



Preparing Data for Submission to the FDA Its More Than SDTM

Bruce W Thompson, PhD, Michael Rippin, Ph.D., Glenn Daughaday M.B.A., Justin Steele, M.B.A. - Clinical Trials and Surveys Corp (C-TASC), Owings Mills, MD;

Abstract

NDA data submissions to the FDA begin with a data check using the OpenCDISC validation tool. The data are further reviewed by FDA staff during a 45-day interval to determine whether the NDA will be permitted to proceed. During this period of time, FDA staff will perform additional levels of data quality checking to ensure that the review will proceed quickly and efficiently. A set of procedures used by the FDA to identify problems of data consistency and completeness can be broken down into five major categories:

Content Analysis:

Evaluation of supplemental and custom domains for important content and determination of whether the supplemental or custom data can be fit into a standard domain

Fuzzy Matching:

Validation that original text has been appropriately transformed to standard values

Missing Data:

Verification that missing data are not treatment-dependant

Statistical Consistency:

Assessment of data consistency using various statistical methods

Event Reconciliation:

Confirmation that time-related events occur sequentially

In this poster we provide an overview on how these reviews can be performed by the sponsor prior to submission.

Introduction

With the advent of submitting data to the FDA in CDISC SDTM format, it has become possible to develop large programs to check the data as they arrive or are being prepared to be delivered.

A current starting point is to use the OpenCDISC checker program to determine the type and severity of errors according to the data that are to be submitted in each domain. Sample output of the SDTM checker are provided below:

Issue Summary			
Rule ID	Message		Found
Error			
SD0002	Null value in variable marked as Required	2	
SD0028	Upper limit must be greater than or equal to lower limit	12	
SD0036	Missing character result when original result provided	131	
SD0073	Referenced Domain not found	144	
Total		289	
Warning			
CT0037	Value for AEBODSYS not found in SOC controlled terminology codelist	2218	
CT0037	Value for MHBODSYS not found in SOC controlled terminology codelist	7529	
SD0006	No baseline result in LB for subject	4	
SD0009	AE is Serious but no qualifiers set to "Y"	60	
SD0026	Missing units on value	1	
SD0027	Missing value although units provided	8	
SD0031	Start date expected when end date provided	1	
SD0058	Variable appears in dataset but is not in SDTM standard	2	
SD0061	Domain referenced in define.xml but dataset is missing	17	
SD0063	SDTM/dataset variable label mismatch	3	

Questions

Using the statistical nature of the domains, it is possible to address some fundamental questions about each domain according to the data that are supposed to be collected:

For instance:

- Are “extra data” appropriately located?
- Is there statistical reliability associated with the entry of information in the xxdecod entry and the text entry of the actual values recorded by the study staff?
- Is there temporal consistency with reporting the start and stop times of exposures and onsets and endings of Adverse Events?
- Is there any evidence that data might not be missing at random?

Write a Series of Oracle and SAS Marcos To Check Incoming SDTM Data

Content Analysis

- SAS Macro Custom Domain
- Table Cat by TRT by Decod by Occur

List-Oriented Frequency Distributions				
Glenn Test 02/28/2012				
Glenn Test 02/28/2012				
Total Number of Subjects: 260				
XXCAT	XXTRT	XXDECOD	XXOCCUR	Record Count
INDIVIDUAL DRUG COUNSELING	INDIVIDUAL DRUG COUNSELING		N	11
			Y	2814

Fuzzy Matching

- Oracle Procedure
- Compare Text to IG list (e.g., MedDRA)

SDTM Domain: AE			
Randomly Sampled Term Matching			
PQ Test 9 120411			
PQ test study 9 120411			
Random Sample %: 100 Low Match Score: 100			
Total Terms: 1168 Total Matches: 1145			
MedDRA Version: 13.1			
AEDECOD	PT_NAME	Average Match Score	
Venous stasis	Venous stasis retinopathy	88.571	
Residual urine	Residual urine volume	88.571	
Hypoxic encephalopathy	Hypoxic-ischaemic encephalopathy	88.714	
Cardiac pacemaker malfunction	Cardiac pacemaker replacement, Cardiac pacemaker insertion, Cardiac pacemaker removal, Cardiac pacemaker evaluation, Cardiac pacemaker revision	80.952	
Anorexia	Anorexia nervosa	80.952	
Syncope vasovagal	Syncope	80.952	
Embolism	Embolism venous, Embolism arterial	80.952	
Postoperative constipation	Post procedural constipation	77.143	
Pacemaker complication	Pacemaker complication	77.143	
Dysphasia	Dysphagia, Disacteriosis	75	
Postnasal drip	Pseudomonas infection, Pseudomyxoma peritonei, Post-anoxic myelodysplasia, Postinfection angina, Pseudoendophthalmitis, Pseudomonal sepsis, Pseudomononucleosis, Postimplantation syndrome, Pseudomonas bronchitis, Postmature baby	75	

Missing Data

- SAS Survival Analysis Macro
- Make Withdrawal Code the “event” and the Event Code the censoring variable

Survival Analysis: Time To Dropout									
PQ Test 9 120411									
PQ test study 9 120411									
ARM XXXXXXX		Placebo		ARM XXXXXXX vs. Placebo					
Reason*	Dropout (Event)	Dropout (Censored)	Dropout (Event)	Dropout (Censored)	Hazard Ratio	95% CI Lower Bound **	95% CI Upper Bound **	P-Value	P-Value (LR)
ADMINISTRATIVE DECISION	2	732	4	730	0.49	0.07	2.49	0.4045	0.3901
ADVERSE EVENT	29	705	24	710	1.17	0.68	2.01	0.5751	
CONSENT WITHDRAWN	124	610	140	594	0.85	0.67	1.09	0.2003	
DEATH	43	691	43	691	0.98	0.64	1.50	0.9230	
DISEASE PROGRESSION	23	711	21	713	1.05	0.58	1.91	0.8595	
INELIGIBILITY DETERMINED	10	724	8	726	1.25	0.49	3.16	0.6432	
LOST TO FOLLOW-UP	16	718	21	713	0.74	0.39	1.42	0.3645	

* The sponsor provided standardized disposition terms include: ADMINISTRATIVE DECISION (6), ADVERSE EVENT(63), CONSENT WITHDRAWN(254), DEATH(8), DISEASE PROGRESSION (44), INELIGIBILITY DETERMINED (18), LOST TO FOLLOW-UP(37), NONCOMPLIANCE(11), OTHER(13), REQUIREMENT FOR ALTERNATIVE THERAPY (18), SUBJECT REQUEST (3)

**Upper bound values of 1 represent a value of 'infinity'

Statistical Consistency

SAS Macro to Review number of times on and off study meds.

Study Drug On/Off Frequency									
Study Multiple Arm Study									
Reference ARMCD: XXX									
Compared Arms	ARM	Total # of Events	Total # of Censored	Hazard Ratio	95% CI Upper Bound	95% CI Lower Bound	P-Value (Chisq)	P-Value (LR)	
YYY vs. XXX	YYY	1	41	0.545	6.00	0.05	0.6197	0.6101	
YYY vs. XXX	XXX	2	46	---	---	---	---	---	
ZZZ vs. XXX	ZZZ	2	44	1.073	7.22	0.16	0.9419	0.9458	
ZZZ vs. XXX	XXX	2	46	---	---	---	---	---	
AAA vs. XXX	AAA	0	53	0.000	0.00	0.00	<.0001	0.0752	
AAA vs. XXX	XXX	2	46	---	---	---	---	---	
BBB vs. XXX	BBB	1	46	0.450	4.97	0.04	0.5148	0.5009	
BBB vs. XXX	XXX	2	46	---	---	---	---	---	
CCC vs. XXX	CCC	2	40	1.185	7.72	0.18	0.8594	0.8648	
CCC vs. XXX	XXX	2	46	---	---	---	---	---	
DDD vs. XXX	DDD	1	47	0.439	4.88	0.04	0.5019	0.4871	
DDD vs. XXX	XXX	2	46	---	---	---	---	---	
EEE vs. XXX	EEE	0	43	0.000	0.00	0.00	<.0001	0.1009	
EEE vs. XXX	XXX	2	46	---	---	---	---	---	
FFF vs. XXX	FFF	49	46	23.439	126.89	4.34	0.0002	<.0001	
FFF vs. XXX	XXX	2	46	---	---	---	---	---	
All Other Arms vs. XXX	All other Arms	58	390	3.536	18.03	0.89	0.1288	0.0308	
All Other Arms vs. XXX	XXX	2	46	---	---	---	---	---	
ARMCD	EXADU	Count							
ZZZ	DOSE ADMINISTRATION ERROR	1							
	ADVERSE EVENT	2							
666	ADVERSE EVENT	1							

Event Reconciliation

- An event must end before another event of the same type can start.
- Oracle Procedure: count number of times two AE intervals of the same type overlap.

AE Reconciliation - Details									
USUBID	ATTEM	AESCCOD	AESTPTC	AESMDC	AESNMF	MYSTPTC	MYMDC	MYMDC	MYMDC
XXXXXXXXXXXX	SHORTNESS OF BREATH	Dyspnea	2005-05-01	2006-09-29		2005-02-02T00:00	2005-09-29		
XXXXXXXXXXXX	SHORTNESS OF BREATH	Dyspnea	2006-09-27	2006-09-29		2005-02-02T00:00	2005-09-29		
XXXXXXXXXXXX	LOW HEMAGLOBIN	Hemoglobin decreased	2005-09-23	2005-11-13		2004-11-19T00:00	2007-12-12		
XXXXXXXXXXXX	LOW HEMAGLOBIN	Hemoglobin decreased	2005-05-23		ATTEP	2004-11-19T00:00	2007-12-12		
XXXXXXXXXXXX	LIGHT HEADS	Dizziness	2005-12-07	2005-12-11		2004-12-19T00:00	2006-11-28		
XXXXXXXXXXXX	LOW HEMAGLOBIN	Dizziness	2005-12-07	2005-12-11		2004-12-19T00:00	2006-11-28		
XXXXXXXXXXXX	PAIN (R) FINGERS (REPETITIVE INJURY ARTHROSIS)	Arthralgia	2006-05-23	2006-05-28		2004-12-12T00:00	2007-12-18		
XXXXXXXXXXXX	PAIN (R) INDEX FINGER (REPETITIVE INJURY ARTHROSIS)	Arthralgia	2006-05-28		ATTEP	2004-12-12T00:00	2007-12-18		
XXXXXXXXXXXX	PAIN (R) ELBOW (REPETITIVE INJURY ARTHROSIS)	Arthralgia	2006-05-29	2006-06-03		2004-12-12T00:00	2007-12-18		
XXXXXXXXXXXX	PAIN (R) SHOULDER (REPETITIVE INJURY ARTHROSIS)	Arthralgia	2006-05-29	2006-06-03		2004-12-12T00:00	2007-12-18		
XXXXXXXXXXXX	BIATERAL HIP PAIN	Arthralgia	2007-05-10	2007-07		2004-12-12T00:00	2007-12-18		
XXXXXXXXXXXX	SPINOAL CYST	Synovial cyst	2007-05-05	2007-07-18		2005-04-02T00:00	2008-04-02		
XXXXXXXXXXXX	LOW BACK PAIN DUE TO SPINOAL CYST	Synovial cyst	2007-05-05		ATTEP	2005-04-02T00:00	2008-04-02		

Conclusions

The standardized format and meanings generated by CDISC SDTM allow the sponsor to generate sophisticated programs to review data before it is submitted to the FDA. The ability to generate and present these types of analyses will allow FDA and the sponsor to effectively communicate problem findings which may reduce the review time.

Acknowledgements

This work was supported by the Food and Drug Administration Contract Number HHSF2232008500241