

Most Common Issues in ADaM Data

Sergiy Sirichenko, Michael DiGiantomasso
PhUSE SDE, Chicago, IL
April 30, 2015

Disclaimer

The views and opinions presented here represent those of the speaker and should not be considered to represent advice or guidance on behalf of the Food and Drug Administration.

Topics

- › New ADaM checks in OpenCDISC
- › Methodology
- › Common issues

OpenCDISC ADaM Checks

- › Introduced in v1.2, 2010
- › Conformance with ADaM IG
 - › Added Metadata checks
- › Used in FDA DataFit (OpenCDISC Enterprise)
- › Free Community version available

New CDISC ADaM v1.3 validation checks

- › 2015-03
- › ADaMIG v1.0
- › ADAE
- › BDS-TTE
- › +75 new checks

New OpenCDISC ADaM Checks

- › Already available for Enterprise clients
- › 73 out of 75 new checks were implemented
 - › 255, 259
- › 10 OpenCDISC checks
 - › Metadata checks
 - › Value Level (VL) metadata checks
 - › SD1228-SD1231

Value Level Metadata Checks

- › To validate study specific info
- › Uses define.xml v2.0
 - › VL Codelists
 - › Mandatory VL
 - › VL Datatype
 - › VL Length
- › Available in Enterprise only

Home

Configure

Design Studies

Manage Standards

Data Structures

Terminology

Specs

Rules

Validate

Report

Administer

Home > Design > Validation Specs > ADaM v1.0

ADaM v1.0

Properties

Domains

Variables

Rules

Remove Selected

since ADaM-VC-1.3

	ID	Message	Metadata	Category	Severity	CDISC
☐	AD0245	CNSR is present but not all of STARTDT, ADT and ADTM are present	✓	Presence	Error	245,246
☐	AD0247	CNSR is populated but not all of STARTDT, ADT, ADTM and AVAL are populated		Consistency	Error	247,248,249
☐	AD0250	EVNTDESC and/or CNSDTDSC is present but CNSR is not present	✓	Presence	Error	250,251
☐	AD0263	*DTF is populated but *DT is not populated		Consistency	Error	263,266
☐	AD0264	*TMF is populated but *TM is not populated		Consistency	Error	264,265

Community Report Rules Tab

opencdisc-report-2015-04-28T18-34.xlsx - Microsoft Excel

	A	B	C	D	E	F
	OpenCDISC	Publisher ID	Message	Description	Category	Severity
22	AD0037	37,286,292,298	Inconsistent value for --GR--N	2-Way-Map_22: Within a study, All *GRyN values must be the same for each unique value of *GRy when primary variable *GRy is populated. ADaM IG 1.0 p.11 general variable naming conventions #10: If any combining of the SDTM character categories is done, the name of the derived ADaM character grouping variable should end in GRy and the name of its numeric equivalent should end in GRyN where y is an integer from 1-9 representing a grouping scheme. (286,292,298 since ADaM-VC-1.3). CDISC created 3 new explicit rules for ADAE which were already covered by this general rule.	Consistency	Error
23	AD0038	38,285,291,297	Inconsistent value for --GR--	2-Way-Map_22: Within a study, All *GRy values must be the same for each unique value of *GRyN when primary variable *GRy is populated. ADaM IG 1.0 p.11 general variable naming conventions #10: If any combining of the SDTM character categories is done, the name of the derived ADaM character grouping variable should end in GRy and the name of its numeric equivalent should end in GRyN where y is an integer from 1-9 representing a grouping scheme. (285,291,297 since ADaM-VC-1.3). CDISC created 3 new explicit rules for ADAE which were already covered by this general rule.	Consistency	Error
30	AD0047	47,49,50,51,52, 55,71,72,88,89, 90,194,195,257 ,260,261,262,2 78	Required variable is not present	unconditional-presence: Variables described in ADaM as Required must be included in the dataset. (since ADaM-VC-1.3: 257-278). Formerly AD1001.	Presence	Error
185	AD0245	245,246	CNSR is present but not all of STARTDT, ADT and ADTM are present	conditional-presence: When CNSR is present, the STARTDT, ADT and ADTM must be included in the dataset. In Time to Event studies, the basis of analysis is the time from a defined starting point to the time of occurrence of the event of interest. (since ADaM-VC-1.3)	Presence	Error
186	AD0247	247,248,249	CNSR is populated but not all of STARTDT, ADT, ADTM and AVAL are populated	conditional-population: When an event is censored, all 4 variables must be populated. CNSR qualifies the AVAL variable. ADT and ADTM are the Analysis date and time of event or censoring associated with AVAL. STARTDT is the original date of risk for the time-to-event analysis. This is generally the time at which a subject is first at risk for the event of interest evaluation. (since ADaM-VC-1.3)	Consistency	Error
			EVNTDESC and/or CNSDTDESC is present but	conditional-presence: In Time to Event studies, EVNTDESC is the Description of the event of interest or an event that warrants censoring. Therefore CNSR should be present. CNSDTDESC Describes the circumstance represented by the censoring date if different from the event date that warrants censoring. If present, CNSR must be present. Note that adding CNSDTDESC provides additional information when there is a difference between the event date that warrants censoring and the actual date of censoring. If the event that warrants censoring is the same as the onset date of the event that represents the censoring date, then the CNSDTDESC column is		

Dataset Summary Issue Summary Details Rules

Ready 32 of 249 records found

Methodology

- › Data
 - › 62 studies
 - › 2013 – 2015
 - › 28 submissions
 - › 22 sponsors
- › Process
 - › Pull validation results
 - › Clean false-positive messages
 - › Summarize validation results

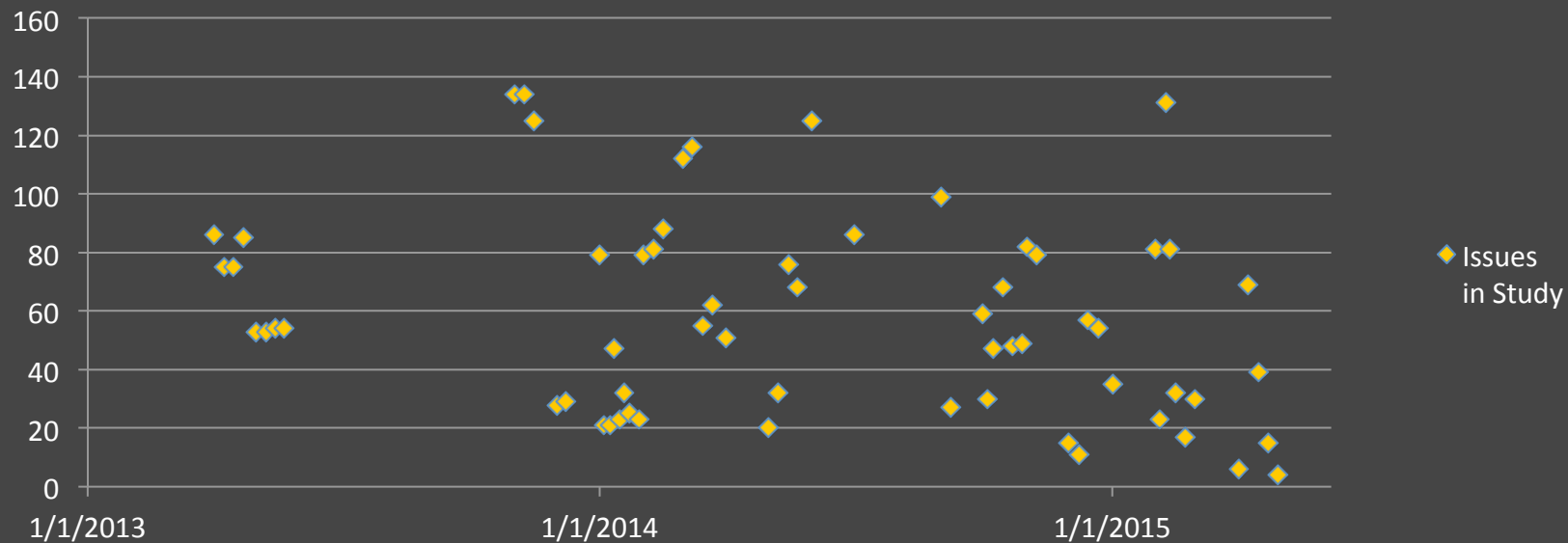
ADaM Data Summary

- › Datasets in a study
 - › 7 to 51
- › Records in a study
 - › 6K (12 datasets) to 36M (12 datasets) (x 6,000)
- › ADaM data is very diverse across studies compared to SDTM data
- › Data quality of ADaM and SDTM data are usually independent due to different teams involved
 - › Statistical programming vs. Data management

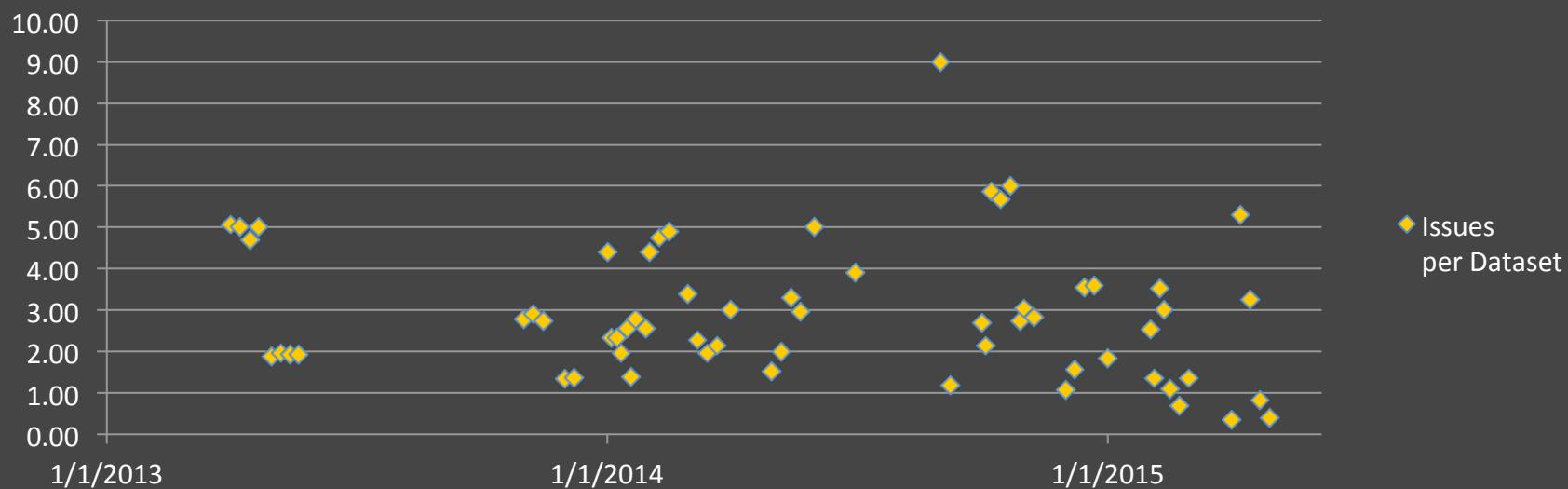
Issues per Study

- › Size of report files in CSV format
 - › 20 KB to 3.6 GB (x180,000)
- › Issues (data points)
 - › 215 to 24,000,000 (x110,000)
 - › Median=331K, Mean=1.86M, StdDev=4.5M, 25%=32K, 75%=1.1M
- › Unique issues
 - › 4 to 134
 - › Median=54, Mean=58, StdDev=34, 25%=30, 75%=81

Issues in Study



Issues per Dataset



False-Positive Messages

- › OpenCDISC validation is limited to ADISL, BDS, BDS-TTE and ADAE
- › Non-BDS domains are not recognized and produce false-positive messages
 - › “Unrecognized domain”
 - › “Required variable is not present: PARAM, PARAMCD”
 - › “Domain referenced in define.xml but dataset is missing”
- › Working on possible solutions to validate non-BDS datasets in future releases

Calculation issues

- › $\text{CHG} \neq \text{AVAL} - \text{BASE}$
- › 74% studies

BASE	AVAL	CHG	Calculated	Exact()	Comment
146	175	146			Obvious errors
1.0485	1.121	0.0725	0.0725	TRUE	False-Positive
36.4444444	36.555556	0.11111111	0.11111112	FALSE	Accuracy issue - tool
0.21	0.24	0.04	0.03	FALSE	Accuracy issue - user
1.09786	1.16244	0.06458	0.06458	FALSE	?

- › $\text{PCHG} \neq (\text{AVAL} - \text{BASE}) / \text{BASE} * 100$
 - › 44% studies
- › $\text{BASE} = 0$ but PCHG is populated
 - › 11% studies

Most common issues

- › Variable label mismatch between dataset and ADaM standard
 - › 79% studies
- › Required variable is not present
 - › 74% studies
 - › TRTP – in 87 datasets across all studies
 - › STUDYID – 44
 - › AESEQ – 6
 - › AESER – 5
 - › AGEU – 3

- › DTYPE value not found in 'Derivation Type' extensible codelist
 - › 71% studies
 - › 41 terms total
 - › WORST – 17
 - › SUMMARY – 11
 - › DERIVED – 10
 - › COPY – 9
 - › IMPUTED – 9
 - › NON-RESPONDER IMPUTATION – 8
 - › LPTCF – 7
 - › SUM – 7
 - › IMPUTED 0 – 7

- › CNSR is present but not all of STARTDT, ADT and ADTM are present
 - › 48% of all studies (not all studies have TTE data)
 - › New check for TTE data
- › Inconsistent value for AVALC
 - › 48% studies
 - › Many false-positive messages
 - › Accuracy issue
 - › 6.2 vs. 6.19865709
 - › Different presentation of AVALC
 - › 15.0 vs. 15
 - › 0 vs. null

Baseline Issues

- › Multiple baseline records exist for a unique USUBJID,PARAMCD,BASETYPE – 31% studies
- › BASE is present but ABLFL is not present – 29%
- › ABLFL = Y, but BASE != AVAL – 23%
- › Inconsistent value for BASEC – 23%
- › BASE or BASEC is populated for a unique USUBJID,PARAMCD but No baseline record exists – 11%
- › Inconsistent value for BASE – 3%
- › BTOXGR is present but ABLFL is not present – 18%

Metadata Issues

- › 81% studies have define.xml v1.0 which cannot support ADaM metadata
- › Codelist mismatched – 65% studies
- › Define.xml/dataset variable type mismatch – 27%
 - › ATM – 55
 - › SRMDT – 53
 - › ADT – 51
 - › RFICDT – 39
- › Variable in dataset is not present in define.xml – 29%
- › Variable in define.xml is not present in the dataset – 6%
- › Dataset is not present in define.xml – 2%

Category Variables Issues

- › Inconsistent value for AVALCAT1 – 29% studies
- › Inconsistent value for PARCAT1 within a unique PARAMCD – 29%
- › Inconsistent value for PARCAT2 within a unique PARAMCD – 5%
- › Inconsistent value for CHGCAT1 – 3%
- › Inconsistent value for BASECAT2 – 2%
- › Inconsistent value for PCHGCAT1 – 2%

Flag variable coding

- › TRTEMFL flag value is not Y or null – 26% studies
- › ANL02FL value is not Y or null – 8%
- › ANL01FL value is not Y or null – 3%
- › ABLFL value is not Y or null – 2%
- › FUPFL flag value is not Y or null – 2%

Illegal Variable Name

- › y is not in [1-9] for (R2)AyLO – 24% studies
- › zz is not in [01-99] for ANLzzFL/FN – 21%
- › y is not in [1-9] for (R2)AyHI – 15%
- › y is not in [1-9] for PARCATy(N) – 15%
- › y is not in [1-9] for CHGCATy – 13%
- › y is not in [1-9] for CRITy(FL/FN) – 13%
- › zz is not in [01-99] for AOCCzzFL – 8%
- › y is not in [1-9] for AVALCATy – 6%
- › xx is not in [01-99] for TRTxxA – 5%
- › xx is not in [01-99] for TRTxxAN – 5%
- › Illegal PARAMCD value – 8%

Inconsistent Value

- › Inconsistent value for AVAL – 27% studies
- › Inconsistent value for PARAM within a unique PARAMCD – 24%
- › Inconsistent value for BASEC – 23%
- › Inconsistent value for PARAMTYP – 16%
- › Inconsistent value for PARAM – 15%
- › Inconsistent value for PARAMN – 10%
- › Inconsistent value for PARAMCD within a unique PARAM – 6%
- › Inconsistent value for ATPT – 5%
- › Inconsistent value for PARCAT2 within a unique PARAMCD – 5%

Is not Numeric Variable

- › *TM is not a numeric variable – 23% studies
- › *DTM is not a numeric variable – 18%
- › *DT is not a numeric variable – 8%
- › Usually due to incorrect usage of variables
 - › AESTENDT="2015-01-01:2015-04-30" (text)

Traceability to SDTM DM data

- › For the same USUBJID, the ADSL.RACE does not equal DM.RACE – 16% studies
- › For the same USUBJID, the ADSL.AGEU does not equal DM.AGEU – 15%
- › For the same USUBJID, the ADSL.AGE does not equal DM.AGE – 11%
- › For the same USUBJID, the ADSL.ARM does not equal DM.ARM – 8%
- › For the same USUBJID, the ADSL.SUBJID does not equal DM.SUBJID – 2%

Traceability to SDTM data

- › ADaM ADAE record key is not traceable to SDTM.AE – 15% studies
- › SDTM.EX is present but neither ADSL TRTEDT nor TRTEDTM are present – 10%
- › SDTM.EX is present but neither ADSL TRTSDT nor TRTSDTM are present – 10%
- › USUBJID value does not exist in the SDTM DM domain – 10%
 - › Integrated data. E.g., DB + OL-ext studies
 - › Wrong study

Other Issues

- › Subject is off treatment (ONTRTFL), but analysis date (ADT) is within treatment period ($TRTSDT \leq ADT$) – 19% studies
- › Subject is on treatment (ONTRTFL), but analysis date (ADT) is after treatment end date (TRTEDT) – 18%
- › $ADY = 0$ – 10%
- › BDS.APERIOD xx does not have a corresponding ADSL.TRxxEDT variable – 15%
- › Secondary variable is present but its primary variable is not present – 13%
- › APHASE is present but APERIOD is not present – 11%

Questions

Sergiy Sirichenko

ssirichenko@pinnacle21.net

Michael Digiantomasso

mike.digian@pinnacle21.net