CDISC Italian User Network 2021

Virtual Event | 3 December 2021

SIMeF
Società Italiana di Medicina Farmaceutica
Italian Society of Pharmaceutical Medicine

IBIG
Italian Blostatistics Group

sas

cdisc

What is the Use of the Controlled Terminology ?

CDISC Italian User Network – 2021-12-03

Thierry Lambert

Outline

- The problem
- Internal Tool and Types of Codelists
- Conclusion

The problem

- The CT = a series of lists of codes (codelists).
  - for each item: an ID, a submission value, a definition, an "NCI preferred term"
  - for each codelist: same + “extensible yes / no” (in this case, the “submission value” is never submitted...)
  - full stop: no information "what is the use of this codelist?"
- The SDTM-IG
  - does contain codelist names for certain variables
  - but it is updated every X years
  - while the CT is updated quarterly
  - moreover, the codelists "pointed to“ by the SDTM-IG are a tiny minority of the CT codelists, when it does not point "in the void" (for example because the codelist has been renamed)
  - an example: SCSTRESC has no codelist in the SDTM-IG, while the CT has... 12!

A solution

- AdClin has developed over time internal tools to help with SDTM mapping
- We present here the current state of our internal CT visualization tool
  - initially just a series of codelists in HTML format
  - now contains the SDTM, the SDTM-IG, and links to the CT, needed to understand how to use the codelists and their "geography"

A simple codelist: ACN

- not extensible
- presentation "Value+Synonyms / Label / Definition / Id"
- addition: link to the variable that uses it

The TESTCD / TEST system

- Example: the first 2 codelists of the CT, TENMW1TC and TENMW1TN
- The items are exactly the same (but not in the same order), except "submission value"
- This is because the value in TENMW1TC is for --TESTCD, while the value in TENMW1TN is for --TEST
- We therefore have in fact only one codelist, but each item has two values, the one of --TESTCD and the one of --TEST (the latter different from "NCI Preferred Term" because –TEST is limited to 40 characters)
- In our viewer, all these "pairs" are a single codelist

Paired --ORRES/--STRESC codelists

- Appeared in CT 2021-06-25
- For the QRS convention (QS, FT, CC): the code is in --STRESC, the decode in --ORRES
  - Example for ECOG: ECOG101OR/ECOG101STR
- Look like TESTCD/TEST, but... how do we pair the items?
  - the Concept Ids are not the same (they are in TESTCD/TEST)
  - in ECOG101OR, “Dead” is the 4th item
  - in ECOG101STR, it is the 6th one
  - nothing found allowing a match by a computer program
  - experimental:
    - sort by concept id (touch wood: they added the items in the same order)
    - human review to confirm the matches
    - if it stays like this, we need a new type of codelist (not yet implemented in our tool)

The subcodelists

- Example: STENRF
- According to the SDTM-IG tables of variables, it applies to the --STRF, --ENRF, --STRTPT and --ENRTPT variables
- But in fact no: there are three sets of codes belonging to STENRF (3 "subcodelists"):
  - codes for --STRF and –ENRF
    - (position if a start or end date vs. a period of time)
  - codes for --STRTPT
    - (positions of a start date vs. a point in time)
  - codes for --ENRTPT
    - (positions of an end date vs. a point in time)
- End result:
  - STENRF as a whole does not apply to any variable
  - STENRF is "cherry-picked":
    - it has subcodelists who independently do their "shopping" among the available codes
    - so some codes appear in several subcodelists (for example "BEFORE" is found in all sub-codelists)

A very productive codelist: NY

- 5 known subcodelists
  - Y
  - N
  - Y N
  - Y N U
  - Y N NA (assumption for --OCCUR)
- Information is often in the text of documents
  - --PRESP and --SPCUFL: in the SDTM
  - therefore semantic annotation of the HTML (“data-” attributes)
- Lists in “Applies to”:
  - from CDASH SUPP variables
  - from codetables
    - example: DD.DDSTRESC where DDTESTCD = HMROIND
    - in 2020-12-18/DD_Codetable_Mapping.xlsx

Codetables

- Available on the CDISC website
- Maintenance frequency unknown
  - but seems lately to follow the quarterly rhythm of the CT
- Sometimes not in line with the CT
  - codes "CNEW"
- Gives relationships between codelists
  - but no information "applies to these variables"
  - DD_Codetable_Mapping.xlsx: NY and AGEU, same variable?
  - are the differences intended?
    - AUTOPIND / DTHCOIND / DTHWIND => NY_YNU
    - HMROIND => NY
- Contains errors
  - visibly maintained manually
  - when the codes do not match the CT, the error is obvious
  - sometimes (above) we don't know if it's a mistake or not

The QRS system

- Three codelists QSCAT, FTCAT and CCCAT
  - QSCAT for QS, FTCAT for FT, CCCAT for … RS!
- Hundreds of codelists like TENMW1TC/TENMW1TN
  - 302 in the CT of 2021-09-24
  - apply to xxTESTCD and xxTEST
- The xxCAT synonym allows to "predict" the codelist for xxTESTCD/xxTEST
- In fact a single codelist "QRSCAT" ("aggregated codelist"), with an additional column "domain"
  - the value of "domain" for an item changes with the CT versions
  - recent: ECOG changed from QS to CC = RS
  - problem on the SDTMIG side: the QS / FT / CC distinction is partially arbitrary, it should not be attached to the domain

The "sliced" codelists

- Compared to "cherry picked":
  - they have sub-codelists
  - but the subcodelists partition the codes, without intersection
- Example: EGSTRESC
  - each value corresponds to a value of EGTESTCD and only one
  - the SDTM mapping to EGSTRESC therefore allows predicting the value of EGTESTCD, which turns out to be just a classification of ECG observations
- Another example: ETHNICC
  - each value of CETHNIC (a SUPPDM of CDASH) predicts the value of DM.ETHNIC
  - except that they forgot "Puerto Rican"

Unit codelists

- Example: we want to code in LBORRESU "mg/mL"
- LB.LBORRESU > UNIT codelist
- In UNIT: find "mg/ml"
- It is a synonym:
  - you must code "g/L"
  - whose preferred term is... "Kilogram per Cubic Meter"!
  - all this to return to mg/mL in the ADaM PARAM variable...
- Long-known solution: UCUM
  - project: code in UCUM and automatically map to the latest CT fad

Information added in some codelists

- LBTESTCD
  - numerator and denominator
- TSPARMCD
  - multiple records
  - core (FDA vs. IG)
  - notes (FDA and/or IG)

Complex cases

- FA
  - the variable FATESTCD has codelists by therapeutic area
  - depending on the value of FATESTCD, the type of FASTRESC varies
  - in particular CVFARS, "sliced" by CVFATSCD
- EG
  - two systems in parallel for the same variables:
  - the variable EGTESTCD follows the codelists EGTESTCD and HETESTCD
  - the variable EGSTRESC follows the codelists EGSTRESC and HESTRESC
- RS: hell
  - RSCAT follows two codelists: ONCRSCAT and CCCAT
  - CCCAT follows the QRS logic, which predicts RSTESTCD codelists
  - but ONCRSCAT has a completely different logic associated with ONCRTSCD and ONCRSR
  - fun fact: according to the codetables, the codelist of RECIST 1.1 non-target responses is not the same when it is associated with iRECIST (not supported by our tool: error message on import)

Conclusion

- Need to add input information
  - Semantic annotation of the HTML of standards
  - Information missing in the codetables (which variables?)
  - Numerators/denominators in LBTESTCD
  - TSPARMCD information
  - A file “standards.rb”: a DSL ("domain-specific language") which allows you to easily enter knowledge
- This information should be maintained by CDISC and publicly available
- The CDISC should have a computer model like ours to validate its own production:
  - Errors in the SDTM IG
  - Errors in the CT