DH12: Metadata-Driven Tool for Creation of a Dataset for Exploratory Analyses (AXDM)

Oct 11, 2017

Alexey Kuznetsov

Content

1. Definitions
2. Dataset for exploratory analyses (AXDM)
3. Types of variables in AXDM
4. Variable naming conventions and rationale
5. Data visualization examples using AXDM
6. Modelling examples using AXDM
7. Creation of AXDM
8. Conclusion

1. Definitions

- **Data visualization**

Data visualization is the visual and interactive exploration and graphic representation of data of any size, type (structured and unstructured) or origin.[1]

- **Data mining**

Data mining is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.[2]

- **Metadata-driven program**

Metadata is data (information) that provides information about other data.[3] And metadata-driven program is the program that decides on how to process the data based on the provided metadata.

REFERENCES

[1]. SAS Institute Australia Pty Limited (Aug 03, 2016). Data Visualisation The What, Why and How. Whitepaper.

[2]. https://en.wikipedia.org/wiki/Data_mining

[3]. https://en.wikipedia.org/wiki/Metadata

2. Dataset for exploratory analyses (AXDM)

ADaM datasets

adae.xpt	6666560
adcm.xpt	40189280
addv.xpt	746720
adeg.xpt	2606255120
adex.xpt	7354000
adfa.xpt	45138480
adlb.xpt	1398837840
admh.xpt	39826800
admqs.xpt	15753600
adpc.xpt	37872000
adpi.xpt	890981920
adpis.xpt	11455298000
adqs.xpt	2376602160
adsl.xpt	2435760
adsv.xpt	2530720
adtte.xpt	9415920
advs.xpt	97746160
adzl.xpt	1203416480

AXDM dataset contents

NAME	LABEL
ADEG_1_HRMEAN_2_30	ADEG_HOLTER ECG EXTRACT_Mean Heart Rate (beats/min)_VISIT 2_30 minutes after the start of infusion
ADEG_1_HRMEAN_2_45	ADEG_HOLTER ECG EXTRACT_Mean Heart Rate (beats/min)_VISIT 2_45 minutes after the end of infusion
ADEG_1_HRMEAN_2_60	ADEG_HOLTER ECG EXTRACT_Mean Heart Rate (beats/min)_VISIT 2_60 minutes after the start of infusion
ADEG_1_HRMEAN_2_90	ADEG_HOLTER ECG EXTRACT_Mean Heart Rate (beats/min)_VISIT 2_90 minutes after the end of infusion
ADLB_CREAT_1	ADLB_Creatinine (umol/L)_VISIT 1
ADLB_CREAT_2	ADLB_Creatinine (umol/L)_VISIT 2
ADLB_CREAT_3	ADLB_Creatinine (umol/L)_VISIT 3
ADPI_WKAVNOW_1	ADPI_Weekly average current pain intensity_Week 1
ADPI_WKAVNOW_2	ADPI_Weekly average current pain intensity_Week 2
ADPI_WKAVNOW_3	ADPI_Weekly average current pain intensity_Week 3
AGE	
SITEGR1	
STRATA	
TRT01P	
USUBJID	

AXDM is a dataset containing 1 line per subject with many variables from various domains (e.g. demographic, disposition, medical history characteristics as well as efficacy, laboratory, vital signs, procedures data for all or selected parameters at each period/visit/timepoint)
The advantage is that you have all the trial data combined in one dataset and you can easily do plotting and modeling of the data.

3. Types of variables in AXDM

- Subject-level

variables coming from the “1-line-per-subject” datasets e.g. ADSL

- Transposed from BDS[1]

variables created by transposing BDS datasets variable(s) AVAL/CHG/PCHG with id variables [PARCATx] <PARAMCD> [AVISIT(N)] [ATPT(N)] by subject

[1] BDS – Basic Data Structure

3. Types of variables in AXDM (2)

- Occurrence flags

Flag variables indicating occurrence of an event for a subject in a period e.g. occurrence of a particular treatment-emergent adverse event (AEDECOD) for a subject or presence of a particular medical history term in a specific system organ class (MHBODSYS) for a subject

- Occurrence counts

Occurrence counts: numeric variables containing number of events that occurred for a subject e.g. number of occurrences of a particular treatment-emergent adverse event (AEDECOD)

- The list can be extended if required for specific exploratory analyses.

4. Variable naming conventions and rationale (Subject-level)

- Subject-level

<Dataset name>_<Orig. variable name>

Example: ADSL_RACE

Rationale:

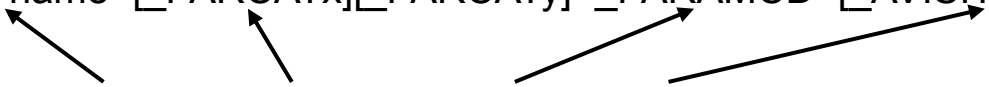
1) Clear traceability to the original dataset,
2) Ability to easily refer to all the variables coming from a particular domain as an array (e.g. ADSL_: in SAS).

Note: CDISC restriction to maximum 8 characters for a variable name length cannot be applied to AXDM.

4. Variable naming conventions and rationale (Transposed BDS)

- Transposed from BDS

<Dataset name>[_PARCATx][_PARCATy]<_PARAMCD>[_AVISIT(N)][_ATPT(N)][_ANLzzFL]



Example: ADEF_RESP_RED30P_5

Rationale:

1) Same as for Subject-level variables +
2) Easy code for creation of the dataset

(e.g.

```
proc transpose data = ADEF delimiter=_;
  by USUBJID;
  var AVAL;
  id PARAMCD AVISITN;
run;)
```

4. Variable naming conventions and rationale (Occurrence)

- Occurrence flags

<Dataset name>_FL[_Standardized term variable name]<_ Standardized term code>

Examples:

ADAE_FL_AEDECOD_HEADACHE

ADMH_FL_MHBODSYS_1

(Standardized term code here is a number since the text within the variable may be too long. The complete description of the term may be provided in the variable label)

- Occurrence counts

<Dataset name>_N[_Standardized term variable name]<_ Standardized term code>

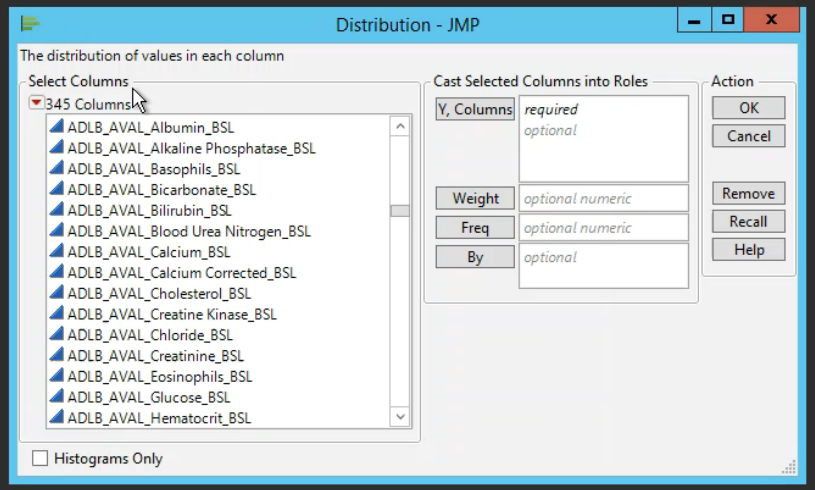
Example:

ADAE_N_AEDECOD_HEADACHE

The same concept as for the flags, but the variable is numeric and represents the counts of occurrences of the AE.

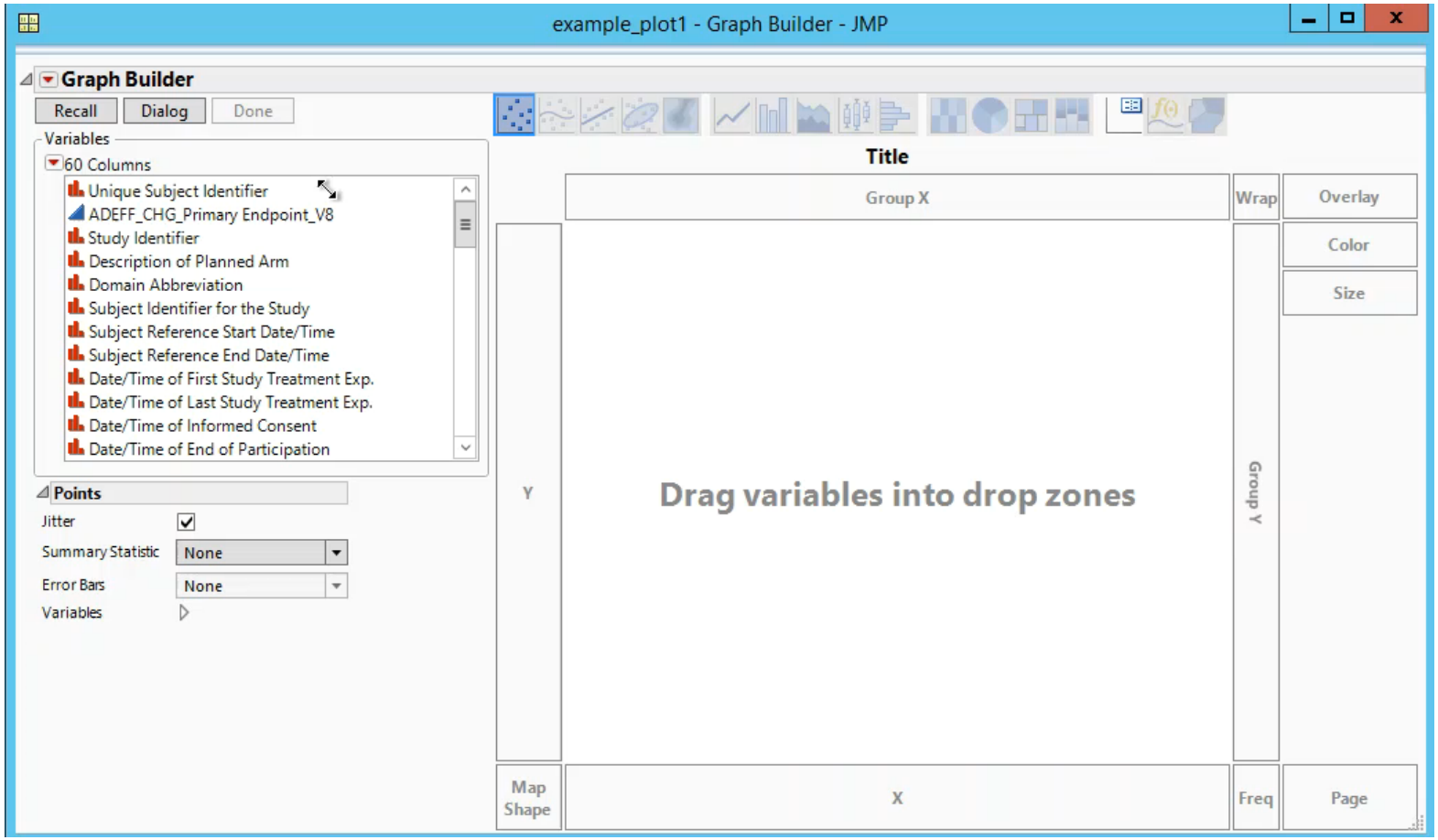
5. Data visualization using AXDM and SAS JMP®

- Example of plotting distributions of baseline laboratory values from AXDM



5. Data visualization using AXDM and SAS JMP®

- Example of plotting primary endpoint on y-axis versus baseline lab value on x-axis by medical history subgroup



6. Modelling in SAS® using data from AXDM

Example:
The following code applies a model with effect selection in the framework of general linear models which is supposed to identify if age, sex, race, medical history event involving any particular body system or baseline creatinine level can explain change from baseline to visit 8 in quality of life score:

```
proc glmselect data=AXDM(where = (RANDFL = "Y" and
TRT01P ne "Placebo"));
  class ADMH_FL_MHBODSYS_:;
  model ADEF_EQ5DTOT_V8_CHG = AGE SEX RACE ADLB_CREAT_BSL
  ADMH_FL_MHBODSYS_: / details=all;
run;
```

7. Creation of AXDM

Metadata-driven approach for creation of AXDM consists of 3 steps:

- creation of metadata for a trial
- automatic generation of a SAS program based on the metadata
- manual fine-tuning the SAS program if required and running it

7. Creation of AXDM (Metadata table)

An example of a metadata table for AXDM:

MEMNAME	BY	ID	IDLABEL	OCCUR	WHERE	VAR
ADSL	USUBJID				RANDFL = 'Y'	
ADQS	USUBJID	PARAMCD AVISITN	PARAM AVISIT		RANDFL = 'Y' & PARAMCD = 'PAINNOW' & ANL01FL='Y'	CHG
ADVS	USUBJID	PARAMCD AVISITN	PARAM AVISIT		RANDFL = 'Y' & not missing (AVAL)	AVAL
ADMH	USUBJID			MHBODSYS	PARCAT2 = 'MEDICAL HISTORY' & not missing (MHBODSYS)	
ADLB	USUBJID	PARAMCD AVISITN	PARAM AVISIT		RANDFL = 'Y' & not missing (AVAL) PARCAT2 = 'CENTRAL LABORATORY'	AVAL

Note:
Parts of the metadata table in black font can be generated automatically based on the contents the SAS library.
The part in red most likely is added by the user accounting for the needs of the analysis.

7. Creation of AXDM (Generated SAS code)

```
data ADSL;
      set temp.ADSL(where =(RANDFL = 'Y'));
run;
%get_dups(ds=ADSL, by= , id=%str());

data ADQS;
      set temp.ADQS(where =(RANDFL = 'Y' & PARAMCD = 'PAINNOW' & ANL01FL='Y'));
      attrib _all_ label=" ";
      length id $40 idlabel $200;
      id=catx("_",trim("ADQS"), "CHG", trim(PARAMCD), trim(AVISITN));
      idlabel=catx("_",trim("ADQS"), "CHG", trim(PARAM), trim(AVISIT));
run;
%get_dups(ds=ADQS, dsout=ADQS_CHG, by= , id=PARAMCD AVISITN,bdupsby=PARAMCD);

proc transpose data = ADQS_CHG_dups (where = (bdups=0)) out=t_ADQS_CHG(drop = _name_);
      by USUBJID;
      id id;
      idlabel idlabel;
      var CHG;
run;
...
```

8. Conclusion

Pros:

- AXDM is convenient to do easily customizable data visualization and high level data mining
- AXDM is more understandable for people with minimal or no CDISC experience
- Semi-automated metadata file generation from SAS library contents eases the creation of such a dataset

Cons:

- The data structure does not cover all the data that could be stored in ADaM/SDTM
- The metadata file has to be prepared by a statistical programmer or a biostatistician with good knowledge of the particular trial data

Questions?

Don’t hesitate to contact me after PhUSE 2017 at Alexey.Kuznetsov@grunenthal.com