core” column
  - The column is in the ADaMIG as part of defining variables - it is not a metadata element
- Multiple hyperlinks are indicated in the examples
  - The ability to include hyperlinks will be driven by the software that the sponsor uses for submissions
- The presentation formats used in this document for metadata are for the purposes of illustration of content only

Questions?