Paper 3790-2019

Automate Process to Ensure Compliance with FDA Business Rules in SDTM Programming for FDA Submission

Xiangchen (Bob) Cui, Hao Guan, Min Chen, and Letan (Cleo) Lin, Alkermes Inc.

ABSTRACT

The U.S. Food and Drug Administration (FDA) has published FDA Business Rules and expects sponsors to submit SDTM data sets that are compliant with the rules and with CDISC SDTMIG. These rules assess whether the data supports regulatory review and analysis. Some of them are specific to FDA internal processes rather than to CDISC SDTM standards. Pinnacle 21 is the most commonly used tool by both the industry and FDA to check compliance with both FDA business rules and CDSIC rules. However, Pinnacle 21 is usually used at a late stage of the SDTM programming development cycle, and it cannot help users to resolve its findings regarding Error and Warming messages, even if it is used at the very early stage.

This paper presents a systematic approach to automate SDTM programming process to ensure compliance with FDA Business Rules. It contains **study data collection design, data collection (edit-checking), standard SDTM programming process, and in-house macros** for automatically reporting and fixing the issues to address non-compliance with FDA Business Rules. It avoids inefficient use of resources for repeated verification of the compliance and resolution of the findings from Pinnacle 21 for these rules. In fact, some of these non-compliant issues are often very costly or too late to be fixed at a late stage. This paper can assist readers to prepare SDTM data sets that are compliant with FDA business rules and with CDISC standards for FDA submission to ensure FDA submission quality, in addition to cost-effectiveness and efficiency.

INTRODUCTION

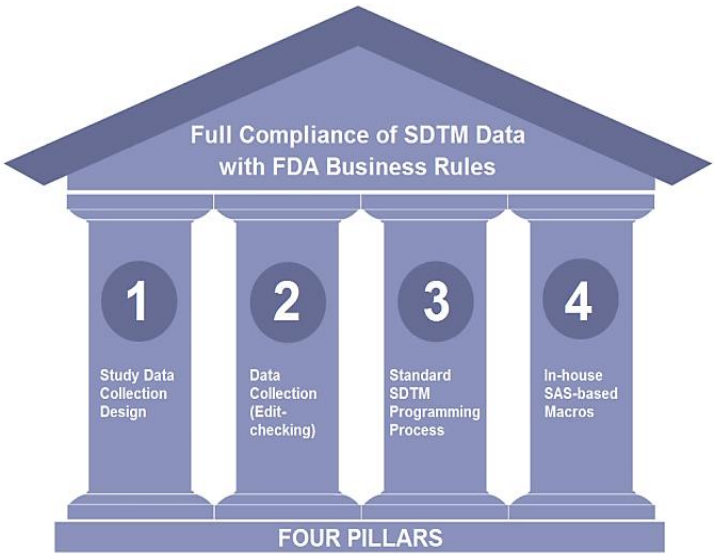
FDA published **FDA Business Rules** [1] in July 2017, December 2017, and October 2018, respectively, and **FDA Validator Rules** [1] in March 2017, December 2017, and October 2018, respectively. There are eighty-five (85) FDA business rules in total, which are categorized into Nonclinical with twenty-eight (28) rules, Clinical Only with twenty-one (21) rules, and Clinical and Nonclinical with thirty-six (36) rules. FDA expects these business rules to be followed where applicable. Per section 8 in **FDA Study Data Technical Conformance Guide** [3], "FDA business rules describe the business requirements for regulatory review to help ensure that study data is compliant and useful and supports meaningful review and analysis. The list of business rules will grow and change with experience and cross-center collaborations. All business rules should be followed where applicable. The business rules are accompanied with validator rules which provide detail regarding FDA's assessment of study data for purposes of review and analysis." Please refer to [1] for the Standards Web page providing links to the most updated business rules and FDA validator rules. These rules are used by the **FDA study data validator** to "ensure data are standards compliant and support meaningful review and analysis". Further, FDA expects that "Sponsors should evaluate their study data before submission against the conformance rules published by an SDO, the eCTD Technical Rejection Criteria for Study Data, and the FDA business rules." Note: SDO stands for standards development organization (CDISC).

FDA "Study Data Technical Conformance Guide" specifies three types of Study Data Validation Rules [3]:

1. Standards Development Organizations (e.g., CDISC) provide rules that assess conformance to its published standards (See www.CDISC.org).
2. FDA eCTD Technical Rejection Criteria for Study Data that assess conformance to the standards listed in the FDA Data Standards Catalog (See above).
3. FDA Business and Validator rules to assess that the data support regulatory review and analysis.

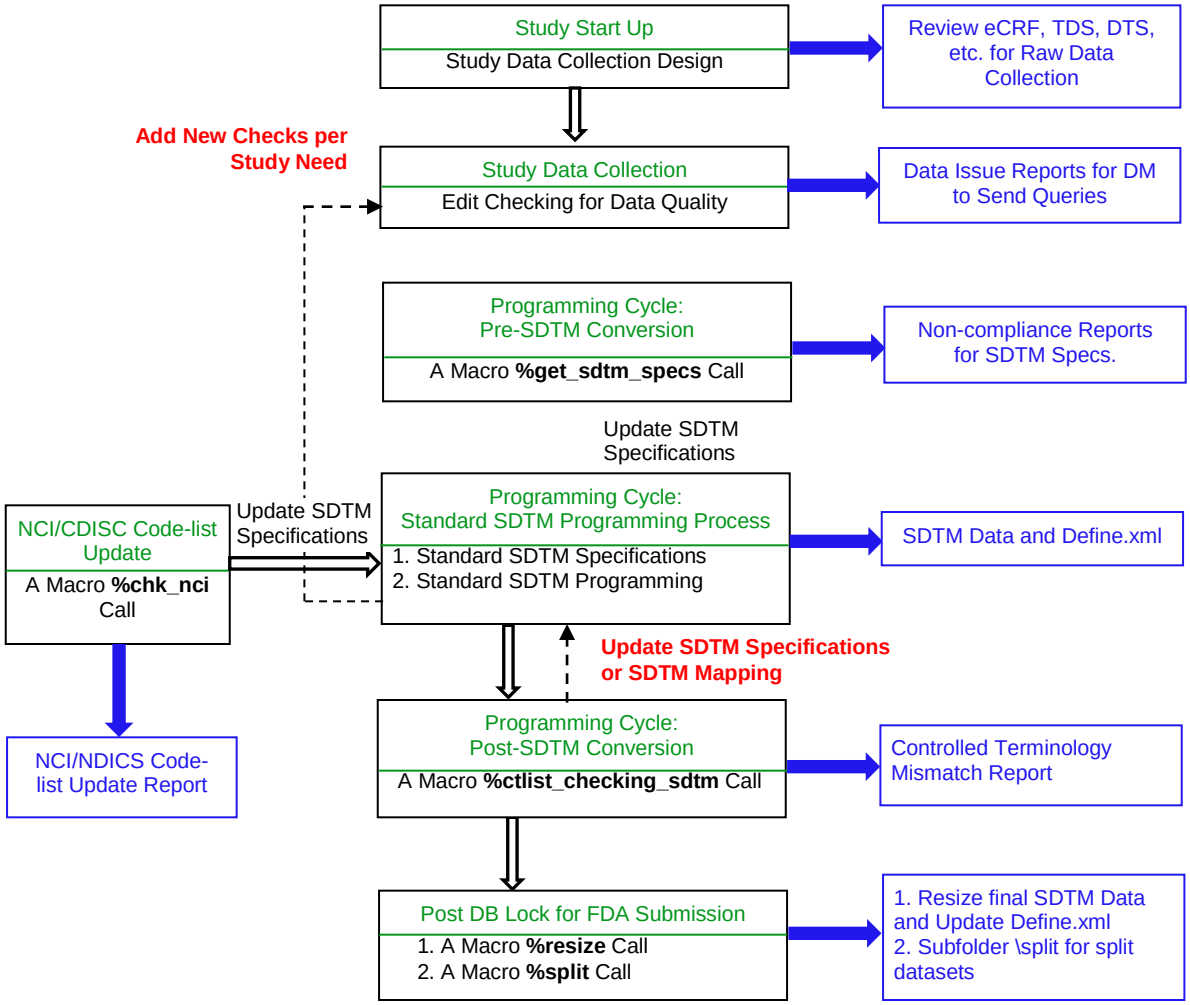
Pinnacle 21 is used by both sponsors and FDA to check compliance with both FDA business rules and CDSIC standards. It is a very useful diagnostic tool for detecting and reporting non-compliance issues of study data. Findings with "Error" and/or "Warming" messages, can be categorized as either data issues or SDTM mapping issues, and sponsors must either correct the data issues and/or explain discrepancies in the SDRG (Study Data Reviewer's Guide), or fix SDTM mapping errors before FDA submission. Hence it cannot help sponsors to automatically resolve the issues of data conformance, even if it is used at the very early stage of SDTM programming development. Furthermore, some of these non-compliant data issues are often very "costly" and/or too late to be fixed at a late stage.

Hence, a proactive approach is warranted for a high quality and cost-effective SDTM programming for preparing FDA submission. This paper presents a systematic approach to automate SDTM programming process to ensure compliance with FDA Business Rules. We focus on the rules of "**Clinical only**" and of "**Clinical and Nonclinical**" to show the reader how to automate the process. It is categorized into four groups based on the solutions to being compliant. They are **study data collection design, data collection (edit-checking), standard SDTM programming process, and in-house SAS-based macros**. Below shows **four pillars of the systematic approach**.



We will illustrate how the systematic approach can help sponsors achieve full compliance of SDTM data with FDA Business rules when it is applied to SDTM programming. We will further explain why this approach is far superior to Pinnacle 21 in ensuring compliance of SDTM data with FDA Business Rules.

Below is a flowchart depicting a systematic approach to automate SDTM programming process to ensure compliance with FDA Business Rules for FDA submission.



Display 1. Flowchart of SDTM Programming Process to Ensure Compliance with FDA Business Rules

STUDY DATA COLLECTION DESIGN

International Council on Harmonization (ICH) and United States Food and Drug Administration (USFDA) emphasize the principles and applications of quality by design (QbD) in pharmaceutical development for current good manufacturing practice (CGMP) regulations in their guidance for the industry. Sixteen (16) FDA Business Rules are identified and addressed in collecting data from a clinical study to support intended analysis among eighty-five (85) rules. Table 1 displays these rules which can be addressed by **Study Data Collection Design**.

Study Data Collection Design includes the development and finalization of eCRF, Data Management Plan (DMP), CRF Completion Guidelines (CCG), training documentation and communication plan of CCG to sites and Clinical Research Associates (CRAs/"monitors"), Trial Design Specifications (TDS) for EDC database build, external data transfer specification (DTS) for lab, ECG, etc.. It is critical to the quality of a clinical study and ensures correct implementation of standards and good practices to be followed in collecting data to support intended analysis during study conduct.

The high quality **Study Data Collection Design** minimizes the chance of EDC re-build and missing critical data collection points, and the number of queries generated for obvious

errors during the study. The inadequate **Study Data Collection Design** causes study delays, and/or study-cost surge, and most importantly jeopardizes both data quality and reliable study results.

The study data manager (DM) is responsible for the development and finalization of these documents, and should work with study team to develop and finalize the **Study Data Collection Design** during study start-up. The Statistical Programmer is one of the key team members for the completion of the Clinical Study Report (CSR), and should be involved in developing and finalizing **Study Data Collection Design** by reviewing and providing comments to DM through the understanding of the study protocol and the knowledge of FDA Business Rules, CDISC guidelines, and ADaM programming derivation for both efficacy and safety table programming. He/she should communicate these sixteen (16) FDA Business Rules shown in Table 1 with the study team to ensure these rules to be met during the study start-up, and timely identify and report any inadequacy from edit-checking and/or controlled terminology inconsistency report in SDTM programming during the study conduct.

In summary, **Study Data Collection Design** automatically helps the sponsors to meet the sixteen (16) FDA Business Rules shown in Table 1.

FDA Business Rule ID	FDA Business Rule
FDAB003	Adverse Events should be coded using MedDRA dictionary.
FDAB006	All death information should be populated for subjects that died during the study including any post treatment follow-up.
FDAB007	All deaths should be an independent row in the adverse event dataset.
FDAB010	All serious adverse events should be flagged.
FDAB028	Screen Failure subjects should have records in Inclusion/Exclusion dataset.
FDAB055	Trial participants should self-report race and ethnicity and they should not be assigned by the study team.
FDAB056	Participants are permitted to designate a multi-racial identity
FDAB057	When collecting ethnicity demographic data from clinical trial participants, the following two minimum choices should be offered: "HISPANIC OR LATINO" or "NOT HISPANIC OR LATINO"
FDAB058	The term "Spanish origin" can be used to collect ethnicity data. If the term "Spanish origin" is collected, it should be mapped to the controlled terminology "HISPANIC OR LATINO" for submission.
FDAB059	When collecting racial demographic data from clinical trial participants, the following five minimum choices should be offered: American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, and White
FDAB060	The term "Hatian or Negro" can be used collect race data. If the term "Hatian or Negro" is collected, it should be mapped to the controlled terminology "BLACK OR AFRICAN AMERICAN" for submission.
FDAB061	When collecting demographic information from clinical trial participants, more detailed/granular choices may be desired with regards to race and ethnicity (ex: clinical trial conducted outside the U.S). In these scenarios, sponsors should consult with the appropriate FDA review division and the FDA guidance document from October 2016 “Collection of Race and Ethnicity Data in Clinical Trials.”

FDA Business Rule ID	FDA Business Rule
FDAB062	When offering more granular/detailed race and/or ethnicity selection options to clinical trial participants, sponsors should ensure that these additional options roll up (or collapse) into the existing five primary race and two ethnicity categories as described in the FDA guidance document from October 2016 "Collection of Race and Ethnicity Data in Clinical Trials."
FDAB069	Drugs and metabolite names in pharmacokinetic datasets should be consistent with naming in other datasets across a submission.
FDAB070	Standardized units should be consistent for a given drug or metabolite across pharmacokinetic datasets within a submission.
FDAB071	A specimen material type should be named consistently across pharmacokinetic datasets within a submission.

Table 1. Sixteen (16) FDA Business Rules to Be Addressed by Study Data Collection Design

Note: in these sixteen (16) FDA Business Rules, except rules FDAB003, FDAB006, FDAB010, and FDAB028, the other twelve (12) rules do not have the corresponding FDA Validator Rules or Pinnacle 21 Validator Rules.

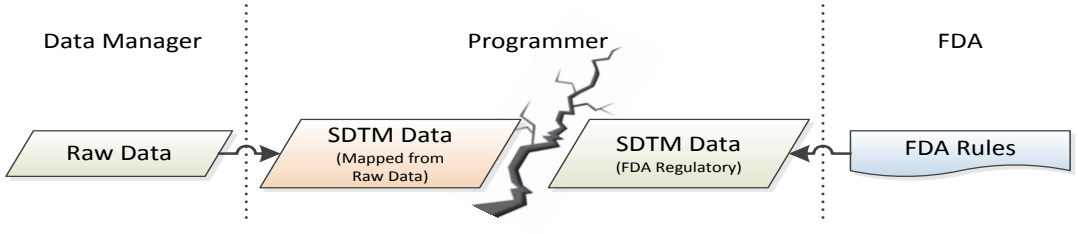
STUDY DATA COLLECTION (EDIT-CHECKING)

Data quality is referred to data integrity (accuracy, completeness, and reliable), which is an integral part of clinical study reports. This is why FDA Guideline for Industry "Oversight of Clinical Investigations —A Risk-Based Approach to Monitoring" [4] identifies it as one of two risks of a clinical study.

> "FDA encourages sponsors to develop monitoring plans that manage important risks to **human subjects** and **data quality** and address the challenges of oversight in part by taking advantage of the innovations in modern clinical trials."

Data Management data cleaning is very critical to achieve study data quality. SAS programming supports Data Management to clean data by providing edit-checking reports to identify "missing data, inconsistent data, data outliers, and potential protocol deviations that may be indicative of systemic or significant errors in data collection" [4].

FDA reviewers usually use SDTM and/or ADaM datasets for reviewing sponsor's application. Moreover, FDA business rules require the compliance of SDTM data, not raw data. SDTM programming is directly mapping raw data into SDTM format, along with very minimal derivation. Hence any raw data issues are directly "populated" into SDTM data if any exists. The plot below shows the relationship among Raw Data, SDTM Data, and FDA Business Rules.



Display 2. Relationship among Raw Data, SDTM Data, and FDA Business Rules

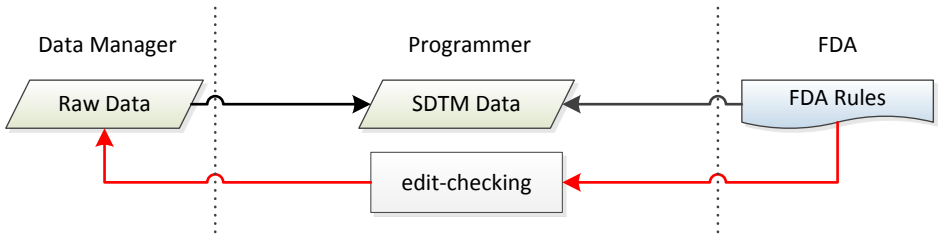
Edit-checking can be categorized as either a standard edit-checking or a study-specific in our standards. The standard edit-checks are developed based on the company's standard eCRFs, and they are supplemental to the basic edit-checks built in EDC system and can be

applied across studies. The study-specific edit-checks are developed for the data from new eCRFs and/or for what is specific to the study, usually efficacy data. An example of study-specific edit-checks for efficacy data is shown in Table 2. The words in bold were raw (EDC) dataset names. These raw datasets are examples from different CRFs of ***EEDSQ2***, ***EEDSQ3***, ***EEDSQ4***, ***EEDSQ5***, ***EEDSQ7***, ***PANSS***, and ***CGIS*** for efficacy data collection. Most of these logic checks are cross-page checks to identify inconsistent data issues from raw data collection for study data manager to send queries for data cleaning. These findings are very critical for the study results.

Edit-check Number	Edit-check	Issue Finding (Yes/No)
1	EEDS Subjects in EEDSQ2 dataset that are not in PANSS dataset	No
2	EEDS Subjects in PANSS dataset that are not in EEDSQ2 dataset	Yes
3	EEDS Subjects in EEDSQ3 dataset that are not in Individual PANSS Score (PANSS) dataset	No
4	EEDS Subjects in Individual PANSS Score (PANSS) dataset that are not in EEDSQ3 dataset	Yes
5	Subjects in OL dataset having post-randomization dose increase but are not in EEDSQ5 dataset	Yes
6	Discontinued EEDS Subjects in EEDSQ7 dataset that are not in DS dataset	Yes
7	Discontinued EEDS Subjects in DS dataset that are not in EEDSQ7 dataset	No
8	Any Individual PANSS item Assessment is Missing at Any Visit	Yes
9	Any Individual CGIS Assessment is Missing at Any Visit	No
10	EEDS found in EEDS4 but not CSSRS_L after randomization	Yes
11	EEDS found in CSSRS_L but not EEDS4 after randomization	Yes

Table 2. An Example of Study-specifc Edit-checks for Efficacy Data

Edit-checking is applied to raw data for data cleaning in order to achieve data quality and further ensure the compliance of SDTM data with FDA business rules. Below plot shows the edit-checking plays an indirect role to automatically ensure compliance of SDTM data with FDA Business Rules.



Display 3. The Role of Edit-checking to Ensure Compliance of SDTM Data with FDA Business Rules

Fifteen (15) rules, shown in Table 3, can be automatically addressed by data collection with the assistance of programmers through the edit-checking reports, which are used by the study data manager for data cleaning.

FDA Business Rule ID	FDA Business Rule
FDAB002	A value for a Toxicity (--TOX) variable should be provided, when a Toxicity Grade (--TOXGR) variable value is greater than 0.

FDA Business Rule ID	FDA Business Rule
FDAB005	Age or age range should be provided for all subjects, except for Screen Failures.
FDAB012	Assessment results should include units whenever a unit of measure is available.
FDAB021	Duplicate records should not be submitted (as constrained by the unique key in the underlying standard).
FDAB023	For Interventional studies all treated Subject should have Exposure data.
FDAB025	Randomized subjects are expected to receive a study treatment.
FDAB038	Timing information of clinical assessments should be submitted.
FDAB039	Upper Limit of Reference Range should be greater than Lower Limit.
FDAB040	Values for the following variables should not be negative: Age, Dose, and Duration of Event, Exposure or Observation.
FDAB041	Variable values should not include non-ASCII or non-printable characters (outside of 32-126 ASCII code range).
FDAB042	When a test is not done, the results should not be populated but the reason should be provided.
FDAB043	When End timepoint is provided, then a related Start timepoint should also be provided.
FDAB044	When Timing info is provided as a reference to a particular timepoint, the timepoint should be also populated.
FDAB045	Where Age is collected, units should be submitted.
FDAB063	The MI domain should contain at least one record for every scheduled tissue for all subjects in the study (i.e., if an organ was examined and found normal, it should have a record indicating NORMAL). In addition unscheduled tissues that were examined should also have a record. Subjects that were not scheduled for examination should not have records unless they were examined. If an organ was scheduled but not examined or no results, there should be a row with a reason not done.

Table 3. Fifteen FDA Business Rules Which Can Be Guaranteed by Data Collection (Edit-checking).

If non-compliant issues identified by Pinnacle 21 during FDA submission preparation are due to any FDA Business Rules from Table 3, it would be too late to fix these data issues due to data base lock. The only solution is to document and explain them in SDRG [5]. For an example of documenting removing duplicates of study baseline records in SDRG, please refer to "***Leveraging Study Data Reviewer's Guide (SDRG) in Building FDA's Confidence in Sponsor's Submitted Datasets***" [6] in PharmaSUG 2017 for details.

Even though Pinnacle 21 is used to check compliance with CDISC and "FDA Business Rules" "just one (1) week after the first subject first visit" [7], its findings regarding "Error" and/or "Warming" messages, which can be categorized as either data issues or SDTM mapping issues, are very difficult to understand by study data managers, in contrast to edit-checking reports, since Pinnacle 21 uses SDTM data as the inputs and has been developed per CDISC IG and FDA Business Rules. Furthermore, data managers are expected to be the experts in study raw data, not SDTM data, or CDISC IG, or FDA Business Rules. Use of reports from Pinnacle 21 to identify SDTM data issues (either data issues or mapping issues), and further trace back to raw data issues for data issues can help data cleaning to some extent, if the customized edit-checks are not available and/or not sophisticated enough. However the technical challenge to the study data manager (s) seriously hinders data cleaning. "Deciphering" Pinnacle 21 findings regarding "Error" and/or "Warming" messages will take an experienced SDTM programmer a significant amount of time if the report has too many findings due to the incomplete study data when Pinnacle 21 is used at the very early stage.

Another deficiency of Pinnacle 21 report is that it is usually designed for standard data issues instead of study-specific data issues, and therefore it cannot meet the study need for particular study specific edit-checks. Table 2 provides a convincing argument that Pinnacle 21 cannot be used as a tool to identify raw data issues due to inconsistent logics of the collected data from different CRF pages for efficacy data of a particular study. **Hence the better tool for the study data manager (s) to clean data is the customized edit-checking (both standard ones and study-specific) reports, not the reports from Pinnacle 21.**

The standard edit-checks have been built in our programming library, along with the customized reports and the standard process re communication with DM. Study-specific edit-checks have been developed through the collaboration with both study Biostatisticians and ADaM/TFL programmers. These tools have been utilized in more than thirty studies for data base locks, clinical study reports, and FDA submission-ready electronic submission packages across different compounds recently.

STANDARD SDTM PROGRAMMING PROCESS

Standard SDTM Programming Process includes two parts, i.e., SDTM Programming Convention (SDTMPC) and SDTM Programming Library (SDTMPL).

SDTMPC is our common practice to do SDTM programming in-house, which governs us how to build SDTMPC, how to train new staff to use SDTMPC, how to validate SDTM programming, and how to prepare electronic submission package.

SDTMPL includes metadata (specification) for SDTM domains, NCI/CDISC Controlled Terminology, SAS-based utility macros, and standard SDTM mapping templates, which are SAS programs for both production and validation, and cSDRG template. Both SDTMPC and SDTMPL have been developed and built per CDISC SDTM IG and FDA Business Rule since mid-2014. They will continue to be enhanced along with the advance of the standards of both CDISC and FDA Business Rule. For the detail of the method that can streamline the process from SDTM programming to FDA electronic submission preparation, please refer to "A Cost-Effective SDTM Conversion for NDA Electronic Submission" [8] in PharmaSUG 2011.Twenty two (22) FDA Business Rules can be met through our Standard SDTM Programming Process. Twelve (12) rules by SDTMPC, and nine (9) rules by SDTMPL. Table 4 and Table 5 display the FDA Business Rules to be met by SDTMPC and SDTMPL, respectively.

FDA Business Rule ID	FDA Business Rule
FDAB004	AE, CE, CM, DS, EG, EX, LB, MH, PC, PP, SE, SV, and VS should be submitted if collected.
FDAB011	All Trial Design data should be submitted as specified in the Technical Conformance Guide (TCG).
FDAB014	Category for Disposition Event (DSCAT) should be populated.
FDAB020	Demographics (DM) and Trial Summary (TS) domains must be submitted.
FDAB026	Records with a baseline flag should have a corresponding standard result with a standardized unit where available.
FDAB027	Required and Expected variables should be submitted.
FDAB029	Standard Character Result should be populated for all completed findings.
FDAB030	Standard Units should be consistent within the same assessment (having the same --TESTCD, --CAT, --SCAT, --SPEC, --METHOD values).

FDA Business Rule ID	FDA Business Rule
FDAB031	Standardized Result in Numeric Format should be populated whenever it is applicable.
FDAB033	Study data should be provided in SAS XPORT v5 (.xpt) format.
FDAB034	Study Start and End Dates should be submitted and complete where collected.
FDAB035	The definition of datasets, variables, and codelists in define.xml should reflect the actual study data.

Table 4. Twelve FDA Business Rules Which Can Be Guaranteed by SDTM Programming Convention

FDA Business Rule ID	FDA Business Rule
FDAB001	A treatment-emergent flag should be submitted.
FDAB008	All exposure records should occur between First and Last Study Treatment dates.
FDAB013	Baseline flags for Laboratory results, Vital Signs, ECG, Pharmacokinetic Concentrations, and Microbiology results should be submitted if the data was collected or can be derived.
FDAB015	Character values should not have leading spaces or only have a period character.
FDAB016	Collection Study Day should be populated when Date/Time of Collection is available.
FDAB022	EPOCH should be included for clinical subject-level observations (e.g., adverse events, laboratory, concomitant medications, exposure, and vital signs).
FDAB032	Start Date/Time of Observation (--DTC) or Study Day of Observation (--DY) should be populated.
FDAB036	The value for Study Day should not be negative for Exposure treatments.
FDAB037	Time Point Reference should be provided, when Reference Time Point is used.

Table 5. Nine FDA Business Rules Which Can Be Guaranteed by SDTM Programming Mapping Templates

IN-HOUSE SAS-BASED MACROS

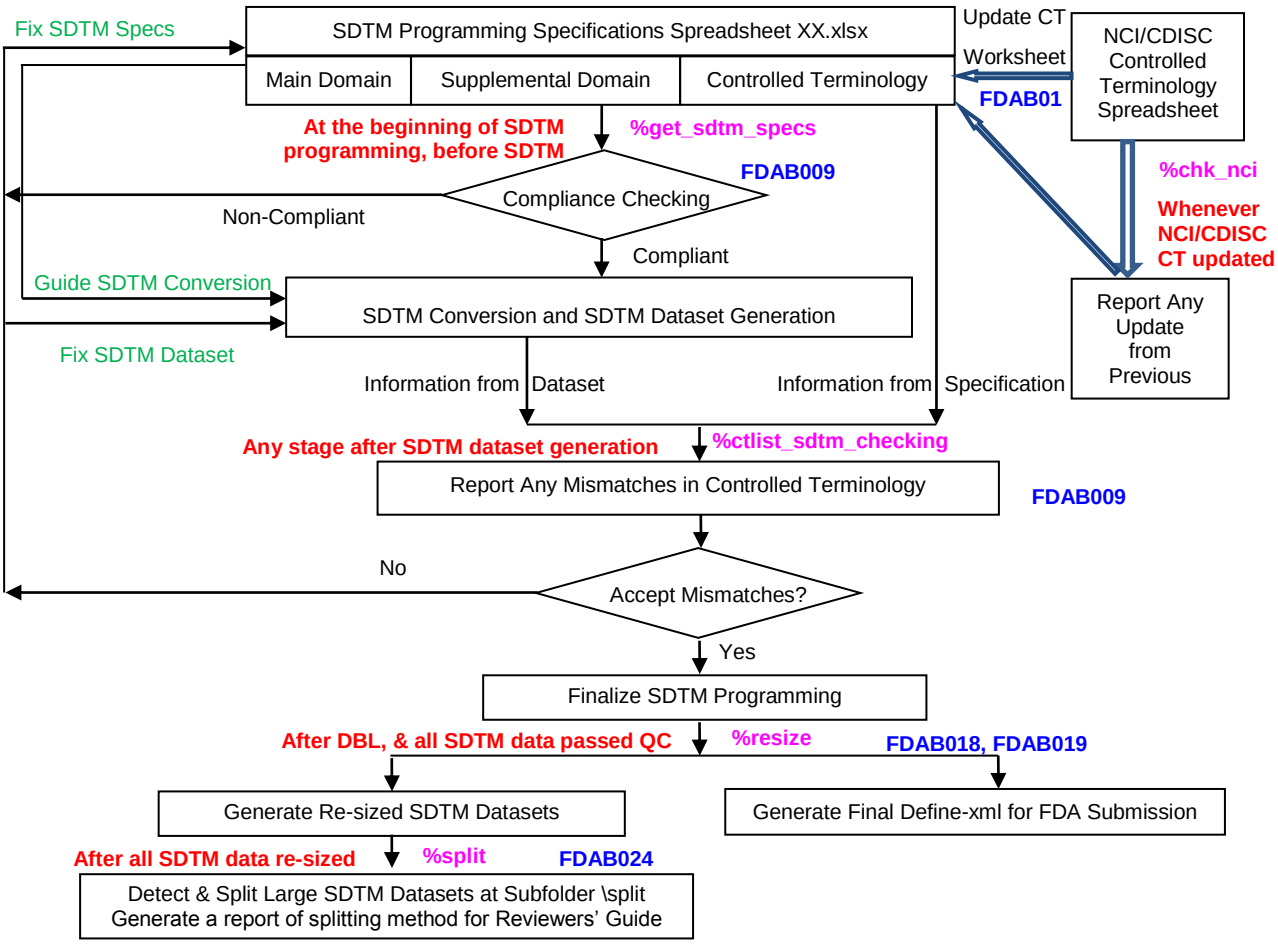
Table 6 below displays five FDA Business Rules and SAS-based macros which help to achieve full compliance of SDTM data with these rules. These macros were developed to accomplish the following:

1. Automatically check SDTM specifications and SDTM datasets against FDA Business Rules, to detect and report the issues to address non-compliance with “FDA Business Rules”
2. Automatically detect and report differences in NCI/CDISC Controlled Terminology between the standard spreadsheet with NCI/CDISC Controlled Terminology in SDTMPL and new release of NCI/CDISC Controlled Terminology Spreadsheet for an automatic/a manual update of controlled terminology worksheet in SDTMPL
3. Automatically resize each character variable length in SDTM datasets and simultaneously update each resized variable length in define.xml for FDA submission
4. Automatically detect whether the SDTM datasets exceed the FDA data-size limitation. If any are detected, the macro tool will automatically split the large datasets into several subsets in a sub-directory \split for FDA submission

FDA Business Rule ID	FDA Business Rule	In-house Macro to Ensure Compliance
FDAB009	All paired variables should have a one-to-one relationship. Examples include Short Name and Name of Test; Parameter Name and Parameter Code or Number; Variable Name and Variable Label, etc.	%Get_Sdtm_Specs %Ctlist_Checking_Sdtm
FDAB017	Controlled terms should use the exact same case used by the terminology maintenance organizations (e.g., MedDRA, CDISC controlled terminology).	%Chk_Nci
FDAB018	A variable's length across a study should be no longer than the maximum length of the actual data (except for SUPPQUAL).	%Resize
FDAB019	SUPPQUAL variable length should be no longer than the maximum length of the actual data within the dataset.	%Resize
FDAB024	Large datasets should be split into smaller datasets no larger than 5 GB in size.	%Split

Table 6. Five FDA Business Rules which Can Be Guaranteed by In-house Macros

PROGRAMMING FLOWCHART FOR IN-HOUSE MACROS TO ENSURE SDTM COMPLIANCE WITH FDA BUSINESS RULES



Display 4. Programming Flowchart for In-house Macros to Ensure the Compliance with FDA Business Rules

Below illustrates how these six SAS-based macros help to achieve full compliance of SDTM data with FDA Business rules.

IN-HOUSE MACROS %GET_SDTM_SPECS AND %CTLIST_CHECKING_SDTM FOR FDA BUSINESS RULE: FDAB009

FDA Business Rule ID	FDA Business Rule
FDAB009	All paired variables should have a one-to-one relationship. Examples include Short Name and Name of Test; Parameter Name and Parameter Code or Number; Variable Name and Variable Label, etc.

A macro %get_sdtm_specs [8] call reads the individual SDTM programming specification in spreadsheet format, and it automatically retrieves metadata information from the specification and converts it into a SAS dataset for SDTM mapping programming. Meanwhile, it checks the compliance of the metadata against both CDISC rules and FDA Business rules, and it reports any finding to the users for correcting the non-compliance in the specification, which covers "FDA Business Rule FDAB009": "All paired variables should have a one-to-one relationship".

Table 7 shows an example of two decode values for LBTEST vs. one value for LBTESTCD. It was from the code-list worksheet tab in LB specification on December 18, 2015. Per the new Controlled Terminology for the coded value of "HDLCLDLC", LBTEST should be mapped to "HDL Cholesterol/LDL Cholesterol".

The macro call reports the non-compliance, which is shown by Table 8. After the old lab test name "HDL Cholesterol/LDL Cholesterol Ratio" was deleted from LB specification, one-to-one relationship between LBTESTCD and LBTEST can be kept to meet "FDA Business Rule FDAB009".

The macro %get_sdtm_specs only handles metadata (programming specification), not raw data. It helps the user to ensure the compliance of SDTM controlled terminology with "FDA Business Rule FDAB009". For the detail of this macro, please refer to "***A Cost-Effective SDTM Conversion for NDA Electronic Submission***" in PharmaSUG 2011.

Variable	Codelist	Order	TESTCD	TEST
LBTESTCD	LBTEST	180	HDLCLDLC	HDL Cholesterol/LDL Cholesterol Ratio
LBTESTCD	LBTEST	180	HDLCLDLC	HDL Cholesterol/LDL Cholesterol

Table 7. An Example of Multiple Decode Values for Lab Test Code HDLCLDLC in Controlled Terminology Worksheet of LB Specifications

Checking Information
Checking: Code Value HDLCLDLC and Decode Value HDL Cholesterol/LDL Cholesterol are not 1:1 mapping for Code-list LBTESTCD
Checking: Code Value HDLCLDLC and Decode Value HDL Cholesterol/LDL Cholesterol Ratio are not 1:1 mapping for Code-list LBTESTCD

Table 8. An Example of Non-compliance Report - Violating 1:1 Mapping (FDA Business Rule FDAB009)

Once SDTM datasets are generated, a SAS macro %ctlist_checking _sdtm call is to check the proper use of Controlled Terminology to ensure the submission quality [9]. The macro compares the controlled terminology and QNAM-QLABEL pair assigned in the SDTM Programming Specifications with ones in the SDTM datasets, detects any mismatches, and generates inconsistency report in RTF format if any exists. Please refer to "***Automatic Consistency Checking of Controlled Terminology among SDTM Datasets,***

Define.xml, and NCI/CDISC Controlled Terminology for FDA Submission" [9] in PharmaSUG 2016 for details about the macro. Table 9 shows an example of data collection in drug screen tests with multiple lab test names ("Urine Cannabinoids" and "Urine 11-nor-9-Tetrahydrocannabinol-9-carboxylic acid") for LBTESTCD='CANNAB'.

DGTEST1	DGORRES1	DGTESCD1	DGTEST2	DGORRES2	DGTESCD2
Urine Cannabinoids	NEGATIVE	UCANNAB			
			Urine 11-nor-9-Tetrahydrocannabinol-9-carboxylic acid	POSITIVE	UCANNAB

Table 9. An Example of Data Collection in Drug Screen Tests: Multiple Lab Test Names for LBTESTCD='CANNAB'

LBTESTCD='CANNAB' had two different LBTEST values, after raw lab dataset was mapped to draft SDTM LB dataset. The value "Cannabinoids" was specified in controlled terminology worksheet in SDTMPL. Table 10 shows the report from a SAS macro %ctlist_checking _sdtm call. It shows the inconsistency between SDTM LB dataset and LB specification for LBTEST. LB test "Urine 11-nor-9-Tetrahydrocannabinol-9-carboxylic acid" should be mapped to "Cannabinoids" in program lb.sas in order to achieve the compliance of FDA Business Rule: **FDAB009**.

Table 10 shows inconsistency report from the SAS macro %ctlist_checking _sdtm call for the example in Table 9.

Domain	Variable	Variable Label	Code Value	Decode Value Label in Dataset	Decode Value Label in Specs.	Codelists In Specs. NOT in Dataset	Codelists In Dataset NOT In Specs.	Different Controlled Terminology
LB	LBTESTCD	Lab Test or Examination Short Name	CANNAB	11-nor-9-Tetrahydrocannabinol-9-carboxyl	Cannabinoids			Yes

Table 10. Non-Consistency Report of Code-Decode Codelists between SDTM Datasets and Specifications

This macro tool can be run at any stage of the programming cycle in order to facilitate finalizing SDTM programming activities earlier.

IN-HOUSE MACRO %CHK_NCI FOR FDA BUSINESS RULE FDAB017

FDA Business Rule ID	FDA Business Rule
FDAB017	Controlled terms should use the exact same case used by the terminology maintenance organizations (e.g., MedDRA, CDISC controlled terminology).

Our SDTM Programming Library (SDTMPL) has the standard spreadsheet with NCI/CDISC Controlled Terminology. It ensures the compliance with FDA Business Rule FDAB017 if the spreadsheet is up to the new standards.

A macro %chk-nci call automatically detects and reports the difference of NCI/CDISC Controlled Terminology between the standard spreadsheet with NCI/CDISC Controlled Terminology in SDTMPL and new release of NCI/CDISC Controlled Terminology Spreadsheet.

Table 11 (a), (b), (c) show the examples of the report from the macro call. Table 11 (a) reported the newly added NCI/CDISC controlled terminology in new version of NCI/CDISC controlled terminology spreadsheet; Table 11 (b) reported the deleted NCI/CDISC controlled

terminology in old version of NCI/CDISC spreadsheet; and Table 11 (c) reported the update of controlled terminology in the new version.

CODELIST_CODE	CODE	CODELIST_NEW	CODEVAL_NEW	DECODEVAL_NEW
C118971	C102118	CCCAT	HAM-A	

(a) Report CODE_IN_NEW_ONLY, for Newly Added NCI/CDISC Controlled Terminology in New Version

CODELIST_CODE	CODE	CODELIST_OLD	CODEVAL_OLD	DECODEVAL_OLD
C100129	C102118	QSCAT	HAM-A	

(b) Report CODE_IN_OLD_ONLY, for Deleted NCI/CDISC Controlled Terminology from Old Version in SDTMPL

CODELIST_CODE	CODE	CODELIST_NEW	CODELIST_OLD	CODEVAL_NEW	CODEVAL_OLD	DECODEVAL _NEW	DECODEVAL_OLD	UPDATE_CODELIST_NAME	UPDATE_CODE_VAL	UPDATE_DECODE_VAL
C65047	C100425	LBTESTCD	LBTESTCD	HDLCLDLC	HDLCLDLC	HDL Cholesterol/LDL Cholesterol	HDL Cholesterol/LDL Cholesterol Ratio			Y

(c) Report CODE_UPDATE, for Updated NCI/CDISC Controlled Terminology

Table 11. Report of NCI/CDISC Controlled Terminology Update

These reports serve as the input for an automatic/a manual update of controlled terminology worksheet in SDTMPL. The automation of detection and report from the macro call ensures the consistency of controlled terminology between one in SDTMPL and one from NCI/CDISC, and further ensures the compliance with FDA Business Rule FDAB017. Please refer to "***Automatic Consistency Checking of Controlled Terminology among SDTM Datasets, Define.xml, and NCI/CDISC Controlled Terminology for FDA Submission***" in PharmaSUG 2016 for details about the macro.

IN-HOUSE MACRO %RESIZE FOR FDA BUSINESS RULES FDAB018 AND FDAB019

FDA Business Rule ID	FDA Business Rule
FDAB018	A variable's length across a study should be no longer than the maximum length of the actual data (except for SUPPQUAL).
FDAB019	SUPPQUAL variable length should be no longer than the maximum length of the actual data within the dataset.

FDA Business Rules FDAB018 and FDAB019 require that "a variable's length or SUPPQUAL variable length should be no longer than the maximum length of the actual data". Re-sizing character variable length from the pre-determined in SDTM to the maximum length of the variable on the actual data values is the common solution to be compliant with FDA rule [10].

A SAS macro %resize call automatically resizes each character variable length in SDTM datasets and simultaneously updating each resized variable length in define.xml. Display 5 (b) shows an example of resized SDTM datasets and their define.xml in eSubmission folder. The reduction of each data file size can be easily seen from it.

Name	Date modified	Type	Size
ae	1/12/2016 9:19 AM	SAS Data Set	1,836 KB
cm	1/12/2016 9:19 AM	SAS Data Set	2,808 KB
da	1/12/2016 9:19 AM	SAS Data Set	2,752 KB
dm	1/12/2016 9:19 AM	SAS Data Set	408 KB
ds	1/12/2016 9:19 AM	SAS Data Set	704 KB
dv	1/12/2016 9:19 AM	SAS Data Set	128 KB
eg	1/12/2016 9:19 AM	SAS Data Set	18,648 KB
ex	1/12/2016 9:19 AM	SAS Data Set	320 KB
ie	1/12/2016 9:19 AM	SAS Data Set	128 KB
lb	1/12/2016 9:19 AM	SAS Data Set	153,720 KB
mh	1/12/2016 9:19 AM	SAS Data Set	1,088 KB
pc	2/8/2016 3:17 PM	SAS Data Set	6,052 KB
qs	1/12/2016 9:19 AM	SAS Data Set	284,184 KB
se	1/12/2016 9:19 AM	SAS Data Set	576 KB
su	1/12/2016 9:19 AM	SAS Data Set	192 KB
suppae	1/12/2016 9:19 AM	SAS Data Set	1,920 KB
suppcm	1/12/2016 9:19 AM	SAS Data Set	3,200 KB
suppda	1/12/2016 9:19 AM	SAS Data Set	4,608 KB
suppdm	1/12/2016 9:20 AM	SAS Data Set	1,472 KB
suppds	1/12/2016 9:20 AM	SAS Data Set	448 KB
suppeg	1/12/2016 9:20 AM	SAS Data Set	320 KB
supplb	1/12/2016 9:20 AM	SAS Data Set	10,816 KB
suppqs	1/12/2016 9:20 AM	SAS Data Set	7,552 KB
suppsu	1/12/2016 9:20 AM	SAS Data Set	192 KB
suppsv	1/12/2016 9:20 AM	SAS Data Set	2,368 KB
sv	1/12/2016 9:20 AM	SAS Data Set	2,432 KB
ta	1/11/2016 5:07 PM	SAS Data Set	128 KB
te	1/11/2016 5:07 PM	SAS Data Set	128 KB
ti	1/11/2016 5:07 PM	SAS Data Set	128 KB
ts	1/11/2016 5:07 PM	SAS Data Set	128 KB
tv	1/11/2016 5:07 PM	SAS Data Set	128 KB
vs	1/12/2016 9:20 AM	SAS Data Set	15,168 KB

(a) SDTM size before re-sizing

Name	Size	Type	Date modified
Project123-SDRG	354 KB	Microsoft Word 97 ...	3/11/2016 1:29 PM
AE	859 KB	SAS Xport Transpo...	2/8/2016 3:34 PM
CM	1,229 KB	SAS Xport Transpo...	2/8/2016 3:34 PM
DA	1,047 KB	SAS Xport Transpo...	2/8/2016 3:34 PM
DM	134 KB	SAS Xport Transpo...	2/8/2016 3:34 PM
DS	384 KB	SAS Xport Transpo...	2/8/2016 3:34 PM
DV	8 KB	SAS Xport Transpo...	2/8/2016 3:34 PM
EG	4,622 KB	SAS Xport Transpo...	2/8/2016 3:34 PM
EX	69 KB	SAS Xport Transpo...	2/8/2016 3:34 PM
IE	7 KB	SAS Xport Transpo...	2/8/2016 3:34 PM
LB	60,611 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
MH	473 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
PC	2,349 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
QS	205,221 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
SE	204 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
SU	40 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
SUPPAE	1,431 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
SUPPCM	1,129 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
SUPPDA	2,245 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
SUPPDM	370 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
SUPPDS	105 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
SUPPEG	53 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
SUPPLB	2,233 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
SUPPQS	1,999 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
SUPPSU	20 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
SUPPSV	475 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
SV	846 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
TA	10 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
TE	4 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
TI	12 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
TS	14 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
TV	6 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
VS	7,449 KB	SAS Xport Transpo...	2/8/2016 3:35 PM
define	442 KB	XML Document	2/8/2016 3:35 PM
define2-0-0	91 KB	XSL Stylesheet	8/15/2013 4:02 PM

(b) SDTM size after re-sizing

Display 5. A Snapshot of SDTM Datasets and their Resized Datasets in XPT format and Their Define.xml for FDA submission readiness

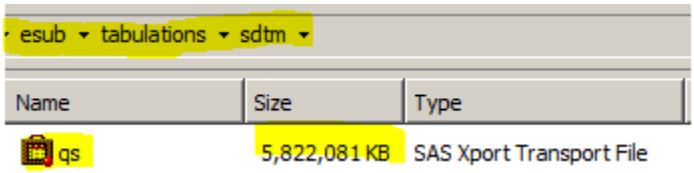
This macro tool was performed after the study database was locked, and all SDTM data passed validation, which makes the re-sizing almost zero impact to our programming activities.

IN-HOUSE MACRO %SPLIT FOR FDA BUSINESS RULE FDAB024

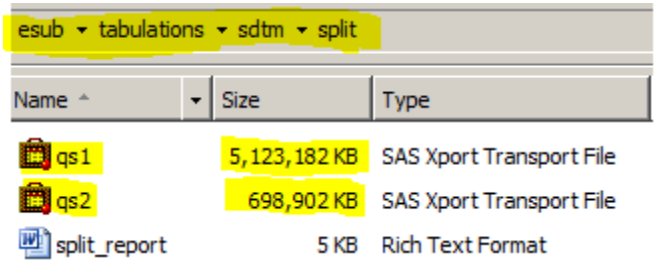
FDA Business Rule ID	FDA Business Rule
FDAB024	Large datasets should be split into smaller datasets no larger than 5 GB in size.

A macro call %split detects whether the SDTM datasets exceed the FDA data-size limitation. If any detected, the macro tool will automatically split the large datasets into several subsets in a sub-directory \split, and output a report for the method splitting the large dataset for Reviewer's guide use.

Display 6 shows the folder for split datasets, and the report for splitting. For a large dataset QS, smaller subsets are named QS1 and QS2.



(a) SDTM dataset QS.XPT which is larger than 5 GB in size (5,822,081KB=5.55GB)



(b) Split datasets QS1.XPT (5,123,182KB=4.89GB) and QS2.XPT (698,902KB=0.67GB)

Dataset Before Splitting	Datasets After Splitting	Total Observations	Rules of Splitting
QS	QS1	8274660	QSCAT in ("C-SSRS BASELINE", "C-SSRS SINCE LAST VISIT", "CGI", "COWS", "HAM-A", "HAMD 17")
	QS2	1128820	QSCAT in ("MADRS")

(c) Method of Splitting datasets QS1.XPT and QS2.XPT

Display 6. Split Large Datasets into Smaller Datasets No Larger Than 5 GB in Size.

FDA BUSINESS RULE AND STUDY DATA VALIDATOR RULE

FDA published both "FDA Business Rules" [1] and FDA Study Data Validator Rules [1] in October 2018. Per section 8 in Study Data Technical Conformance Guide [3], "The business rules are accompanied with validator rules which provide detail regarding FDA's assessment of study data for purposes of review and analysis."

The spreadsheet of FDA Study Data Validator Rules provides the details for each of Business Rule and its Study Data Validator Rules. Some of the Business Rules have multiple Study Data Validator Rules, and some do not have any corresponding Study Data Validator Rules. Ten Business Rules without Study Data Validator Rules are not covered by Pinnacle 21 Report. Our solution to them is the first pillar "**Study Data Collection Design**". Also there are fourteen Business Rules without supporting from Pinnacle 21 Report.

CONCLUSION

This paper presents a systematic approach to automate SDTM programming process to ensure compliance with FDA Business Rules. This systematic approach is composed of these four pillars: **study data collection design, data collection (edit-checking), standard SDTM programming process, and in-house macros**. It illustrates how each pillar can help sponsors achieve full compliance of SDTM data with FDA Business rules. It further explains why this approach is far superior to Pinnacle 21 re ensuring compliance of SDTM data with FDA Business Rules. The sharing of hands-on experiences in this paper is to assist

readers to apply this methodology to prepare FDA Business Rule compliant SDTM datasets for FDA submission in order to ensure the technical accuracy and submission quality, in addition to cost-effectiveness and efficiency.

REFERENCES

[1] https://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/ucm2005545.htm

[2] CDISC Submission Data Standards Team. "Study Data Tabulation Model Implementation Guide: Human Clinical Trials". November 2013. http://www.cdisc.org/sdtm

[3] U.S. Department of Health and Human Services, Food and Drug Administration, Study Data Technical Conformance Guide: Technical Specifications Document. November 2017. Available at http://www.fda.gov/downloads/ForIndustry/DataStandards/StudyDataStandards/UCM384744.pdf.

[4] U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER). "Oversight of Clinical Investigations — A Risk-Based Approach to Monitoring". Available at https://www.fda.gov/downloads/Drugs/Guidances/UCM269919.pdf

[5] Pharmaceutical Users Software Exchange, Study Data Reviewer's Guide Final Work Packages: SDRG Package v1.2 2015-01-26. Available at http://www.phusewiki.org/wiki/index.php?title=Study_Data_Reviewer%27s_Guide.

[6] Xiangchen (Bob) Cui, Min Chen, and Letan (Cleo) Lin. "Leveraging Study Data Reviewer's Guide (SDRG) in Building FDA's Confidence in Sponsor's Submitted Datasets", PharmaSUG, May 2017.

[7] Sergiy Sirichenko. Pinnacle 21,"Common Programming Errors in CDISC Data", PharmaSUG, May 2017.

[8] Xiangchen (Bob) Cui, Scott Moseley, Min Chen. "A Cost-Effective SDTM Conversion for NDA Electronic Submission", PharmaSUG, May 2011.

[9] Min Chen, Xiangchen (Bob) Cui. "Automatic Consistency Checking of Controlled Terminology among SDTM Datasets, Define.xml, and NCI/CDISC Controlled Terminology for FDA Submission", PharmaSUG, May 2016.

[10] Xiangchen (Bob) Cui, Min Chen. "A Practical Approach to Re-sizing Character Variable Lengths for FDA Submission Datasets (both SDTM and ADaM)", PharmaSUG, May 2016.

ACKNOWLEDGMENTS

Appreciation goes to Sondra Smyrnios for her valuable review and comments.

CONTACT INFORMATION

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

Name: Xiangchen (Bob) Cui, Ph.D.
Enterprise: Alkermes, Inc.
Address: 852 Winter Street
City, State ZIP: Waltham, MA 02451
Work Phone: 781-609-6038
Fax: 781-609-5855
E-mail: xiangchen.cui@alkermes.com

Name: Hao Guan, M.S.
Enterprise: Alkermes, Inc.
Address: 852 Winter Street
City, State ZIP: Waltham, MA 02451
Work Phone: 781-609-6813
Fax: 781-609-5855
E-mail: hao.guan@alkermes.com

Name: Min Chen, Ph.D.
Enterprise: Alkermes, Inc.
Address: 852 Winter Street
City, State ZIP: Waltham, MA 02451
Work Phone: 781-609-6047
Fax: 781-609-5855
E-mail: min.chen@alkermes.com

Name: Letan (Cleo) Lin, M.S.
Enterprise: Alkermes, Inc.
Address: 852 Winter Street
City, State ZIP: Waltham, MA 02451
Work Phone: 781-609-6380
Fax: 781-609-5855
E-mail: letan.lin@alkermes.com