

Science For A Better Life

Remapping of Codes (and of course Decodes) in Analysis Data Sets for Electronic Submissions

Joerg Guettner, Lead Statistical Analyst

Bayer Pharma, Wuppertal, Germany



Agenda/ Content

- Introduction
- Codelists – the place to store the remapping information
- Metadata
- Workflow to update codes and decodes
- Conclusion

Introduction

- During a life cycle of a project codes are subject to change
- Two main reasons that make a remapping of codes necessary:
  - FDA requirement[1]
    (variable names and codes in analysis data sets should be consistent across studies and where feasible, the NCI CDISC Vocabulary should be used)
  - Integrated analyses
    (consistent approach for analyses)

[1] US Food and Drug Administration. Guidance for industry: study data specifications

Introduction

Prominent example: laboratory tests (codelist LBTEST)

- First release of CDISC controlled terminology: < 100 terms
- Meanwhile: > 700 terms

Handling for laboratory tests not present in codelist LBTEST at the time of analysis:

- Extend codelist by adding sponsor defined term

Problem:

- Sponsor defined terms need to be updated in case that CDISC introduce controlled terms for these laboratory tests

=> Code remapping needed

Introduction

- Analysis data sets following Analysis Data Model (ADaM) have often pairs of corresponding variables containing a decode and a code, e.g. AVISIT and AVISITN (analysis visit)
  - In case of a necessary remapping both (code and decode) have to be updated

- Identifying corresponding variables maybe tricky due to limitation of eight characters for variable names, e.g. LBMETHOD and LBMTHODN (method of test or examination)

Introduction

What is needed for the workflow?

- the remapping information
- the codelist of a variable
- which variables represent a pair of corresponding variables, containing a decode and a code

Codelists – the place to store the remapping information

- Bayer uses several repositories to store codelists:
  - Global Medical Standards / Therapeutic Area Standards
  - Project Standards
  - Analysis Data Sets (also on project level)

- Advantage:
  - All studies share the same codelists (and do the same remappings).

- Important restrictions:
  - It is not allowed to delete Codes.
  - Meaning can not be changed.
    (e.g. COLD: Common Cold ≠> Chronic Obstructive Lung Disease)

Codelists – the place to store the remapping information

- Due to these restrictions it is possible to store the remapping information in the codelists as obsolete codes may not be deleted

- To distinguish between active and retired codes and for traceability, additional administrative variables needed,
  - STATUS: A – active, R – retired
  - REASON: short description for changes on the record
  - SYSDATE: date and time of last change of the record

- Remapping information can be stored in just one additional variable
  - UPMAP: remapping information

Codelists –
the place to store the remapping information

Extract of Codelist LBTEST

FMTNAME	START	LABEL	TYPE	UPMAP	STATUS	reason	sysdate
LBTEST	ETHANOL	Ethanol	C		A	creation	28FEB2011 :17:21:38
LBTEST	ETHYLALC	Ethyl Alcohol	C	ETHANOL	R	updated feb 28 2011	28FEB2011 :17:21:38
LBTEST	FAC7	Factor VII	C	FACTVII	R	Update request 2011-03-30	30MAR201 1:16:11:24
LBTEST	FACTVII	Factor VII	C		A	Creation	20DEC201 0:07:32:30

Codelists –
the place to store the remapping information

Limitations:

- Only one-to-one mapping possible, not one-to-many
- Remapping to a different codelist is not possible

Metadata

- Bayer's production area is strongly metadata-based, i.e.
  - Data must comply with metadata
  - Checks during transfer to production that
    - all codelists used in the data exist
    - all codes used can be decoded
- Metadata available as SAS data sets
- Metadata used in the workflow for
  - to identify the codelist used by a variable
  - to identify the pairs of corresponding variables containing a decode and a code

Metadata

- Bayer system did not allow to add variable in the metadata without changing the underlying system
- Existing variable had to be used: COMMENTS
- To distinguish between normal comments and variable containing the associated code:
  - use variable name in square brackets and uppercase at end of comment e.g. [LBTESTCD]

Metadata

Extract of Metadata for Analysis Dataset ADLB

VARSEQ	SASNAME	LABEL	TYPE	OUTFORM	CODLST	DESCRIPT	COMMENT
11	LBTESTCD	Lab Test of Examination Short Name	C	8	LBTEST	LB.LBTESTCD	
12	LBTEST	Lab Test of Examination Name	C	40		LB.LBTEST	[LBTESTCD]
23	PARAM	Parameter Description	C	200		New code based on combination of LBTEST/LBTESTCD, LBSTRESU, LBSPEC, LBMETHOD and …	[PARAMCD]
24	PARAMCD	Parameter Code	C	8	X_PARAMC	New code based on combination of LBTEST/LBTESTCD, LBSTRESU, LBSPEC, LBMETHOD and …	
49	AVISIT	Analysis Visit Description	C	40		Windowed value of VISIT according to rules in SAP	[AVISITN]
50	AVISITN	Analysis Visit Number	N	9.4	Z_AVISIT	Windowed value of VISITNUM according to rules in SAP	

Metadata

Why this extra efforts to use code to remap the decode?

- Check on the content of the variable containing the code, but not on variable containing the decode
- Cases where code and decode do not match 100%
- Real world example:
  Unit ‘DA’ misspelled ‘Da’

Workflow to update codes and decodes

Requirements:

- Formats as SAS data sets
- Remapping information stored in additional variables in the formats
- Metadata as SAS data sets
- Codelist of a variable stored in the metadata
- Pairs of corresponding variables containing a decode and a code stored also in the metadata

Workflow to update codes and decodes

Workflow:

0. Add remapping information in formats data sets
1. Search the codelists for codes to be remapped
2. Identify the datasets and variables that use codes to be remapped in the metadata
3. Update the identified variables and datasets

Workflow to update codes and decodes

0. Add remapping information in formats data sets

- At Bayer done by different teams (global, project, project statistical analysts)

Workflow to update codes and decodes

1. Search the codelists for codes to be remapped

- Search for variable UPMAP populated in the codelists
- In case of multiple remappings (e.g. A remapped to B, B remapped to C), only latest remapping information should be kept (A remapped to C)
- Result:
  - the codelists with codes to be remapped
  - codes to be remapped
  - and code to be mapped to

Workflow to update codes and decodes

2. Identify the datasets and variables that use codes to be remapped in the metadata

- Identify datasets with variables using codelists containing codes to be remapped based on results of first step
- Results:
  - data sets using codelists with codes to be remapped
  - corresponding variable pairs containing code and decode

Workflow to update codes and decodes

3. Update the identified variables and datasets

- Search for codes to be remapped in identified variables and data sets
- Update codes and decodes where necessary

Conclusion

- Codes and decodes can be easily remapped with this workflow
- Limitation: AVAL / AVALC in ADaM can not be updated with this workflow
  - mixture of character and numeric values or even codelists



Science For A Better Life

Thank you!