PharmaSUG 2015 - Paper SS08-SAS

Getting Rid of Bloated Data in FDA Submissions

Ben Bocchicchio, SAS Institute, Cary, NC
Frank Roediger, SAS Institute, Cary, NC

ABSTRACT

As the management of FDA submission data has become more automated, the amount of data that the FDA has to keep track of has grown rapidly. In an attempt to manage this growing mountain of data, the FDA issues Technical Specifications Documents that stipulate rules that submission data should follow to help the review process proceed as smoothly as possible. The most recent version, *Study Data Technical Conformance Guide*, was issued in December 2014, and contains a link to 314 validation rules that are being proposed to take effect with the Prescription Drug User Fee Act VI (PDUFA VI) in 2017.

This presentation includes a survey of the validation rules and identifies resources that you can use to ensure that submission data complies with them. We provide a detailed discussion of 12 high-level rules that apply to folders, transport files, and data sets; we also provide utility programs to help evaluate whether your submission complies with those rules.

Currently, *FDAC036 – Variable length is too long for actual data* is a rule that is not often observed. We explain why it is so important to comply with this rule and provide access to a utility program that ensures that your character variables are just wide enough to store the longest value that they need to contain.

Another rule, FDAC016, is that data sets whose transport files are larger than 1 GB must be split into several data sets and that the corresponding transport files do not exceed the 1 GB limit. We explain the rationale behind this rule and discuss the problems that the FDA encounters when this rule is not observed.

INTRODUCTION

THIS IS AN IMPORTANT DISCLAIMER: The content of this presentation is not made under the aegis or endorsement of the FDA. The authors are employees of SAS Institute, Inc., and have worked as consultants with the FDA on a data warehouse project to load and manage a data repository where FDA reviewers can readily access submission data. We have first-hand experience with the problems with loading submission data into the FDA's warehouse caused by noncompliance with the *Guide*'s rules. We want to help applications submitted to the FDA avoid these problems. The identification of validation rules and checks and the utility programs that we provide are initial contributions to that objective.

The *Study Data Technical Conformance Guide* (the *Guide*) that the FDA published in December 2014 describes the submission requirements that become effective with PDUFA VI (PDUFA V, which is currently in effect, expires in September 2017). The *Guide* contains a link to a spreadsheet with 314 validation rules and it defines additional rules in the body of the document. For example, "There should be one dataset per XPORT file" is a rule that appears on p.6 in the *Guide* but is not specifically contained in the spreadsheet's list of rules.

The FDA is careful to explain in the *Guide* that the rules are not inflexible dogma: at the top of each page in the *Guide* is the heading "Contains Nonbinding Recommendations." Also, the *Guide* frequently contains language such as "Sponsors should discuss… with the review division," which is clearly an invitation for sponsor/FDA dialog. The FDA also makes clear that it considers the current version to be a work in progress: as updates become available, the FDA will post revisions of the document on its Web site.

If these rules are still two years away and they will likely be revised, why should we pay attention to them now?

The FDA published the *Guide* because it identifies submission data problems that currently cause operational delays. Even though its rules will not be active until 2017, conforming them prior to that date reduces problems the FDA commonly encounters when reviewing submission data. Furthermore, two years is not a particularly long time to establish practices for producing submission data, especially if you do not yet have working practices in place to produce submission data that meets a significant number of the rules.

Our utility programs are a rudimentary way to determine whether currently-produced submission packages conform to some of these rules. Before we discuss our utility programs, however, we present a method to identify existing automated resources that can already test whether submission data follows the FDA's validation rules.

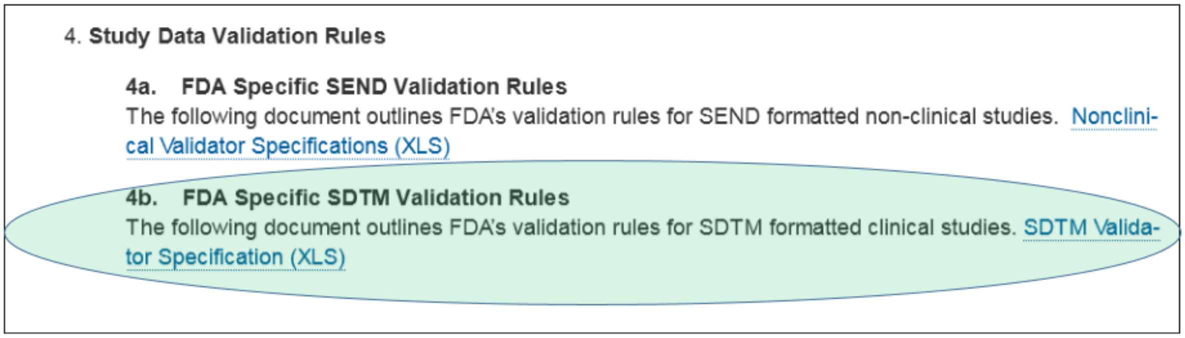
IDENTIFYING OPENCDISC CHECKS FOR FDA'S STUDY DATA TABULATION MODEL (SDTM) VALIDATION RULES

OpenCDISC has a Web site where you can download its validation checks. By reconciling OpenCDISC's list of validation checks with the FDA's validation rules, you can determine which FDA rules cannot be verified by the OpenCDISC validation checks and will need to be handled separately.

FDA'S VALIDATION RULES

The FDA's SDTM Validation Rules are available in a spreadsheet that can be accessed from a link on the following page: http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm (see Display 1). Clicking the link activates a dialog that allows you to open the spreadsheet or download it to your computer.

Display 1: Link to the SDTM Validation Specification Spreadsheet



4. **Study Data Validation Rules**

4a. FDA Specific SEND Validation Rules
The following document outlines FDA's validation rules for SEND formatted non-clinical studies. Nonclinical Validator Specifications (XLS)

4b. FDA Specific SDTM Validation Rules
The following document outlines FDA's validation rules for SDTM formatted clinical studies. SDTM Validator Specification (XLS)

If you choose to download the spreadsheet, the **SDTM validation rules 1.0.xls** file is copied to your **\Downloads** directory.

The spreadsheet contains one row for each rule, with the following columns:

- **FDA Rule ID** – a 7-character identifier in the form, "FDAC" followed by 3 digits (for example, *FDAC001*)
- **Message** – a brief notification to use when the rule fails (for example, *Missing DM dataset*)
- **Description** – an explanation of the rule's purpose (for example, *Demographics (DM) dataset must be included in every submission*)
- **Domains** – the SDTM domains for which the rule applies (for example, *DM* or *ALL*)
- **Severity** – whether failing the rule is an *Error* or a *Warning*
- Three columns that indicate to which SDTM version(s) the rule applies:
  - 3.1.1
  - 3.1.2
  - 3.1.3 (also, 3.1.2 A1)

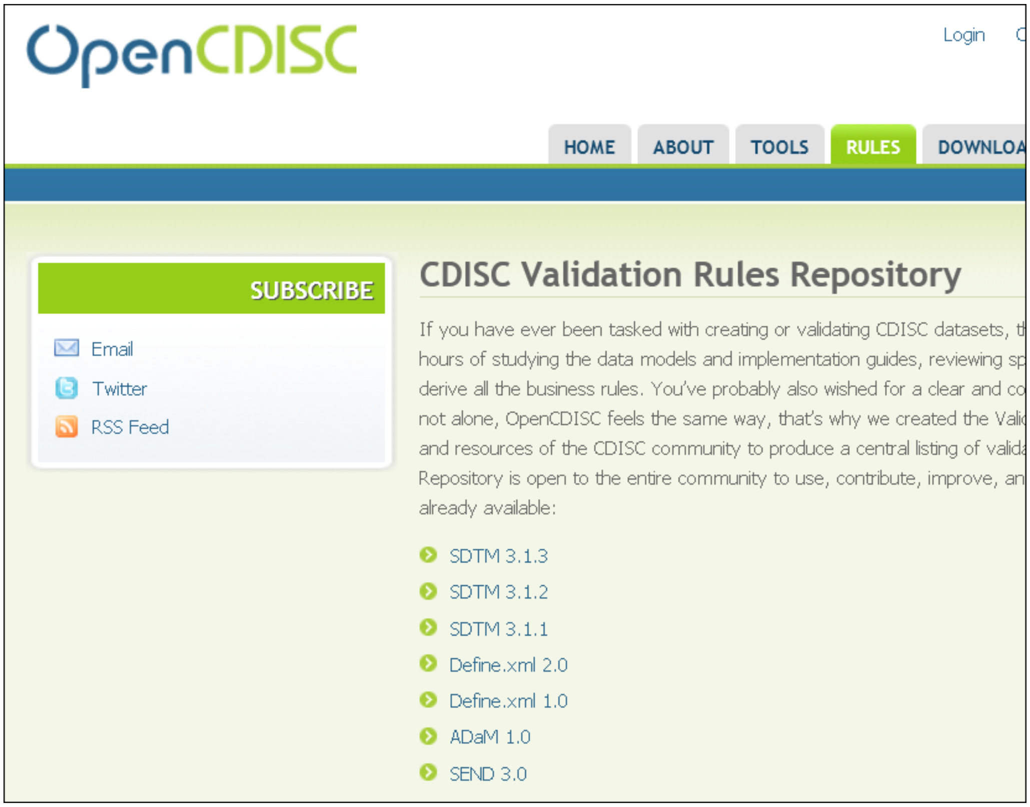
The SAS code that converts the spreadsheet into a SAS data set, as well as the other code that is referred to in this paper, is available on the SAS Drug Development Forum (https://communities.sas.com/community/support-communities/sas-drug-development), a Web page where users can get help with—and post tips about—pharma programming issues. The Forum was set up specifically to help SAS Drug Development users, but it also contains information about more general pharma issues as well.

Reading the spreadsheet into a SAS data set makes it feasible to reconcile the FDA's SDTM validation rules with the OpenCDISC validation checks.

OPENCDISC VALIDATION CHECKS

OpenCDISC has automated processes to check SDTM rule compliance on its Web site. The automated processes can be downloaded by clicking on the home page's **Download** tab and can be listed by clicking on the home page's **Rules** tab (see Display 2).

Display 2: OpenCDISC Rules Tab



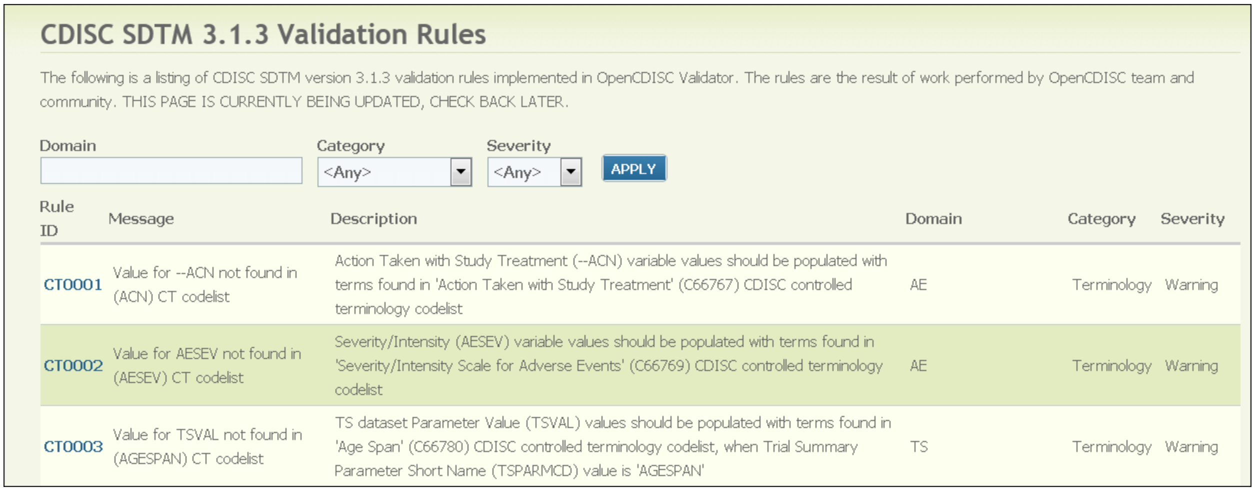
CDISC Validation Rules Repository

If you have ever been tasked with creating or validating CDISC datasets, t
hours of studying the data models and implementation guides, reviewing sp
derive all the business rules. You've probably also wished for a clear and co
not alone, OpenCDISC feels the same way, that's why we created the Vali
and resources of the CDISC community to produce a central listing of valid
Repository is open to the entire community to use, contribute, improve, an
already available:

- SDTM 3.1.3
- SDTM 3.1.2
- SDTM 3.1.1
- Define.xml 2.0
- Define.xml 1.0
- ADaM 1.0
- SEND 3.0

Clicking on the **SDTM 3.1.3** link takes you to the OpenCDISC page that documents the rules that are specific for SDTM 3.1.3 (see Display 3).

Display 3: OpenCDISC SDTM 3.1.3 Validation Rules



CDISC SDTM 3.1.3 Validation Rules

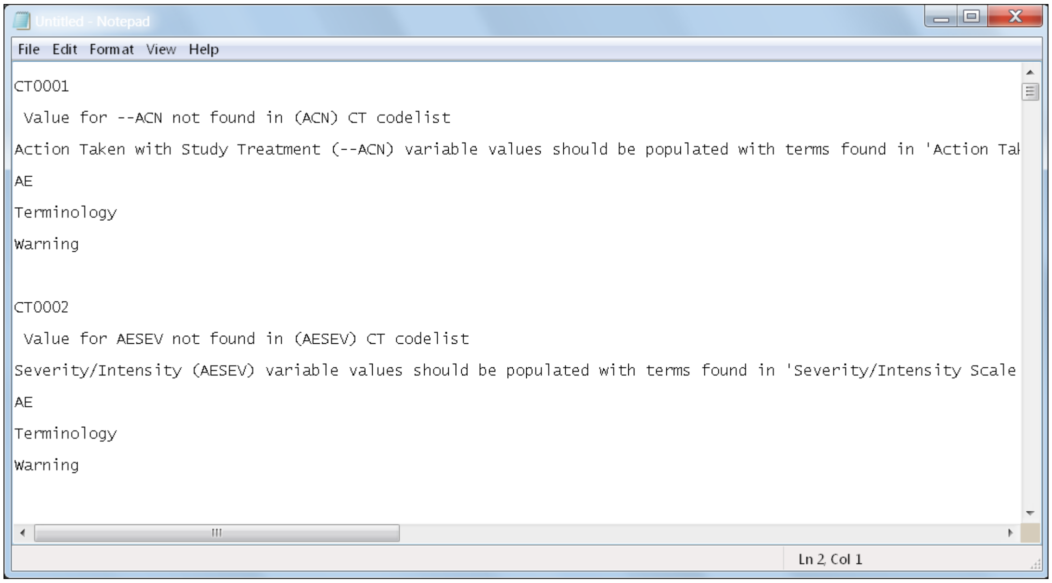
The following is a listing of CDISC SDTM version 3.1.3 validation rules implemented in OpenCDISC Validator. The rules are the result of work performed by OpenCDISC team and community. THIS PAGE IS CURRENTLY BEING UPDATED, CHECK BACK LATER.

Rule ID	Message	Description	Domain	Category	Severity
CT0001	Value for --ACN not found in (ACN) CT codelist	Action Taken with Study Treatment (--ACN) variable values should be populated with terms found in 'Action Taken with Study Treatment' (C66767) CDISC controlled terminology codelist	AE	Terminology	Warning
CT0002	Value for AESEV not found in (AESEV) CT codelist	Severity/Intensity (AESEV) variable values should be populated with terms found in 'Severity/Intensity Scale for Adverse Events' (C66769) CDISC controlled terminology codelist	AE	Terminology	Warning
CT0003	Value for TSVAL not found in (AGESPAN) CT codelist	TS dataset Parameter Value (TSVAL) values should be populated with terms found in 'Age Span' (C66780) CDISC controlled terminology codelist, when Trial Summary Parameter Short Name (TSPARMCD) value is 'AGESPAN'	TS	Terminology	Warning

OpenCDISC documents 243 SDTM 3.1.3 validation rules. It would be convenient to be able to automatically download the contents of this Web page, but the purpose of the page is to dynamically view all the rules (or a subset of the rules using the Domain, Category, and Severity filters). However, the page does allow you to select its contents using click-and-drag and copy your selection to a text file.

Note: Different browsers can affect how the copied contents appear after you paste them into a text editor. We chose to use Internet Explorer to view the OpenCDISC page when we copied its contents and Notepad when we pasted the copied contents. We saved the Notepad contents as a .txt file. This Internet Explorer/Notepad combination rendered the Web page's contents in the following form, which is conveniently parse-able (see Display 4).

Display 4: OpenCDISC Web Page Contents Converted to a Text File



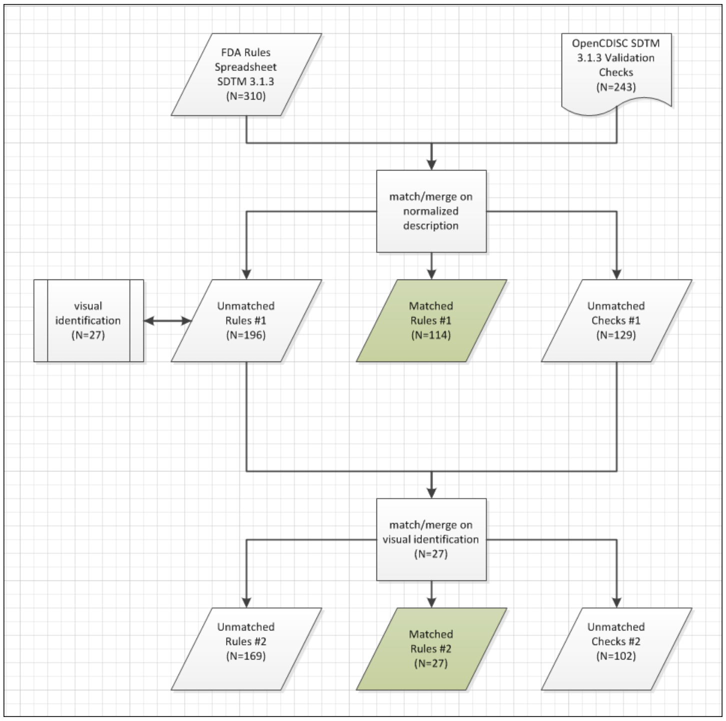
```
CT0001
 Value for --ACN not found in (ACN) CT codelist
Action Taken with Study Treatment (--ACN) variable values should be populated with terms found in 'Action Tak
AE
Terminology
Warning

CT0002
 Value for AESEV not found in (AESEV) CT codelist
Severity/Intensity (AESEV) variable values should be populated with terms found in 'Severity/Intensity Scale
AE
Terminology
Warning
```

The DATA Step code that we used to parse the .txt file and convert its contents into a SAS data set is available to review and download from the SAS Drug Development Forum.

RECONCILING FDA RULES WITH OPENCDISC VALIDATION CHECKS

With both the FDA Rules and the OpenCDISC validation checks represented in SAS data sets, it is feasible to identify which rules are already covered by existing checks. The process is outlined in Figure 1 (the actual code is available on the SAS Drug Development Forum).

Figure 1: Diagram of Rules/Checks Reconciliation Process


FDA Rules Spreadsheet SDTM 3.1.3 (N=310)
OpenCDISC SDTM 3.1.3 Validation Checks (N=243)
match/merge on normalized description
visual identification (N=27)
Unmatched Rules #1 (N=196)
Matched Rules #1 (N=114)
Unmatched Checks #1 (N=129)
match/merge on visual identification (N=27)
Unmatched Rules #2 (N=169)
Matched Rules #2 (N=27)
Unmatched Checks #2 (N=102)

The process is a rudimentary matching of the FDA's SDTM 3.1.3 rules with OpenCDISC's SDTM 3.1.3 checks based on the descriptions. To reduce any false negatives, we normalized the descriptions from both sources by converting them to uppercase and by compressing multiple consecutive blanks to single blanks. This resulted in 114 matches. We then visually compared the unmatched records from both sources to identify strongly similar descriptions. From this comparison, we identified 27 more matches.

Although more than half of the FDA's SDTM 3.1.3 rules remained unmatched, we stopped the reconciliation effort at this point. We bundled the results and stored them on the SAS Drug Development Forum. We hope to continue this effort and make more progress with the reconciliation (a colleague wants to tackle the problem with SAS Enterprise Miner/Text Miner). Although we at SAS have established this foundation, this is not exclusively a SAS initiative: anyone who wants to contribute to this effort on the SAS Drug Development Forum is welcome to do so.

FOUR UTILITIES TO CHECK FOR FDA SDTM VALIDATION RULES

Even if we had matched up all the OpenCDISC checks with the FDA rules, there would still be work to do: there are only 243 OpenCDISC SDTM 3.1.3 checks, but there are 310 FDA SDTM 3.1.3 rules, which leaves 67 rules without checks. Also, some rules are discussed by the FDA in the body of the *Guide* but are not listed in the FDA rules spreadsheet.

We will identify 12 rules and discuss how we have bundled them into four utility programs (see Table 1).

Table 1: Four Utility Programs to Check Submissions for FDA Rules

Utility Program	Source	Rule
Folders	p. 23	Data is not submitted in the correct directory structure
	p. 23	Directory structure contains empty folders
	p. 23	Directory structure has additional subfolders
Transport Files	p. 6	More than one data set in a transport file
	p. 6	Data set name/transport file name are not the same
Data Sets	p. 13	Data set label is not unique within the study
	p. 7	Label (variable or data set) contains illegal characters
	FDAC014	Domain table should have at least one record
	FDAC016	Dataset is greater than 1 GB in size
	FDAC036	Variable length is too long for actual data
Split Data Sets	FDAC072	Invalid dataset name for split domain
	p. 11	Split data sets do not match the source data set

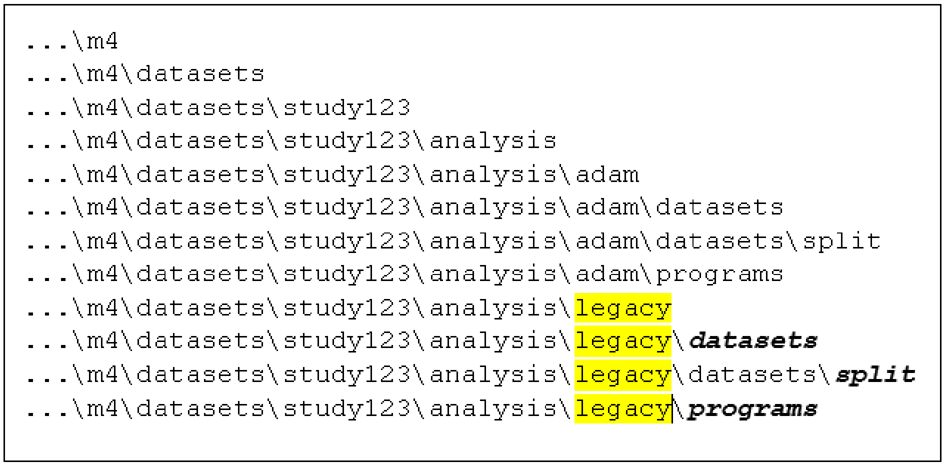
UTILITY PROGRAM #1: CHECKING FOLDERS

The FDA has specific expectations regarding the folder structure that organizes submission data (see *Guide,* pp. 23-25). When submissions comply with this structure, FDA reviewers, data managers, and data administration software can readily locate the study data and efficiently move it along the review process; submissions without this structure require time-consuming and wasteful manual intervention.

The directory structure that the FDA expects submissions to use is elaborate – Display 5 presents a portion of the folders for M4 (non-clinical data). The reason for so many folders is to provide an established location for every type of file that will be in any submission. If a submission does not have any files to put into a particular folder, the unused folders (and their paths) should not be part of the submission deliverable. For example, in Display 5, if a submission does not have non-clinical legacy data, there would not be any need to include the ***highlighted*** folders; there also would not be any need to include the \legacy folder because its only purpose is to provide a location for the empty ***highlighted*** folders. When place-holding empty folders are included in a submission structure, reviewers have to traverse meaningless paths and then need to determine whether a folder was intentionally or mistakenly left empty.

Additional folders should not be added to a submission directory structure. The *Guide* recommends, "If you feel that additional folders are needed, please consult with the appropriate center in advance for guidance" (p. 23).

Display 5: Some Folders for Storing Non-Clinical (M4) Files



```
...\m4
...\m4\datasets
...\m4\datasets\study123
...\m4\datasets\study123\analysis
...\m4\datasets\study123\analysis\adam
...\m4\datasets\study123\analysis\adam\datasets
...\m4\datasets\study123\analysis\adam\datasets\split
...\m4\datasets\study123\analysis\adam\programs
...\m4\datasets\study123\analysis\legacy
...\m4\datasets\study123\analysis\legacy\datasets
...\m4\datasets\study123\analysis\legacy\datasets\split
...\m4\datasets\study123\analysis\legacy\programs
```

A utility program to verify that these rules are being followed is available on the SAS Drug Development Forum.

UTILITY PROGRAM #2: CHECKING TRANSPORT FILES

Most often, the FDA moves submission data sets in the V5 transport files that they were received in. This is efficient, but it also can cause some problems, namely: how can you be sure what a V5 transport file contains without converting its contents back into SAS data sets? For this reason, the FDA specifies that a transport file should contain only one data set and that the data set should have the same name as the transport file that contains it. For example, if a transport file is named **ae** (with an extension of **.xpt**), it should contain only one data set and that data set should also be named **ae**. For the most part, submissions comply with these rules, but when they do not, the FDA's automated data management and review processes must be supplemented with time-consuming manual tasks.

A utility program that confirms that these two transport files rules are being observed is available on the SAS Drug Development Forum.

UTILITY PROGRAM #3: CHECKING DATA SETS

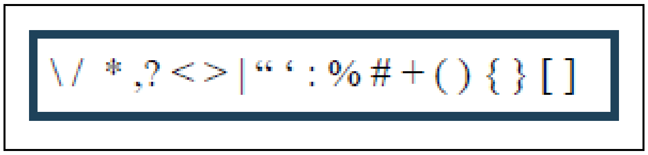
Even if a submission correctly creates V5 transport files and correctly locates them in the appropriate set of standard directories, the review process can still get bogged down with issues related to the data sets themselves. Another utility program (also available on the SAS Drug Development Forum) checks for data set-related problems

The data set label is an item of metadata that can provide supplemental information about a SAS data set. Because the CDISC SDTM standard domain names are limited to two characters (six characters for SUPP-- domains), a data set label can be a useful way to clearly identify the domain (for example, the data set label **Drug Accountability** is a more complete description of a domain than just the domain name, **DA**). This is particularly true for custom domain names that start with **X**, **Y**, or **Z**. In order for a label to be meaningful, however, it must be associated with only one domain, so the FDA has stipulated that data set labels need to be unique within a study.

File names, data set names, and variable names cannot contain any of the characters listed in Display 6 (directory names do not need to be tested for these characters because as long as a submission uses the correct directory structure, its directory names will not contain these characters).

Note: These characters cannot be used in any submission file name, not just in V5 transport file names. These characters are considered illegal because their use in names causes operational problems.

Display 6: Characters That Are Not Valid in Names



\/ * ,? < > | “ ‘ : % # + ( ) { } [ ]

In addition to these characters, two other characters that cannot be used in names: **&** (ampersand) and **spaces** (especially consecutive spaces). These characters are illegal because their use violates the naming requirements for the FDA's repository from which reviewers retrieve submission data.

Domain data sets that do not have any rows are an operational problem for the FDA because certain tests (especially cross-domain references) depend on actual data, not just on metadata. For this reason, the FDA specifies that every submission domain have at least one row (if a domain is not used in a study, it should be omitted from the submission package and not included as an empty shell table).

Large data sets pose several operational problems for the FDA:

- moving very large files around the FDA's internal networks takes a great deal of time and increases the likelihood of time outs, which require manual intervention and re-starts;
- the FDA has protocols that mandate when and where within its internal networks security scans need to be performed; extremely large files take longer to scan than smaller files;
- extremely large data sets (and extremely large V5 transport files) require that automated processes be coded to handle these outliers; this increases the complexity of the automated processes and causes a disproportionate expansion of development, testing, and maintenance efforts.

For this reason, the FDA has indicated that submissions are not to contain a file that is bigger than 1 GB. Certain domains (notably **Laboratory Test Results – LB** and **Clinical Events – CE**) are customarily large and frequently exceed the 1 GB threshold. When this happens, sponsors should split the large original data set into "smaller data sets no larger than 1 GB" (*Guide*, p. 6).

Bloated character fields occur when a field length is greater than what is required to store that field's longest value. For example, character field **XYZ** has a length of **$200** (the maximum size for a data set that will be converted into a V5 transport file) but its longest value is only **$25**, which means that each row in the data set has at least 175 padded bytes – and that is only for field **XYZ**. This type of bloating is a concern for the FDA because it poses the same operational problems as large data sets – but there is no content that makes the size of the bloated character fields worthwhile. For this reason, the FDA stipulates, "The allotted length for each column containing character (text) data should be set to the maximum length of the variable used across all datasets in the study" (*Guide*, p.7).

Verification that all the data sets in a study comply with these data sets rules can be obtained using utilities that can be downloaded from the SAS Drug Development Forum.

UTILITY PROGRAM #4: CHECKING SPLIT DATA SETS

If a submission data set is legitimately large (that is, it does not have any bloated character fields), the FDA requests that it be included in both its original form *and* as split data sets. It might seem curious that the FDA wants both the large (greater than 1 GB) original data set and its constituent splits, since such a large original file can cause operational problems. However, without the original file, the FDA has no way to verify that the splits contain all the data from the original. After the original file is verified to be completely contained in the splits, the FDA does not need to send it to every location on its internal networks.

References to split data sets are located in several sections of the *Guide*. If you have a submission that requires split data sets, we recommend that you read all the related sections carefully. The SAS Drug Development Forum contains a downloadable utility program that checks whether a submission complies with the two rules for split data sets in Table 1.

CONCLUSION

We wrote this paper because we want as many pharma professionals as possible know about the FDA's *Study Data Technical Conformance Guide.* We provided a rudimentary inventory of the FDA's rules for submission data and indicated which ones are currently available in OpenCDISC's downloadable validation checks. In addition, we've developed basic utilities that check a submission for its compliance with some of the other rules (these utilities are available on the SAS Drug Development Forum).

There is still a great deal to do. In fact, attaining a fully automated, validated, and compliant submission is probably a mirage. The FDA itself "recognizes that it is impossible or impractical to define *a priori* all the relevant validation rules for any given submission" (*Guide*, p.26). Nevertheless, we can all contribute to incremental changes in our practices that will make the process of reviewing submission data go more smoothly.

REFERENCES

Gaffney, Alexander. "With PDUFA VI Negotiation Process Fast Approaching, BIO Takes Critical Look at Regulations." Regulatory Affairs Professional Society. August 6, 2014.
Available at http://www.raps.org/Regulatory-Focus/News/2014/08/06/19967/With-PDUFA-VI-Negotiation-Process-Fast-Approaching-BIO-Takes-Critical-Look-at-Regulations/

OpenCDISC. Available at http://www.opencdisc.org.

SAS Drug Development Forum. Available at https://communities.sas.com/community/support-communities/sas-drug-development.

U.S. Department of Health and Human Services, Food and Drug Administration, *Study Data Technical Conformance Guide*: *Technical Specifications Document*. December 2014. Available at http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm.

Wilson, Todd Allen. "Inside Health Policy - Industry Looks At 21st Century Cures To Set Stage For PDUFA VI." Friends of Cancer Research. December 19, 2014. Available at http://www.focr.org/news/inside-health-policy-industry-looks-21st-century-cures-set-stage-pdufa-vi.

ACKNOWLEDGMENTS

We would like to thank our colleagues in the SAS Life Sciences Consulting Group for their unflagging and inspiring desire to automate, and the SAS Solutions OnDemand Technical Communications group (especially Jensen Mabe) for their cheerful willingness to mold our draft into its final form.

CONTACT INFORMATION

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

Name: Ben Bocchicchio
Enterprise: SAS Institute, Inc.
Address: 3362 Building Q, 801 SAS Campus Dr.
City, State ZIP: Cary, NC 27513
Work Phone: 1-919-531-3704
E-mail: ben.bocchicchio@sas.com

Name: Frank Roediger
Enterprise: SAS Institute, Inc.
Address: 3358 Building Q, 801 SAS Campus Dr.
City, State ZIP: Cary, NC 27513
Work Phone: 1-919-531-0519
E-mail: frank.roediger@sas.com