Paper DS07

Common Data Quality Issues Seen During FDA's JumpStart Service

Aaron Belowich, IBM, Silver Spring, MD, USA
Andrea Arbelaez, U.S. Food and Drug Administration, Silver Spring, MD, USA

ABSTRACT

JumpStart is a FDA Center for Drug Evaluation and Research (CDER) service which provides data quality and data analytic support to reviewers. Since 2013, the JumpStart team has assessed the data quality of Study Data Tabulation Model (SDTM) data for over 170 New Drug Applications (NDA) and Biologics License Applications (BLA) using automated tools. The JumpStart data quality assessment reviews the composition and quality of clinical trial data to identify potential issues and errors that may affect a review. As part of the service, the JumpStart team investigates these findings and helps determine their potential impacts for review teams. This paper discusses some of the common findings from the JumpStart investigations that most impacted application reviews. We share these findings to inform applicants of the identified data quality issues to encourage them to submit higher quality data in the future, thereby increasing the efficiency of the review.

INTRODUCTION

The JumpStart service provides CDER review teams with an assessment of the composition, conformance, and quality of the clinical trial data submitted within an application. Using automated tools and additional investigation, the JumpStart team reviews errors, issues, and warnings that may impact the review. JumpStart empowers reviewers to make more data-driven decisions on data quality and readiness for analysis. Some issues warrant further scrutiny and/or may hinder a reviewer's ability to fully analyze the safety and efficacy of the investigational product. Ambiguity in the data, coding discrepancies, or faulty metadata may require clarification from applicants, resulting in delays. This paper will highlight nine common, impactful issues that the JumpStart team has identified during its data fitness assessment.

IMPLEMENTATION OF DEATH FLAG AND DEATH INFORMATION

An important safety evaluation of NDA and BLA applications is a thorough review of death related data from clinical trials. In order to perform this analysis, FDA reviews death information across several domains, including Demographics (DM), Death Details (DD), Adverse Events (AE), and Disposition (DS). Information related to a subjects' death may be located in other domains, however it is expected and most commonly located in DM, DD, AE and DS. Inconsistent and/or missing death information may prevent the review team from a complete safety evaluation.

Consider the following records from *Table 1* in the DM, AE, and DS domains:

Table 1: Inconsistent Death Information

DOMAIN	USUBJID	RFSTDTC	RFENDTC	DTHFL	DTHDTC
DM	XYZ-123	2016-08-12	2016-11-11		
DM	ABC-456	2017-03-12	2017-07-06	Y	2017-07-06

DOMAIN	USUBJID	AEDECOD	AESER	AESDTH	AESTDTC
AE	XYZ-123	Stroke	Y	Y	2016-12-01
AE	ABC-456	Heart Failure	Y	Y	2017-07-06

DOMAIN	USUBJID	DSCAT	DSDECOD	EPOCH	DSSTDTC
DS	XYZ-123	Disposition Event	Death	Follow-Up	2016-12-05
DS	ABC-456	Disposition Event	Death	Treatment	2017-07-06

In the example above, subject XYZ-123 appears to have a stroke which results in death (AESDTH = 'Y') on 2016-12-01, but the disposition domain indicates the subject died on 2016-12-05. Discrepancy between AESTDTC and DSSTDTC for XYZ-123 can make it difficult to ascertain the causality of death. Using the Death Flag (DTHFL) to review subjects who died would exclude subject XYZ-123. Subject ABC-456 has consistent death information across all three domains.

ALIGNING SUBJECT REFERENCE DATES WITH EXPOSURE INFORMATION

The Demographics (DM) domain contains important variables to indicate when subjects began and ended treatment. *Table 2* below lists these variables and their descriptions, as defined by SDTM IG 3.3:

Table 2: Start and End Date Derivations

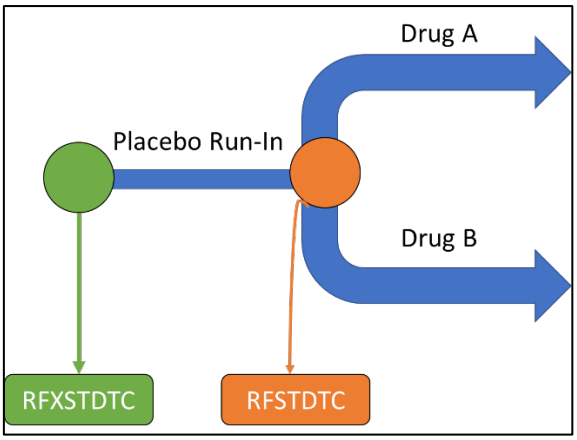
Variable Name	Variable Label	Description
RFSTDTC	Subject Reference Start Date/Time	Reference start date or start date and time for the subject. Usually equivalent to date/time when subject was first exposed to study treatment. Required for all randomized subjects; will be null for all subjects who did not meet the milestone the date requires, such as screen failures or unassigned subjects.
RFENDTC	Subject Reference End Date/Time	Reference end date/time for the subject. Usually equivalent to the date/time when subject was determined to have ended the trial, and often equivalent to date/time of last exposure to study treatment. Required for all randomized subjects; null for screen failures or unassigned subjects.
RFXSTDTC	Date/Time of First Study Treatment	First date/time of exposure to any protocol-specified treatment or therapy for the subject.
RFXENDTC	Date/Time of Last Study Treatment	Last date/time of exposure to any protocol-specified treatment or therapy for the subject.

The important distinction between the two "Start" variables (RFSTDTC, RFXSTDTC) plays a critical role throughout the SDTM data package. The latter variable, Date/Time of First Study Treatment (RFXSTDTC) represents the earliest date/time, by subject, to any exposure captured in the Exposure (EX) domain. This may include treatments during the run-in period. The Subject Reference Start Date/Time (RFSTDTC) variable represents the date/time of the first exposure *to study treatment*. This

variable excludes records from a run-in period and represents Study Day 1. All study day variables (--DY) throughout the domains use RFSTDTC in its calculation.

Consider the following study design in *Figure 1* in which subjects enter a placebo run-in period before being randomized to either Treatment Drug A or Treatment Drug B:

Figure 1: Differentiating RFXSTDTC and RFSTDTC



Likewise, for the "End" variables (RFENDTC, RFXENDTC), the Date/Time of Last Study Treatment (RFXENDTC) represents the subject's last date to any exposure in the EX domain. The Reference End Date/Time (RFENDTC) represents the subject's last date to and protocol-specified treatment in the EX domain. Often, RFXENDTC and RFENDTC are identical. Per SDTM guidance[1], RFSTDTC and RFENDTC are provided for all randomized subjects and remain null for unassigned subjects and screen failures.

INCLUSION OF ALL SCREEN FAILURES IN THE INCLUSION/EXCLUSION (IE) DOMAIN

As part of its standard review process, FDA assesses potential bias in patient selection by analyzing screen failures. Review teams evaluate whether screen failures have valid reasons for being excluded from the trial. This information is captured in the Inclusion/Exclusion domain (IE). The inclusion and exclusion criteria are stated in the study protocol .

Consider the examples below in *Table 3* from the Inclusion/Exclusion (IE) domain:

Table 3: Examples from the Inclusion/Exclusion Domain

USUBJID	IETESTCD[A]	IETEST[B]	IECAT[C]	IEORRES[D]
SMF-629	IN02	Demonstrated ability to properly operate the eDiary	INCLUSION	N
DFJ-918	EX05	Participation in another investigational trial during the 30 days prior to the Run-in Period or during this study	EXCLUSION	Y

[A] The IETESTCD variable is the code pertaining to the specific inclusion/exclusion criteria as described in sponsor documentation
[B] The IETEST variable is the specific inclusion/exclusion criteria as described in sponsor documentation
[C] The IECAT variable indicates whether the criteria in IETEST is an "Inclusion" or "Exclusion" criteria
[D] The IEORRES variable indicates whether the subject met the criteria ("Y") or did not meet the criteria ("N") as written in the IETEST variable

In the example above, subject SMF-629 did NOT meet the required inclusion criteria IN02 and subject DFJ-918 did meet exclusion criteria EX05; as such, both subjects are considered "Screen Failures" in the Demographics (DM) domain.

MISSING REFERENCE RANGE IN LABORATORY (LB) DOMAIN

An important laboratory safety analysis is the review of abnormal test results. Review teams evaluate lab data to check if any elevated readings may be related to administration of the study treatment. In order to assess if a lab reading is elevated, the review team compares the lab result (LBSTRESN) against it's upper limit of normal (LBSTNRHI). When LBSTRESN is greater than LBSTNRHI, the reading is "High" and may warrant further investigation.

Consider the following set of ALT records in *Table 4* from the Laboratory (LB) domain:

Table 4: Missing Reference Ranges in LB

DOMAIN	USUBJID	LBSEQ	LBTESTCD	LBSTRESN	LBSTNRLO	LBSTNRHI	LBNRIND
LB	ABC-123	1	ALT	42	30	65	Normal
LB	ABC-123	2	ALT	84	30	65	High
LB	ABC-456	1	ALT	71			
LB	ABC-456	2	ALT	36			

The first two records have reference ranges (LBSTNRLO, LBSTNRHI) present. As a result, the second record is identified in the Reference Range Indicate (LBNRIND) as "High", as the result (84) is greater than the upper limit of normal (65). Results within the reference range, as in the first example, have "Normal" values in LBNRIND.

The second set of records, for subject ABC-456, are missing reference range. This prevents the review team from assessing whether the results are normal, high, or low. Some reference ranges may vary from test to test and subject to subject, so it is not possible to extrapolate the reference range from one subject to another. Missing reference ranges prevent a complete and thorough analyses of analyzing a subject's laboratory results.

GUIDANCE FOR IMPLEMENTING REFERENCE PARTICIPATION END DATE (RFPENDTC)

According to the SDTM Implementation Guide[1], the Reference Participation End Date (RFPENDTC) should be derived as the "Date/time when subject ended participation or follow-up in a trial, as defined in the protocol, in ISO 8601 character format. Should correspond to the last known date of contact." Essentially, this variable is intended to serve as a catch-all; RFPENDTC is populated by searching all SDTM datasets and finding the last date to correspond with this subject. In a shortened example, consider the following records in *Table 5* from the Adverse Events (AE), Disposition (DS), Exposure (EX), and Subject Visits (SV) domains:

Table 5: Scanning SDTM Data for Last Date

DOMAIN	USUBJID	AEDECOD	AESER	AESTDTC	AEENDTC
AE	XYZ-654	Influenza	Y	2015-04-13	2015-04-22

DOMAIN	USUBJID	DSCAT	DSDECOD	EPOCH	DSSTDTC
DS	XYZ-654	Disposition Event	COMPLETED	Follow-Up	2015-12-05

DOMAIN	USUBJID	EXTRT	EXDOSE	EXSTDTC	EXENDTC
EX	XYZ-654	Drug A	100	2015-02-10	2015-12-01

DOMAIN	USUBJID	VISIT	EPOCH	SVSTDTC	SVENDTC
SV	XYZ-654	Visit 10	Follow-Up	2015-12-15	2015-12-15

In this example, the last date of contact for Subject XYZ-654 occurs at Visit 10 in the SV domain (2015-12-15). Therefore, the Reference Participation End Date (RFPENDTC) is populated in the Demographics (DM) domain accordingly, as seen in *Table 6*:

Table 6: Populating RFPENDTC

DOMAIN	USUBJID	RFSTDTC	RFENDTC	RFPENDTC	ACTARM
DM	XYZ-654	2015-02-10	2015-12-01	2015-12-25	Drug A

The RPENDTC may come from any domain, and different subjects may have RFPENDTC populated from different domains (One subject may have RFPENDTC populated from Concomitant Medications while another subject has RFPENDTC populated from Disposition).

GUIDANCE FOR IMPLEMENTING ACTARM

The Demographics (DM) domain contains two treatment arm variables and respective code variables – Planned Arm (ARM) and Actual Treatment Arm (ACTARM). Consider the following examples of randomized subjects in *Table 7* – assume only Subject XYZ-123 has exposure records:

Table 7: Incorrectly Populating ACTARM

USUBJID	RFSTDTC	RFENDTC	ARM	ACTARM
XYZ-123	2016-08-12	2016-11-11	Drug A 50mg	Drug A 50mg
XYZ-135			Drug B 100mg	Drug B 100mg

In this example, Section 4 of the SDRG said that Subject XYZ-135 was randomized and not treated. However, the presence of a study drug in the ACTARM ("Drug B 100mg") indicates the subject was, in fact, treated. In addition, all subjects with study drug(s) in the ACTARM variable should have treatment start and end dates populated. This discrepancy is confusing – perhaps the subject was treated but is missing exposure records. Using the ACTARM variable to create analyses would include this subject, when the subject should be excluded. To more easily differentiate treated and not treated subjects, not treated subjects can have null values for ACTARM to indicate they were, in fact, not treated. As seen in *Table 8,* use the ARMNRS (Reason Arm and/or Actual Arm is Null) to provide further clarification for why the subject was not treated*:*

Table 8: Using ARMNRS to Show Subjects Not Treated

USUBJID	RFSTDTC	RFENDTC	ARM	ACTARM	ARMNRS
XYZ-135			Drug B 100mg		Assigned, Not Treate

The ARMNRS can be populated with the values of "SCREEN FAILURE", "NOT ASSIGNED", "ASSIGNED, NOT TREATED", or "UNPLANNED TREATMENT".

MEDICAL HISTORY AFTER SUBJECT START DATE

Medical History (MH) records contain relevant prior events and medical information for individual subjects. Captured before the study, these records may help contextualize new events and observations during the study. Records occurring after the first dose of study treatment (RFSTDTC) are misplaced in the MH domain – placing records in MH may preclude the event from being considered "treatment emergent". *Table 9* contains two examples:

Table 9: MHSTDTC after RFSTDTC

DM			MH	
USUBJID	RFSTDTC	RFENDTC	MHDECOD	MHSTDTC
123123	2011-07-29	2012-08-06	Coronary Artery Disease	2012-05
456456	2010-10-30	2011-01-15	Incisional drainage	2010-12

ADVERSE EVENT OUTCOMES AND END DATES

Adverse event outcomes (AEOUTs) are assigned from an extensible code list with values such as "Recovered/Resolved", "Recovering/Resolving", "Not Recovered/Not Resolved", "Fatal", or "Unknown". The outcome of a given adverse event determines whether an End Date (AEENDTC) should be provided. Consider the following examples in *Table 10* from the AE domain:

Table 10: Missing AEENDTC

AEDECOD	AEOUT	AESTDTC	AEENDTC
Fatigue	NOT RECOVERED/NOT RESOLVED	2016-01-24	
Headache	RECOVERED/RESOLVED	2016-01-25	2016-01-25
Erythema	RECOVERED/RESOLVED	2016-03-21	
Nasal Congestion	NOT RECOVERED/NOT RESOLVED	2016-05-14	2016-05-16

It may be contradictory to say that an adverse event is not recovered but then to provide an end date, or to say an adverse event has resolved but not to provide an end date.

INCONSISTENT STANDARD UNITS IN LABORATORY (LB) DOMAIN

Consider the following records in *Table 11* of Albumin records from the Laboratory (LB) domain:

Table 11: Inconsistent LBSTRESU

DOMAIN	USUBJID	LBSEQ	LBTESTCD	LBSTRESN	LBSTRESU	VISIT
LB	ABC-123	1	ALB	4.7	g/dL	WEEK 4
LB	ABC-123	2	ALB	4.3	g/dL	WEEK 8
LB	ABC-456	1	ALB	38	g/L	WEEK 4
LB	ABC-456	2	ALB	47	g/L	WEEK 8

Sometimes submitted datasets might have both conventional and SI units in the LBSTRESN values as shown above. In this scenario, calculating a mean ALB value by visit without converting all subjects' lab results to the same unit would give an inaccurate finding.

CONCLUSION

Since its inception in 2013, the JumpStart team has uncovered hundreds of different issues with respect to SDTM conformance from sponsors. The JumpStart team helps reviewers understand the significance of these issues and how they may impact the review process. This paper presents nine of the common issues identified by the JumpStart team that sponsors may find useful as they continue their processes of developing SDTM datasets. Implementing good data practices to address these and other issues will help FDA with an efficient review process for NDA/BLA applications.

REFERENCES

[1]Study Data Tabulation Model Implementation Guide: Human Clinical Trials, Version 3.3, CDISC, 2018, https://www.cdisc.org/standards/foundational/sdtmig/sdtmig-v3-3. Accessed 30 January 2019.

ACKNOWLEDGMENTS

Jesse Anderson, U.S. Food and Drug Administration, Silver Spring, MD, USA
Dr. Jose Delgado Donayre, U.S. Food and Drug Administration, Silver Spring, MD, USA
Dr. Tejas Patel, U.S. Food and Drug Administration, Silver Spring, MD, USA
Dr. Lilliam Rosario, U.S. Food and Drug Administration, Silver Spring, MD, USA
Dr. Alan Shapiro, U.S. Food and Drug Administration, Silver Spring, MD, USA
Sarah Ferko, IBM, Silver Spring, MD, USA
David Jacobs, IBM, Silver Spring, MD, USA
Kathryn Matto, IBM, Silver Spring, MD, USA

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Andrea Arbelaez
US Food and Drug Administration
10903 New Hampshire Ave.
Silver Spring, MD 20904
Work Phone: 301-796-3770
Email: Andrea.Arbelaez@fda.hhs.gov