Paper 855-2017

Preparing Analysis Data Model (ADaM) Data Sets and Related Files for FDA Submission with SAS®

Sandra Minjoe, Accenture Life Sciences; John Troxell, Accenture Life Sciences

ABSTRACT

This paper compiles information from documents produced by the U.S. Food and Drug Administration (FDA), the Clinical Data Interchange Standards Consortium (CDISC), and Computational Sciences Symposium (CSS) workgroups to identify what analysis data and other documentation is to be included in submissions and where it all needs to go. It not only describes requirements, but also includes recommendations for things that aren't so cut-and-dried. It focuses on the New Drug Application (NDA) submissions and a subset of Biologic License Application (BLA) submissions that are covered by the FDA binding guidance documents. Where applicable, SAS® tools are described and examples given.

INTRODUCTION

The purpose of this paper is to describe how to assemble analysis data and related files for the submission of NDAs and most BLAs to FDA CDER and CBER. The deliverables discussed are analysis datasets, other files related to analysis datasets, analysis programs, data definition files (define.xml) and the Analysis Data Reviewers Guide (ADRG).

The material included here is based on requirements described in the two December 2014 FDA Binding Guidance documents:

- Providing Regulatory Submissions in Electronic Format — Submissions Under Section 745A(a) of the Federal Food, Drug, and Cosmetic Act
- Providing Regulatory Submissions In Electronic Format — Standardized Study Data

Three other FDA documents that are related to these binding guidance documents and contain material relevant to this paper are:

- Data Standards Catalog v4.5.1 (08-31-2016)
- Study Data Technical Conformance Guide v3.2 (October 2016)
- Technical Rejection Criteria for Study Data (Revised 11142016)

Additional documents used to compile this paper are published by the Clinical Data Standards Interchange Consortium (CDISC), the Computational Sciences Symposium (CSS) workgroups, and the Japan Pharmaceuticals and Medical Devices Agency (PMDA).

The References section of this paper contains links to websites where all of these documents can be downloaded.

ANALYSIS DATA AND OTHER RELATED DATA

Let's begin by defining "analysis data" and other related data.

ANALYSIS DATASET DEFINTIONS

The Analysis Data Model Implementation Guide (ADaMIG) v1.1 defines three different types of datasets: analysis datasets, ADaM datasets, and non-ADaM analysis datasets:

> **Analysis dataset** – An analysis dataset is defined as a dataset used for analysis and reporting.
>
> **ADaM dataset** – An ADaM dataset is a particular type of analysis dataset that either:
>
> (1) is compliant with one of the ADaM defined structures and follows the ADaM fundamental principles; or

(2) follows the ADaM fundamental principles defined in the ADaM model document and adheres as closely as possible to the ADaMIG variable naming and other conventions.

Non-ADaM analysis dataset – A non-ADaM analysis dataset is an analysis dataset that is not an ADaM dataset. Examples of non-ADaM analysis datasets include:

- an analysis dataset created according to a legacy company standard
- an analysis dataset that does not follow the ADaM fundamental principles.

This same document includes a figure showing the relationships of these types of datasets:

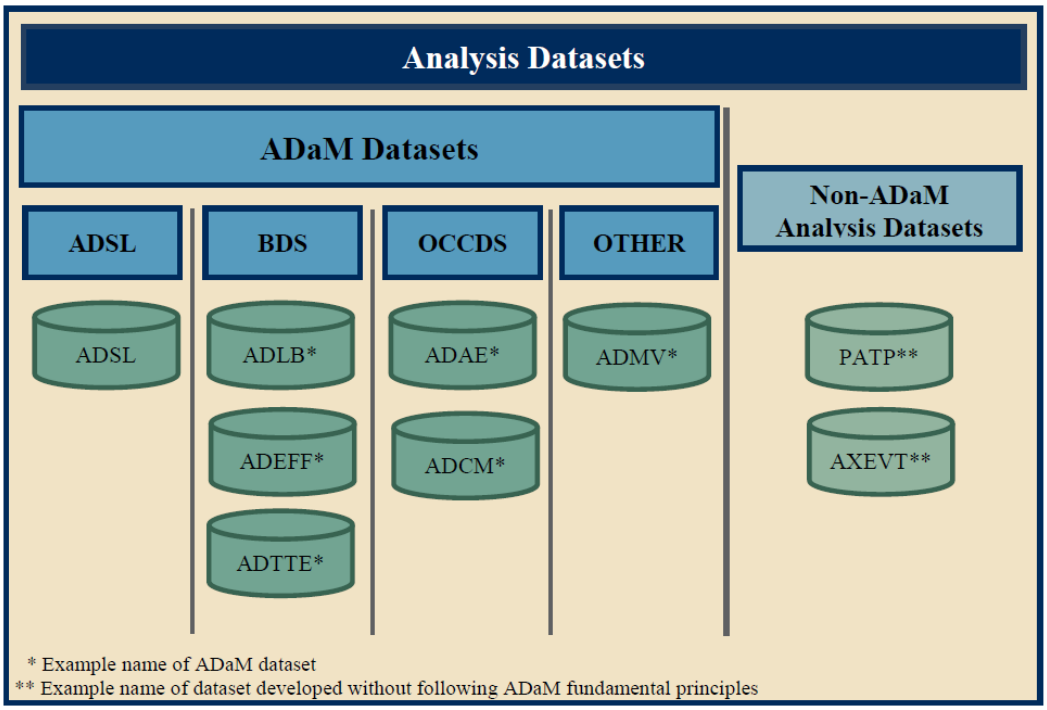

Figure 1: copy of "*ADaMIG v1.1 Figure 1.6.1 Categories of Analysis Datasets*"

Basically, an analysis dataset is either an ADaM dataset or a non-ADaM analysis dataset. There are three standard structural classes of ADaM datasets:

- ADSL (Subject-Level Analysis Dataset)
- BDS (Basic Data Structure)
- OCCDS (Occurrence Data Structure) if using ADaMIG v1.1; or ADAE (Adverse Event Analysis Dataset) if using ADaMIG v1.0.

Occasionally, there may be an analysis need which no standard structure can address. For example, no standard structure enables generation of a correlation matrix of time-varying dependent variables. In that case, the unmet analysis need can be addressed by designing a dataset with a non-standard structure. Such a dataset is an ADaM dataset only if follows all of the ADaM fundamental principles and other ADaM conventions. These true ADaM datasets that cannot follow a standard ADaM structure are considered to be members of the ADaM Other class of ADaM datasets.

A non-ADaM analysis dataset is any analysis dataset that is not compliant with ADaM. Non-ADaM analysis datasets are not broken down into structures or classes the way ADaM datasets are.

STANDARDS ACCEPTED BY FDA

The FDA Data Standards Catalog v4.5.1 (08-31-2016) lists all supported and required standards. For analysis data, the only standards included are ADaM v2.1 and ADaMIG v1.0.

The FDA Study Data Technical Conformance Guide (SDTCG) v3.2 states that they will also accept standards described in the following CDISC Therapeutic Area User Guides (TAUGs):

- Chronic Hepatitis C
- Dyslipidemia
- Diabetes
- QT Studies
- Tuberculosis

These TAUG standards are developed quickly and are often finalized before ADaM documents can be updated.

WHY DOES FDA WANT ADAM?

Standards are developed and used for many reasons, including to increase efficiency. The FDA SDTCG states that ADaM facilitates their review, simplifies programming steps necessary for performing analysis, and promotes traceability from analysis results to ADaM datasets to SDTM datasets.

Specifics not mentioned in the FDA documents are that reviewers have been receiving more and more ADaM data and are getting used to using it. They've also been provided training and tools to help them use this data in their reviews.

FORMAT OF DATASETS SUBMITTED TO THE FDA

The FDA SDTCG v3.2 states that the only way electronic datasets can be submitted to the FDA is in the file format of SAS Transport Format v5. These transport files can be created using SAS PROC COPY or in a DATA step. Native SAS datasets such as those with extension "sas7bdat", as well as transport files created using SAS PROC CPORT, are not accepted.

Although SAS PROC COPY allows multiple SAS datasets to be combined into a single transport file, FDA requires that for submission each SAS dataset be converted into a SAS transport file. Moreover, the name of the transport file must have the same name as the dataset. For example, adae.sas7bdat must be converted to adae.xpt.

Information Beyond the FDA Documents

The reason for the requirement of this "old" version of the SAS transport file is that SAS v5 transport is an open file format. In other words, data can be translated to and from SAS v5 transport and other commonly used formations without the use of programs from SAS Institute (or any other specific vendor).

Because the v5 file format is so old, it doesn't understand many of the newer features of SAS. In fact, this is the reason CDISC data standards such as SDTM and ADaM restrict dataset and variable names to 8 characters, dataset and variable labels to 40 characters, and character variable lengths to 200 characters or less. Longer versions of any of these items will be truncated and/or an error message will be generated when the transport file is created.

Also watch out for newer or user-specified SAS display formats. Any format that isn't known to SAS v5 transport will be lost when the transport file is created. For dates, this means displays of the date, such as via the SAS Viewer or PROC PRINT, will show the number of days since Jan 1, 1960, the underlying content of the date variable. For times and datetimes, this means that a number of seconds will appear. Only use formats that are standard in SAS V5.

To ensure that no data or formatting is lost when creating the SAS transport file, consider using a validation process such as:

(1) Create a SAS dataset

(2) Create a SAS v5 transport file from the SAS dataset using SAS PROC COPY or the DATA step. For example:

```
libname adam          "C:\desktop\data\adam";
libname xptfile xport "C:\desktop\data\xport\adsl.xpt";

data xptfile.adsl;
   set adam.adsl;
run;
```

(3) Convert the SAS v5 transport file into a new SAS dataset. For example:

```
libname xptfile xport "C:\desktop\data\xport\adsl.xpt";
libname new           "C:\desktop\data\new\";

data new.adsl;
   set xptfile.adsl;
run;
```

(4) Use SAS PROC COMPARE to compare the new dataset with the original version to check for discrepancies. For example:

```
libname adam "C:\desktop\data\adam";
libname new  "C:\desktop\data\new\";

proc compare base=adam.adsl compare=new.adsl printall;
   title "Comparison of adam.adsl (BASE) and new.adsl (COMPARE)";
run;
```

SIZE REQUIREMENTS FROM THE FDA

Another requirement found in the FDA SDTCG v3.2 is that the allotted length for each variable containing text be set to the maximum length needed by that variable. Artificially setting all text variables to a length of 200 makes the dataset much larger and more difficult for the reviewers to work with.

FDA SDTCG v3.2 sets the maximum size of a submitted dataset to 5 gigabytes (GB). Many different tools can be used to do a review, and not all of them can handle datasets larger than 5GB. Larger datasets must be split, and both versions (split and non-split) must be submitted. There is a separate directory to hold the split datasets.

Information Beyond the FDA Documents

There are at least two reasons why programmers may not set character variable lengths appropriately:

- Setting the variable lengths appropriately requires some effort, involving consideration of CDISC and sponsor standards as well as examination of collected and derived data values.
- Programmers, especially those with an Oracle background, may not be aware of how SAS allocates memory for character variables. In SAS, variable length 200 always uses 200 bytes of storage for that variable on every record, even if the actual data value on a record is only 1 character or null. In contrast, Oracle's VARCHAR200 data type allocates only as much storage as required by the actual data value on a given record.

When trimming SAS variable lengths to the minimum necessary to contain the maximum actual data values, it is best to look across all datasets rather than in only one dataset at a time. This is because data processing such as SET and MERGE statements can result in inadvertent truncation if lengths of

variables with the same name vary across datasets. Also, in some cases, it may pay to anticipate future uses such as data integration when setting variable lengths.

The FDA split rule described above was put in place to handle the CDISC Study Data Tabulation Model (SDTM) data requirement that all data of the same type be put into a single dataset. For example, all laboratory tests are required to be part of domain LB, even if that means the dataset will be larger than 5 GB.

In ADaM, there is no requirement that all data of the same type be put into a single dataset. Not only do smaller datasets not require splitting at submission time, they are nimbler and can reduce program run times. When it makes sense, consider creating multiple smaller, focused datasets rather than fewer large, cumbersome ones.

DATASET SUBMISSION LOCATION

The FDA SDTCG v3.2 includes this figure to show where to put all data and related submission items:

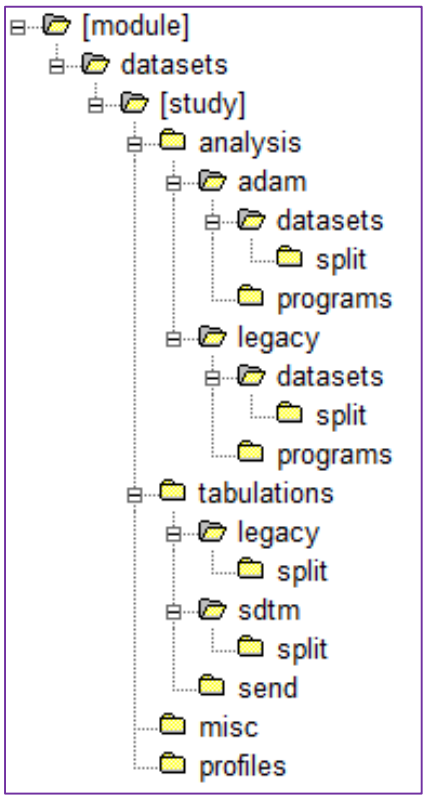

Figure Error! Use the Home tab to apply 0 to the text that you want to appear here.**2: copy of "*FDA SDTCG v3.2 Figure 1: Folder Structure for Study Datasets*"**

Additionally, ADaMIG v1.1 describes that for ease of use with the define file and in the eCTD folder structure, all analysis datasets for a study should be kept in a single folder, either **adam** or **legacy**, using the following rules:

- If a set of analysis datasets includes an ADaM-compliant ADSL dataset (as required for a CDISC-conformant submission), then the whole set of analysis datasets for that study belongs in the **adam** folder

- If not, the whole set of analysis datasets for that study belongs in the **legacy** folder.

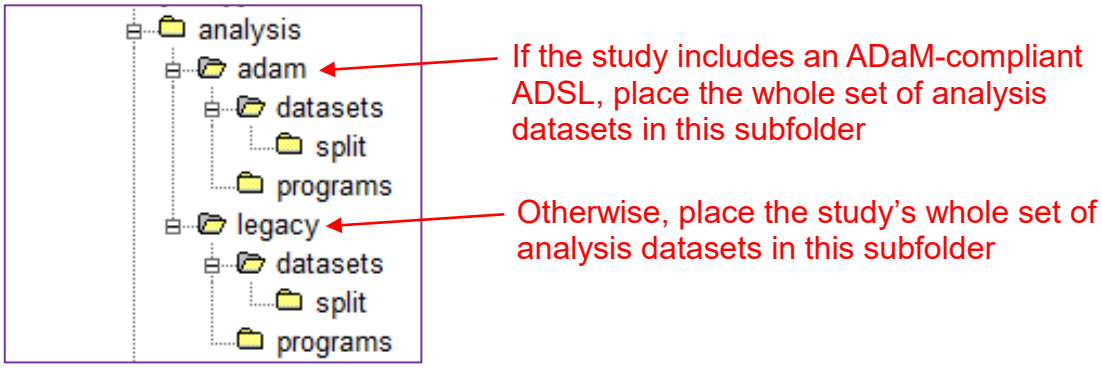

Figure Error! Use the Home tab to apply 0 to the text that you want to appear here.**3: Analysis Dataset Submission Folders**

Information Beyond the FDA and CDISC Documents

Although the FDA binding guidance documents say that ADaM is only required in NDA and BLA submissions for studies that start after December 17, 2016, ADaM can be submitted for other studies. Reviewers have tools and training to support the use of this standard, and it could theoretically speed up the time it takes for them to do their review.

FDA Data Standards Catalog (v4.5.1) does not yet include ADaMIG v1.1, only ADaM v2.1 and ADaMIG v1.0. As of this writing, FDA is evaluating ADaMIG v1.1 for use with their tools. In the interim, check with the relevant FDA reviewing division if you want to submit datasets following ADaMIG v1.1, because they may allow a waiver.

WHICH DATASETS TO CREATE AND SUBMIT

The CDISC ADaM standard requires ADSL. ADaMIG v1.1 states that it is up to the sponsor to determine what other analysis datasets are created.

The FDA Technical Rejection Criteria for Study Data document states that ADSL is required in the NDA and BLA submission for all studies starting after December 17, 2016. The FDA SDTCG states that sponsors should submit ADaM datasets to support key efficacy and safety analyses.

Information Beyond the FDA Documents

Based on the text in the FDA documents, a sponsor may choose not to submit any datasets other than ADSL and those used for key efficacy and safety analyses. This is risky, because a reviewer may ask for additional datasets during review. The sponsor would then need to submit quickly these additional datasets, and potentially slow down the review time. A safer solution is to discuss with the review division, perhaps at a pre-NDA or pre-BLA meeting, which datasets to include in the submission.

MISCELLANEOUS DATA

Figure 2 includes a folder called **misc**. The FDA SDTCG v3.2 specifies that miscellaneous datasets, which don't qualify as analysis, profile, or tabulation datasets, should be put in this folder.

Although not specified in the SDTCG, miscellaneous datasets would include any data not captured in SDTM but used to create ADaM datasets. Look-up tables, such as a list of prohibited concomitant medications, and deviations collected somewhere other than on the CRF are examples of this miscellaneous data.

ANALYSIS PROGRAMS

Recall that all the analysis datasets for a study are placed in either the **adam** or **legacy** datasets folder. Within each of these folders, at the same level as the **datasets** folder, is a **programs** folder. The FDA SDTCG states that the **programs** folder is where to put programs used to create analysis datasets, tables, and figures associated with primary and secondary efficacy.

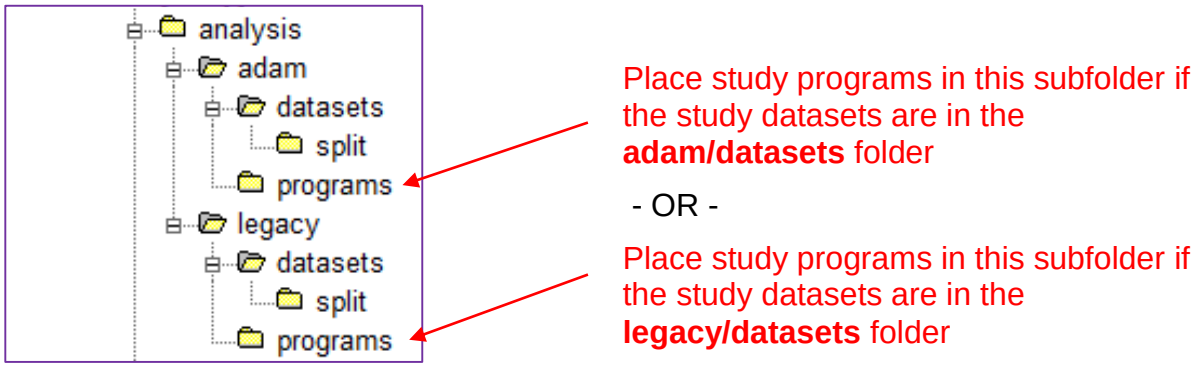

Figure Error! Use the Home tab to apply 0 to the text that you want to appear here.**4: Analysis Programs Submission Folders**

The FDA SDTCG document describes that the purpose of these programs is to understand the process and confirm analysis algorithms. This implies that that programs not expected to be run directly on the FDA system. The SDTCG requires that submitted programs to be ASCII text files (*.txt) or PDF files (*.pdf).

Information Beyond the FDA Documents

The practical impact may be illustrated with an example. When a SAS program called adtte.sas is prepared for submission, it would become adtte.txt or adtte.pdf. Some reviewers may take snippets of code to replicate the sponsor's analysis results and modify them to test alternate approaches.

Although not specifically stated in the FDA SDTCG, consider submitting at least all programs used to create the submitted datasets and key analyses. If not submitting all programs, be prepared to provide them for any FDA Reviewer requests.

To make submitted programs as easy as possible for FDA Reviewers to read and use, consider including robust comments and using non-macro language as much as possible. Also, it may not be necessary to include the table program code that put the results into specific places on the table. In other words, the program that was actually used to create the table may not be the program that is submitted.

It is worth noting that the Japanese PMDA regulatory agency has similar text in their Technical Conformance Guide. In addition, that PMDA document includes text about submission of full complex programs including macros: "…if submission of the macro program is difficult or submission of the program itself is difficult because the creation of the dataset or program was outsourced, the submission of specifications that show the analysis algorithm would be sufficient." Although the FDA SDTCG doesn't contain this text, it might be something to discuss with the relevant FDA reviewing division before blindly submitting complex and macro-driven programs.

DATA DEFINITION FILES (DEFINE.XML)

The data definition (define) file describes the metadata of submitted electronic datasets. The DSTCG states that the data definition file is "arguably the most important part of the electronic dataset submission for regulatory review". It also states that "An insufficiently documented data definition file is a common deficiency that reviewers have noted."

DEFINE CONTENT

CDISC has useful document packages on define.xml that can be downloaded for free. In addition to robust specifications, these document packages each include examples of how to lay out a define.xml file. The Analysis Results Metadata Specification v1.0 for Define-XML v2 (Jan 2015) contains examples and instructions for creating all the metadata needed for an analysis dataset submission:

- Dataset-level Metadata
- Variable-level Metadata
- Parameter Value-level Metadata, when appropriate
    - Note that Value-Level Metadata is essential for describing ADaM Basic Data Structure datasets containing metadata that vary according to analysis parameter
- Results-level Metadata (recommended for critical analyses)
- Controlled terminology and codes
- Links to other documents, such as
    - Statistical Analysis Plan (SAP)
    - Analysis Data Reviewers Guide (ADRG)

DEFINE VERSION

The FDA Data Standards Catalog v4.5.1 lists define.xml v1.0 and define.xml v2.0. The define.pdf is not included in the Data Standards Catalog v4.5.1, but it was a former standard and might be allowed via a waiver.

The DSTCG recommends using the standard define.xml v2.0. One reason for this recommendation is that version 2.0 allows printing of the define.xml file, something reviewers regularly need to do. Additionally, define.xml v1.0 only included dataset-level and variable-level metadata, because it was written before any of the current ADaM documents and designed specifically for the submission of SDTM data. The define.xml v2.0 added value-level metadata. The Analysis Results Metadata Specification v1.0 for Define-XML v2 (Jan 2015) added results-level metadata, and is the best option to accompany ADaM datasets.

SET OF DEFINE FILES

The define.xml file is very difficult to read in its native form, since it contains both textual content and XML code and symbols. It needs a stylesheet to allow the XML code to render properly for human consumption. CDISC has provided in their packages an example stylesheet that works across many browsers. It is not required that this CDISC-provided stylesheet be used; however doing so can help ensure that a submission reviewer will see the define in the layout that the sponsor intended.

A define.html and define.pdf may also be provided. The define.pdf can be useful for printing.

Below is an example of some typical define files. Note that they are shown here along with the ADaM datasets.

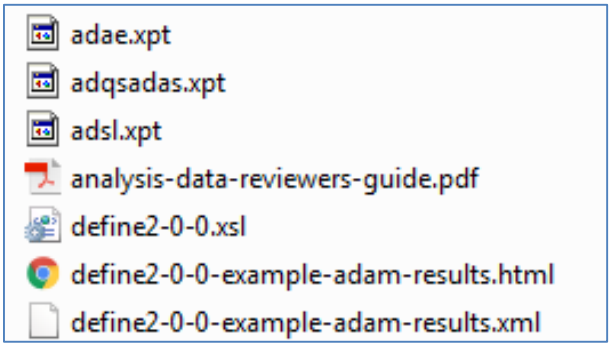

Figure 5: Example of Define Files

DEFINE SUBMISSION LOCATION

Because of technology constraints, links sometimes don't work when referencing material in a different folder. This means that each folder with datasets must have its own define file. For analysis data, this means the define file is located in the appropriate **datasets** folder:

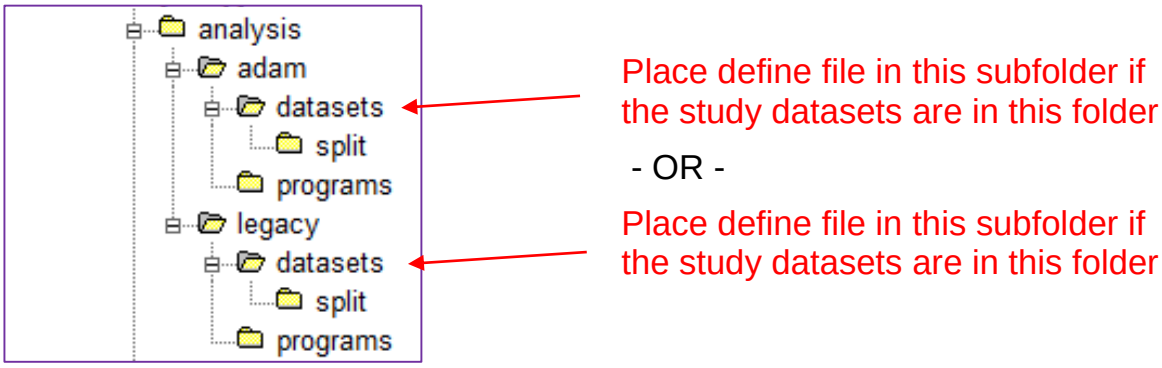

Figure Error! Use the Home tab to apply 0 to the text that you want to appear here.**6: Folder for**

ANALYSIS DATA REVIEWERS GUIDE (ADRG)

The Analysis Data Reviewers Guide (ADRG) is one of the newer components in submissions of analysis data.

ADRG PURPOSE

The introduction of the CSS ADRG Completion Guideline describes that the purpose of the submitted ADRG is to provide "FDA Reviewers with additional context for analysis datasets (AD) received as part of a regulatory submission." It goes on to state that the "ADRG purposefully duplicates limited information found in other submission documentation (e.g., the protocol, statistical analysis plan, clinical study report, define.xml) in order to provide FDA Reviewers with a single point of orientation to the analysis datasets." It also notes that "submission of a reviewer guide does not obviate the requirement to submit a complete and informative define.xml document to accompany the analysis datasets."

The DSTCG states "The ADRG provides FDA reviewers with context for analysis datasets and terminology, received as part of a regulatory product submission, additional to what is presented within the data definition file (i.e., define.xml)." and also "It should be noted that the submission of an ADRG

does not eliminate the requirement to submit a complete and informative define.xml file corresponding to the analysis datasets."

The Analysis Data Reviewers Guide (ADRG) package was created by the Computational Sciences Symposium (CSS). A zip file with a template, guidelines for completion, and examples can be downloaded from phusewiki.org, and the CDISC Analysis Results Metadata Specification v1.0 for Define-XML v2 also contains an example ADRG.

ADRG CONTENT

The ADRG is set up with standard sections and leading questions to prompt on what to say.

The section on **Dataset Processing** is a good place to explain any complex data flows. For example, the figure below shows the dependencies for a suite of ADaM datasets. Here ADAE, ADLB, and ADTR are used to create ADTTE; then ADTTE and ADBASE are used to create ADEFF:

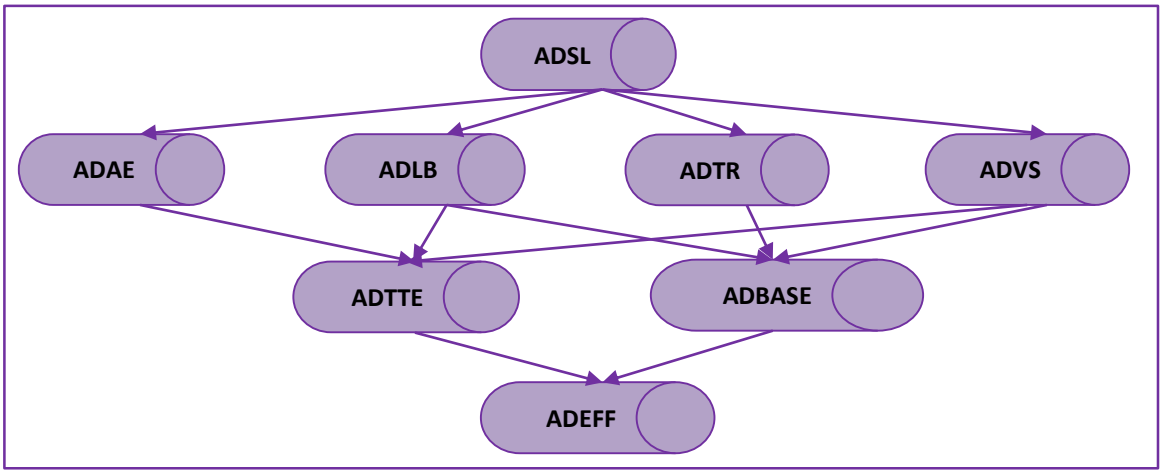

Figure 7: Complex Data Flow Diagram Example

The section on **Conformance** is the place to describe any conformance checks that were run, and explain any issues found.

ADRG SUBMISSION LOCATION

The DSTCG recommends that an ADRG be included as part of any analysis data submission. Like the define files, it is submitted in the same folder as the analysis datasets:

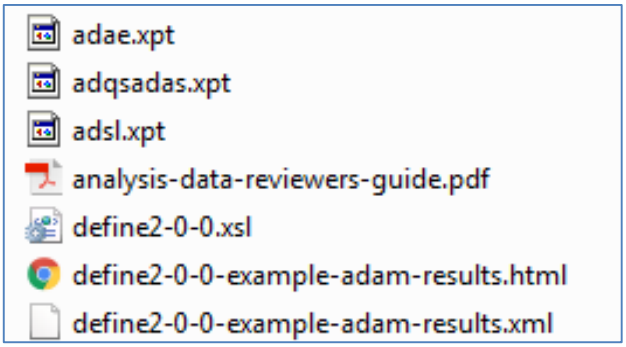

Figure 8: Example of a Folder with an ADRG File

SUMMARY

For ADaM data, the following figure summarizes what to submit where:

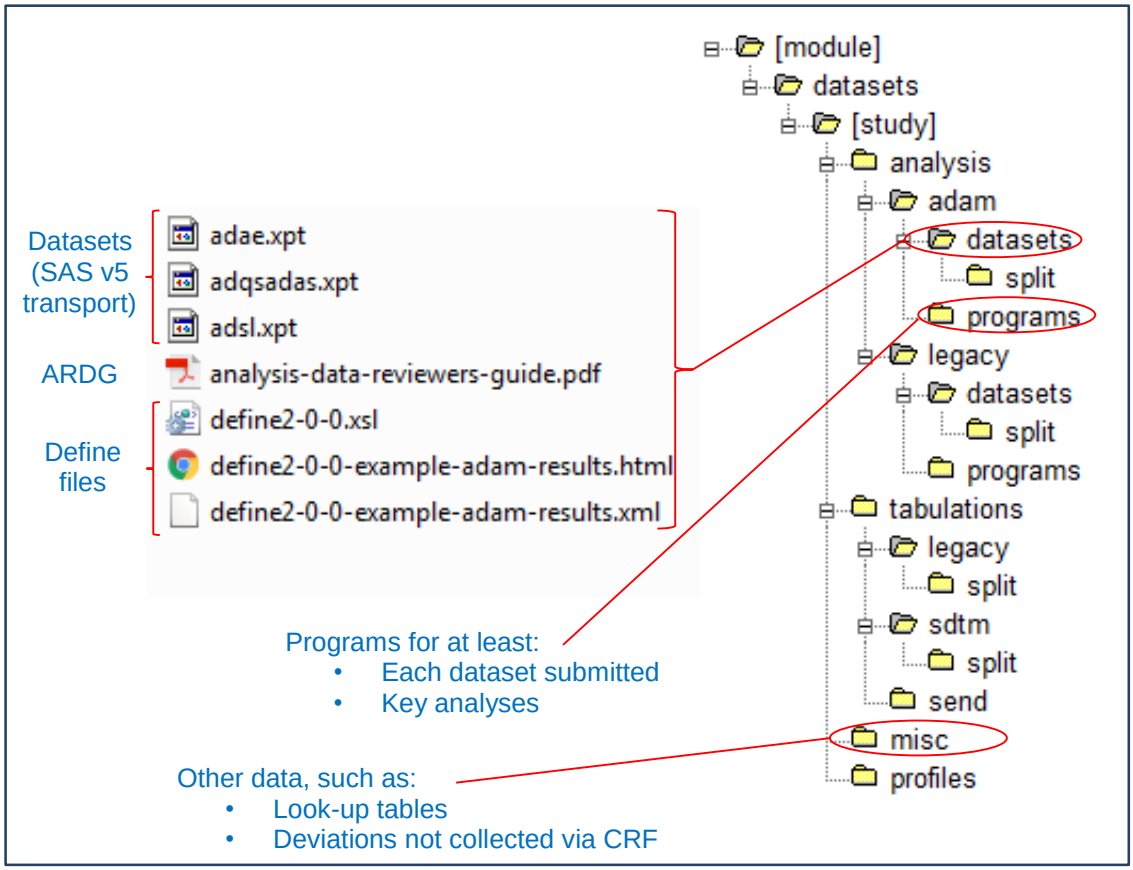

Figure 9: Summary of Submission Folder Locations and Content

The **datasets** folder holds not only ADaM data, but also the define files and the ADRG. Submit a SAS v5 transport file for each ADaM dataset, not the actual SAS datasets themselves. Include at least ADSL and datasets used for key analyses, as negotiated with the review division. Include at least define.xml and define.xsl.

The **programs** folder holds all submitted programs. Each program should be a text file (extension .txt) or a pdf (extension .pdf). Don't submit programs with extension .sas.

The **misc** folder holds data used to create ADaM that is not in the SDTM folders.

Take advantage of the reference documents from FDA, CDISC, CSS, and PMDA for additional details.

REFERENCES

United States Food and Drug Administration. 2017. "Study Data Standards Resources." Accessed January 30, 2017. http://www.fda.gov/ForIndustry/DataStandards/StudyDataStandards/default.htm.

- This site contains all the FDA documents referenced in this paper. It is also where you'll find email addresses to ask questions to CDER/CBER.

Clinical Data Interchange Standards Consortium. 2017. "Analysis Data Model (ADaM)." Accessed January 30, 2017. https://www.cdisc.org/standards/foundational/adam.

- This site contains all the CDISC ADaM documents, including the Analysis Results Metadata.

Clinical Data Implementation Standards Consortium. 2017. “Define-XML” Accessed January 30, 2017. https://www.cdisc.org/standards/foundational/define-xml.

- This site contains all the CDISC define.xml documents.

PhUSE wiki. 2017. “Optimizing the Use of Data Standards.” Accessed January 30, 2017. http://www.phusewiki.org/wiki/index.php?title=Optimizing_the_Use_of_Data_Standards.

- This site contains the ADRG package.

Japan Pharmaceuticals and Medical Devices Agency. 2017. “Notification No. 0427001”. Accessed February 25, 2017. https://www.pmda.go.jp/files/000206449.pdf.

- This site contains the English translation of the PMDA Technical Conformance Guide.

CONTACT INFORMATION

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

Sandra Minjoe
Accenture Life Sciences
sandra.minjoe@Accenture.com

John Troxell
Accenture Life Sciences
john.troxell@Accenture.com