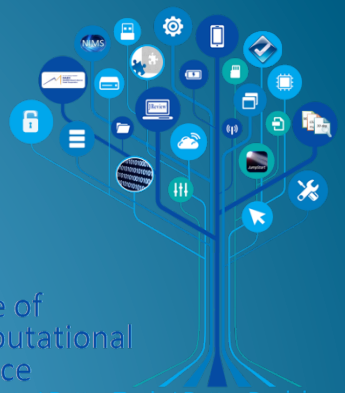
Common Errors in Loading SDTM Data to the Clinical Trials Repository

Why Getting it Right Matters



Crystal Allard, OCS
Special Assistant to the Director

Office of Computational Science
Better Data | Better Tools | Better Decisions

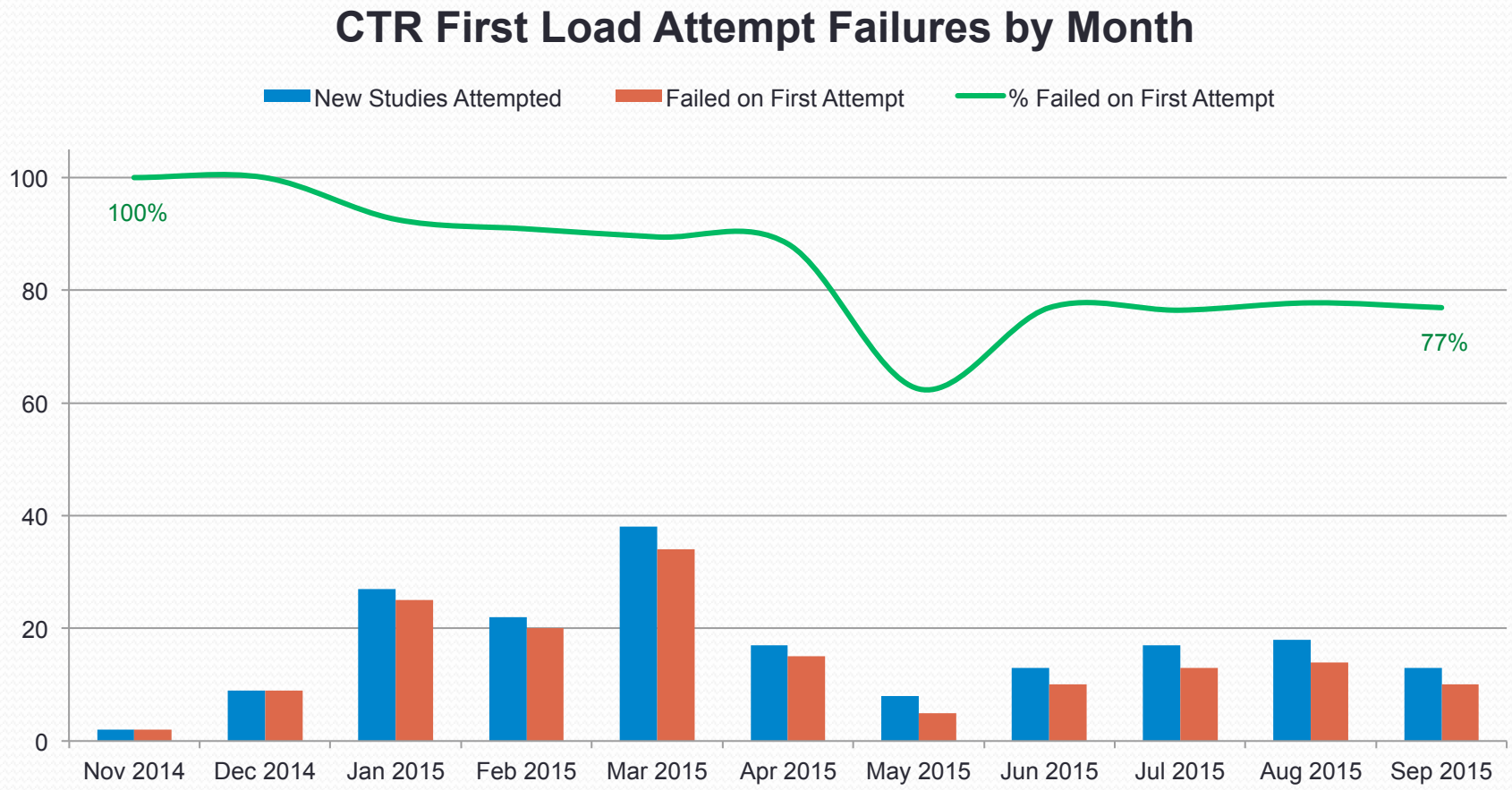
The Janus Clinical Trials Repository

- **Data warehouse** for **subject-level clinical trial data**
- Developed jointly with the National Cancer Institute(NCI).
- Operational at the FDA's White Oak Data Center since January 2015.
- **Conceptual model** publicly available
- Model informed by Biomedical Research Integrated Domain Group (**BRIDG**) Domain Model (see http://www.bridgmodel.org/).
- Designed to receive data in CDISC SDTM[1] format and can support other study data exchange standards, such as CDISC SEND, HL7 v3, FHIR,[2] and RDF.[3]
- Study data validation rules published on the Study Data Standards Resources page on fda.gov.
- Download the CTR Conceptual Model

Loading Data to CTR 2014 - 2015

- Loaded all data from applications that received a JumpStart service in 2014 and 2015
- In FY 2016, will load all SDTM data for applications associated with JS
- Eventually, all SDTM data received at FDA will be loaded to CTR

CTR Data Loading Stats 2014 - 2015



CTR Data Loading Stats 2014 - 2015

8 Studies (5%) did not load to Janus

- 4 studies (2.5%) are missing define files
- 4 studies (2.5%) have significant data issues and we cannot fix them without more information from sponsor/sponsor resubmission

86 Studies (55%) had major blocking issues

- 65 studies (42.7%) had issues within the Define files
- 76 studies (49%) had issues within the TS domain data

CTR Data Loading Stats 2014 - 2015

17 Studies (11%) misc. loading issues

- Issues within other domain files that prevented the initial load or software problems that prevented the initial load

28 studies (18%) missing TS domain files

- Had to create the TS domain file in order to load the studies

CTR Blocking Rules

Issue	Blocking Issue Category	Description
Duplicate trial element codes	Known Anomaly	ETCD values should be unique for different elements in Trial Elements (TE) domain. Duplicate values for ETCD (Element Code) in the Trial Elements domain. The SDTM IG Rule 15 for Trials Elements "Note that Elements that have different start and end rules are different Elements and must have different values of ELEMENT and ETCD.For instance, Elements that involve the same treatment but have different durations are different Elements. The same applies to non-treatment Elements. For instance, a washout with a fixed duration of 14 days is different from a washout that is to end after 7 days if drug cannot be detected in a blood sample, or after 14 days otherwise." This is checked by existing data fitness rules: FDAC218 and FDAC219 . This issue was encountered in a number of studies that were loaded in production.
Duplicate Trial Inclusion Exclusion Criteria	Known Anomaly	IETESTCD values should be unique within a category, sub category and version of Trial Inclusion Exclusion criteria in the Trial Inclusion Exclusion (TI) domain. There was 1 study where this issue was encountered.
Duplicate Trial Visits	Known Anomaly	Visit numbers should be unique within an arm in the Trial Visits (TV) domain. Rule 1 in the SDTMIG for the Trial Visits domain explains that.
Trial Summary duplicate records	Known Anomaly	Trial summary records should be uniquely identifiable using TSPARMCD and TSSEQ, if there are duplicate records for the combination of TSPARMCD and TSSEQ values then the study load will fail
Invalid class names for domains	Known Anomaly	Class names specified for domain are not correct. The actual observation class of a domain does not match its class as defined by the CDISC define standard. For the standard SDTM domains the load software takes the domain class as specified by the standard. For custom domains if the class cannot be correctly determined then study cannot be loaded. This issue happened with several studies, especially for standard SDTM domains as CTR load software was using define for domain class for all domains.
Duplicate STUDYID values	Known Anomaly	Multiple values for STUDYID variable in dataset files.
Trial summary Data Type Number	Data Type	TSVAL for this variable needs to be a number. This happens for a number of trial summary parameter codes such as PLANSUB, ACTSUB, RANDQT which are required to be a number.
Trial Summary Date time	Data Type	This happens for TSPARMCDs SSTDTC, SENDTC. TSVAL for this variable needs to be an ISO8601 date time value as follows: YYYY-MM-DD where [YYYY] four digit year, [MM] two digit month, [DD] two digit year.
Trial Summary Data Type number or ISO8601 duration	Data Type	This happens for TSPARMCDs such as AGEMIN and AGEMAX. TSVAL for this variable should be a number or ISO8601 duration as follows: PnYnMnDTnHnMnS or PnW where:[P] precedes the alphanumeric duration.[n] represents is a number >= 0 [W] is used as week designator (e.g., P6W represents 6 weeks of calendar time).
Trial Summary ISO8601 Duration	Data Type	This happens for TSPARMCDs such as LENGTH, SDMDUR, CRMDUR. TSVAL for this variable should be an ISO 8601 duration as follows: PnYnMnDTnHnMnS or PnW where:[P] precedes the alphanumeric duration.[n] represents is a number >= 0 [W] is used as week designator (e.g., P6W represents 6 weeks of calendar time).
Special Characters	Special Characters	Special Characters in the data sets that can result in errors during load

Examples of Define File Errors

Error	Description
StudyName element is same for two different studies under the same product application number	Two different studies have the same StudyName element in define.xml for the same product application number
Invalid XSL file	The XSL file included with the study data is invalid. This results in an error in creating the define html
Required attributes and elements missing in define.xml	Required attributes and elements are needed to be present in define.xml.
Datasets included in the study data folder but not described in define.xml	CTR uses define.xml to load study data. If a domain is not described in define.xml then it will not be loaded.
define.xml refers to a define.xsl file that does not exist	define.xml refers to a define.xsl file that does not exist in the study folder. CTR uses the referred define.xsl to render define.xml into html for display on the study summary application.
Path to the dataset files in define.xml does not match the dataset file name exactly	CTR uses the href attribute in each leaf element in the ItemGroupDef element of define.xml to locate the dataset file for a specific domain. If the href attribute does not match the dataset file name then CTR cannot load the study.

Examples of Trial Summary Errors

Error	Description
Treatments not associated correctly with the dose and route data	Trial summary values for treatment need to be associated with the dose and the route trial summary values. If there are multiple values for treatment and dose and route then TSGRPID values should be used to associate which treatment goes with which dose. If multiple treatments are included in a study and this association is not provided then load software assumes that each dosage group is associated with each dosage group.
Trial summary records are not unique based on TSPARMCD and TSSEQ	Trial summary records should be uniquely identifiable using TSPARMCD and TSSEQ, There cannot be duplicate records for the combination of TSPARMCD and TSSEQ values
Trial Summary Value (TSVAL) for TSPARMCD AGEMIN uses invalid format	AGEMIN is required to be in the ISO 8601 format. If a single age is listed (i.e., 18), the age should be in the format of P##Y, where the ## are the age noted. In this example, 18 should be P18Y.
Trial Summary parameter ACTSUB has invalid format	The value for param ACTSUB has to be a number or null
Trial Summary Value (TSVAL) for TSPARMCD AGEMAX uses invalid format	AGEMAX is required to be in the ISO 8601 format. Values such as UNLIMITED, NONE, No maximum age are invalid
Invalid format for TSPARMCD RANDQT	Randomization quotient needs to be a numerical value.
Trial summary parameters SSTDTC and SSENDTC are of invalid format	Trial summary parameters specifying date/time values need to be in proper date time format.

Examples Other Types of Errors

Error	Description
Duplicate records in observation domains	There are duplicate records with in any of the findings,events or interventions domains. Multiple records have the same sequence number value for the same USUBJID within a domain.
Special characters in data set files	There are special characters in the data set files
Invalid Trial Design domain	Valid Trial Design domains based on the SDTM IG are TS, TE,TV, TA and TI. Any other domain with a class of Trial Design will result in load errors
Domain names should be upper case	The name attribute in the ItemGroupDef elements in define.xml should be upper case.
Duplicate USUBJID in the demographics domain	Duplicate values for USUBJID in the demographic domain cannot be loaded. Each records in the DM domain is expected to identify a unique subject
Invalid date values	Date/time values in SDTM are required to follow the ISO 8601 date/time format. These invalid formats can result in invalid data in the CTR warehouse and cause some of the analysis panels to not work.
Special characters such as "-","&","/" in the Product application number	Special characters in the product application name can result in errors loading the study into CTR.

Major Benefits of Data Submissions Without Errors

- Enables auto-loading of data to CTR from

  EDR ->Portes ->CTR
- Minimization of information requests to sponsors
- Minimizes the number of versions of data
- Enables Automated Analyses of Data
- Many, Many More!

Coming Up for Janus CTR

- Use with JumpStart Service
- Training for reviewers
- Much Much More!

Questions?