Paper PO40

Some Statistical Programming Considerations for e-Submission

Hang Pang, Sanofi-Aventis Pharma. Inc., Bridgewater, NJ

ABSTRACT

Electronic submissions for new drug application (NDA) to the regulatory agencies have been widely used in the pharmaceutical industry. This paper discusses the main components that need to be prepared and communicated to the statistical programmers for the e-submission: SAS data transport (XPORT) files, data definition document (define.pdf), and statistical review aids. The FDA (Food and Drug Administration) and CDISC (Clinical Data Interchange Standards Consortium) guidelines for define.pdf are briefly introduced. Currently, no standard guideline is available for the statistical review aids. This paper presents some suggestions for what components should be included, and how to standardize and generate templates for high quality statistical review aids to meet the FDA requirements.

KEY WORDS: e-Submission, Define.pdf, Statistical Review Aids, SAS Data Transport Files

INTRODUCTION

Most pharmaceutical companies, when filing an NDA with the regulatory agencies, have used electronic submission. In addition to generating the derived analysis datasets and SAS outputs of TLGs (tables, listings and graphs), statistical programmers usually need to prepare the following main components for the e-submission: SAS data transport files, data definition document (define.pdf), and statistical review aids.

E-submission formats required for archiving by the agency are based on the type of files submitted. SAS datasets used to perform the analysis can be submitted in different formats. SAS transport (XPORT) files are currently accepted by the FDA. Other formats (e.g. XML) are also acceptable. An important part of preparing the XPROT files is addressing the data size limitations set by the FDA.

SAS output files are a second type of files found in the e-submission. The text and image files are usually submitted in PDF format, such as the tables, listings and graphs.

A third type of file submitted is the Data Definition Document, otherwise known as the define.pdf. There are two comparable structures for this file. In 1999 the FDA proposed a format of a define.pdf that consists of five columns (Variable Name, Variable Label, Type, Code and Comment). CDISC (Clinical Data Interchange Standards Consortium) proposed a format for define.pdf with seven columns (two additional columns Origin and Role). The CDISC structure is based on the SDTM (Study Data Tabulation Model)/SDS (Submission Data Standards) V3.1 (2004) models, and the Analysis Dataset Model (ADaM) V1.0. Both FDA and CDISC formats are acceptable by FDA.

A fourth type of file submitted is known as a Statistical Review Aid. In addition to the data, output and define.pdf, some e-submissions request statistical review aids. The purpose of the review aids is to allow the reviewer to easily generate SAS outputs for the primary endpoint and/or the major secondary efficacy endpoints. Currently, no standard guideline is available for the statistical review aids.

SAS DATA TRANSPORT FILES

SAS datasets can be submitted as SAS transport (XPORT) format, also known as version 5 SAS transport formats. The SAS transport (XPORT) files are open format published by SAS Inc., and can be created by PROC XCOPY in SAS version 5 and by the XPORT engine in SAS version 6 or higher.

Currently, the XPORT file is accepted by FDA, which can be adopted on different operating systems or different software. Other formats (e.g. XML) are also accepted by FDA. CDISC (2005) has recently published the define.xml.

The data file size limitation was originally set at 25 MB by FDA (Jan. 1999), however, the file size limit increased to <= 50 MB in 2003. Currently, 100 MB or less per file is also acceptable by FDA for some projects. The sponsor needs to contact the agency reviewers to confirm the maximum file size. It is recommended this communication with the agency occur prior to the data file preparation.

If the dataset size is over the limit, FDA guidance (2003) recommends splitting the dataset in columns. The most important variables will be in one table (file) and less important variables in another table (file). Each of the files needs to have the identifiers that can be used to link the files together as needed. Example of these identifiers may be study number, subject number, visit number, visit date, etc.

Existing FDA guidance documents have discouraged the use of formats catalog for user-defined formats and suggest that character variables with meaningful values be used instead. Analysis datasets should use these character variables where appropriate. However, there are some cases where a numeric version of a categorical variable is required for statistical purpose, e.g. a statistical model may require numeric 0/1 variable as indicators. For all formatted variables, decode variables can be added, which contain the formatted value prior to the creation of the XPORT file.

Since the FDA uses SAS version 5 transport files for all datasets, variable names should not be longer than 8 characters, and all character variables must be 200 characters or less.

Correcting data issues through hardcoding within the analysis dataset program should be avoided. Data should be cleaned at the source. However, should hardcoding be required, details around the hardcoding need to be part of the data specifications document (define.pdf) submitted to the agency.

SAS OUTPUT FILES

Statistical programmers also need to prepare SAS outputs for reporting (TLGs). Tables and listings can be generated and output in a number of different formats (LST files or RTF files for tables and listings, ESP files for graphics, etc.). These files can be easily converted to the PDF files for submission purposes.

DATA DEFINITION DOCUMENT

The FDA established the regulatory basis for electronic submission of data in 1997 with the publication of regulations on the use of electronic records in place of paper records (21 CFR Part 11). In 1999, the FDA standardized the file format (SAS version 5 transport files) for electronic submission of the clinical data collected in clinical trials with the first of series of guidance documents that describe the submission of clinical data and data definition files in PDF format (Define.pdf) or XML format (Define.xml).

CDISC (2005) has recently published the Case Report Tabulation Data Definition Specification (CRT-DDS, define.xml) version 1.0. To increase the level of automation and improve the efficiency of the Regulatory Review process, define.xml can be used to provide the Data Definition Document in a machine-readable format. For more information about the define.xml, you may check web site http://www.cdisc.org/models/def/v1.0/index.html.

The Data Definition Document (or define.pdf) contains two main parts: a Table of Contents (TOC) and a collection of Data Definition Tables (DDT). The TOC identifies each dataset that is submitted and its location within the submission. Whereas the Data Definition Tables provide detailed information about the variables contained within each of the datasets submitted and any derivations of variables within the dataset.

Example of Table of Contents (TOC) based on FDA format:

Datasets For Study 3001		
Dataset	**Description**	**Location**
AE	Adverse events	..\3001\data\ae.xpt
CM	Concomitant medication	..\3001\data\cm.xpt
DM	Demographics	..\3001\data\dm.xpt

Example of Data Definition Table (DDT) based on FDA format:

Study 3001 - AE.xpt, Adverse Event				
Variable	**Label**	**Type**	**Codes**	**Comments**
USUBJID	Unique Subject identification	Char		STUDY || CENTER || PATID
STUDY	Study number	Char		
CENTER	Study center	Char		
SEX	Sex of subject	Char	F = Female M = Male	

CDISC described the datasets into two classes: Study Data Tabulations (SDT) and the Analysis Datasets (AD).

- Study Data Tabulations (SDT): Datasets containing data collected during the study and organized by clinical domain. These datasets are described by CDISC SDTM/SDS V3.1 (June 2004).
- Analysis Datasets (AD): Datasets used for statistical analysis and reporting by the sponsor. These datasets are described by CDISC Analysis Dataset Model (ADaM) standards V1.0 (Dec. 2004).

Example of Table of Contents (TOC) based on CDISC format:

Datasets For Study 3001					
Dataset	**Description**	**Structure**	**Purpose**	**Key variables**	**Location**
AE	Adverse events	One record per subject, per event	Tabulation	USUBJID, AETERM, AESEQ	..\3001\data\ae.xpt
CM	Concomitant medication	One record per subject, per medication, per instance	Tabulation	USUBJID, ODCD, REPNO	..\3001\data\cm.xpt
DM	Demographics	One record per subject	Tabulation	USUBJID	..\3001\data\dm.xpt

Example of Data Definition Table (DDT) based on CDISC format:

Study 3001 - AE.xpt, Adverse Event, Feb. 10, 2005, One record per subject, per event						
Variable Name	**Variable Label**	**Type**	**Controlled Terms or Format**	**Origin**	**Role**	**Comments**
USUBJID	Unique subject identifier	Char		Sponsor Defined	Identifier	STUDYID || INVID || SUBJID
STUDTID	Study number	Char		(CRF) ae.studyid	Identifier	
INVID	Investigator identifier	Char		(CRF) ae.invid	Qualifier	
SEX	Sex	Char	M,F,U	(CRF) dm.sex	Qualifier	

Variables in the dataset need to be ordered. The variables whose roles are defined as either “Identifier”, “Selection”, followed by project common variables are listed first, all remaining variables are then listed in alphabetical order. The variables list in the define.pdf should be in the same order as they appear in the dataset.

The Role column determines the type of information conveyed by the variables, which can be classified into the following major roles in the SDTM/SDS V3.1 (June 2004):

- Identifier variables: identify the study, the subject, the domain, and the sequence number of the record.
- Topic variables: specify the focus of the observation (e.g. name of a lab test);
- Timing variables: describe the timing of the observation (e.g. start date and end date);
- Qualifier variables: include additional illustrative test, or numeric values that describe the results or additional traits of the observation (e.g. units). The list of Qualifier variables will vary depending on the type of observation and the specific domain.

The CDISC SDTM/SDS V3.1 roles defined above should also be used where appropriate in Analysis Datasets (AD). Additional AD roles defined in ADaM V1.0 (Dec. 2004) include:

- Selection variables: Variables that are frequently used to subset, sort, or group data for reporting purposes: e.g. Treatment, Age, Gender, and Race...
- Analysis variables: Variables tabulated and or summarized for analysis purposes.
- Support variables: Provide useful background or reference information.

There are some macros available to generate the define.pdf. Following are suggestions for creating or choosing an appropriate macro:

- Checks for data structure: Check the data structure, and make sure all datasets and variables are labeled. The length of variable name should be <= 8. Otherwise, the XPT file will not be created.
- Allows for columns to store both a short and long variable name: In the new version of CDISC SDS V3.1, there are two columns for the variable names: short name and long name. Short name should include <= 8 character field name. The <= 8 character limitation is currently required due to the limitation of the SAS V5 transport format currently required by FDA. A more flexible format is expected to replace this in the near future. The long name of variable may correspond to the variable name from the operational database or a long name generated from SAS version 6 or higher.
- Simultaneous file creation: The macro can generate Define.pdf and XPT files for all the datasets at the same time. This will help to ensure the data files created are current with the specifications defined. Additionally, it will allow for more efficient processing.
- Flexibility in creation of define.pdf: Allow for both the FDA and CDISC formats. The user has an alternative to choose one or the other.
- Addition of comments from external files: The “Comments” column could be easily added. A TEXT file or .CSV file with a special sign as the delimiter is recommended. The comma (‘,’) should not be used as the delimiter, because this causes problems if the text includes a comma.

- Proper pagination: When using ODS RTF to generate the define.pdf, caution is required because there is a bug for ODS RTF in SAS V8.2. If the Code or Comment is too long, it will automatically go to the next page, and the text will be truncated. SAS V9 may fix this bug of ODS RTF. Using ODS HTML will avoid the truncation problem. First, using ODS HTML generate the define.htm file, then convert it to WORD document or define.pdf file in MS WORD.

STATISTICAL REVIEW AIDS

After the e-submission, the agency review and NDA approval is highly dependent on the data within the submission. The more confidence the reviewers have in the data, the less time they will spend trying to validate it in the review process. In order to build reviewers confidence in the submitted data, the sponsor needs to provide the data in a structure that is standard, easy to understand, and has appropriate documentation.

The reviewing statisticians usually perform the following tasks. 1) Replicate or verify the sponsor's analysis, results, and conclusions, with particular emphasis on the primary and secondary efficacy endpoints; 2) Test the validity and robustness of the sponsor's analyses and assumptions.

The statistical review aids help the FDA reviewer get the key results quickly, e.g. the primary efficacy endpoint and major secondary efficacy endpoints. Components of the statistical review aids include:
1) Data XPT file: Only include the variables needed in the model (eff.xpt).
2) Programs and statistical models for the primary efficacy analysis (eff.sas).
3) Table for the primary efficacy endpoint in the clinical study report (CSR) (eff_table.pdf).
4) Key efficacy analysis specifications (eff.pdf): A text description documenting the analysis performed. It can reference the statistical methodology used, assumptions made, population used and other descriptive data, which includes:

a) Project/Study name.
b) Table name and title of the table: An unique identifier for this analysis.
c) Clinical endpoints: Show the reason for performing this analysis. For example, primary endpoint, secondary endpoints, exploratory analysis, subgroup…
d) Analysis dataset and location: The name of the analysis dataset used for the statistical review aids. In most of cases, this will be a single derived dataset.
e) Contents of the analysis dataset.
f) Decodes for the formats, e.g. treatment group, population variable, and visit…
g) Sample program code: The program that uses the analysis dataset to generate the analysis results, not the programs that create the analysis datasets.

A sample of statistical review aids specifications:

Primary Efficacy Analysis Specifications For Statistical Review Aids

Project: Proj1234
Study: 3001
Table: ef_0001t
Title: Analysis of change from baseline in primary endpoint at week 12 – ITT population
Clinical Endpoint: Primary efficacy endpoint
Analysis Dataset and Location: /../eff.xpt

Variables needed for this table as listed below are all from dataset eff.xpt

#	Variable	Type	Len	Pos	Format	Informat	Label
4	AGE	Num	8	16	4.	4.	Informed consent age
3	EFFTYPE	Char	2	116	$EFFTYPE.		Efficacy type (Raw, Imputed)
5	BLVA	Num	8	40			Baseline value
6	CHBL	Num	8	48			Change from baseline
7	SEX	Char	3	118	$SEXS.		Sex
11	ITT	Num	3	128			Intent-to-treat population
8	POOL	Char	7	121			Center pooling
12	PP	Num	3	131			Per protocol population
9	STRATA	Num	8	80	8.	8.	Derived stratum
2	TIMEPT	Num	8	0	TIMEPT.	8.2	Reporting visit number
10	TRTR	Num	8	88			Randomized treatment
1	USUBID	Char	12	104			Derived unique subject ID

Statistical Analysis Procedures:

```
/************************************************************************
*                          Decodes/Formats                              *
*                                                                       *
* TRTR:   0 = Placebo                                                   *
*         1 = Active 10 ug/day                                          *
*         2 = Active 20 ug/day                                          *
*         3 = Active 30 ug/day                                          *
*                                                                       *
* ITT: 1=True, 0=False for the ITT population                           *
* PP:  1=True, 0=False for the per-protocol population                  *
*                                                                       *
* EFFTYPE: I = Imputed data for last observation carried forward (LOCF) *
*          R = Raw data                                                 *
*                                                                       *
* TIMEPT:    0 = Baseline                                               *
*            4 = Week 1 / Visit 4                                       *
*            5 = Week 2 / Visit 5                                       *
*            6 = Week 4 / Visit 6                                       *
*            7 = Week 8 / Visit 7                                       *
*            8 = Week 12 / Visit 8 or Early Termination Visit (LOCF)    *
*          888 = Early Termination                                      *
*          999 = Final Visit Completer                                  *
*                                                                       *
* STARTA:  1 = Previous therapy 1                                       *
*          2 = Previous therapy 2                                       *
*                                                                       *
************************************************************************/

/*************************************************
*                    Beginning of Code           *
*************************************************/

** Allow program to run without format catalog **;
options nofmterr;

** Assign libname for sas data sets **;
libname ads "c:\temp\ads";

** Load data from the transport file to the working library **;
filename indat "c:\temp\eff.xpt";

proc cimport lib=ads file=indat;
run;

** Subset data to support week 12 and LOCF data **;
proc sort data=ads.eff out=final;
  by trtr usubid;

  ** ITT population and the time point **;
  where itt eq 1 and timept eq 8 and efftype eq 'I';
run;

title "Proj1234/3001: SAS outputs for the primary efficacy endpoint";

** Produce baseline mean for table ef_0001t **;
proc means data=final(where=(chbl>.z)) n mean;
  var blva;
  class trtr;
  output n=N mean=BLMean;
run;

** Produce inferential statistics & p-value **;
proc mixed data=final;
  class trtr pool strata sex;
```

```
  model chbl = trtr pool strata sex blva age;

  ** LS Means Differences, 95% CI, and p-value for Comparisons **;
  estimate "Active 10 ug/day - Placebo" trtr          -1 1 0 0 / cl;
  estimate "Active 20 ug/day - Placebo" trtr          -1 0 1 0;
  estimate "Active 30 ug/day - Placebo" trtr          -1 0 0 1;
  estimate "Active 20 ug/day - Active 10 ug/day" trtr 0 -1 1 0;
  estimate "Active 30 ug/day - Active 10 ug/day" trtr 0 -1 0 1;
  estimate "Active 30 ug/day - Active 20 ug/day" trtr 0 0 -1 1;

  ** LS Means & 95% CI for Treatment **;
  lsmeans trtr / cl alpha=0.05;
run;

/*********************************************************
*                        End of Code                     *
*********************************************************/
```

CONCLUSION

The analysis datasets submitted as SAS transport files should be directly usable by reviewing statisticians with little or no data manipulation and programming. The datasets should be compatible with currently available statistical software and compatible with future standard software tools that may be developed. The standards and models by FDA or CDISC for the submitted datasets are extremely important for efficient e-submission.

When submitting clinical data to the FDA, the data definition document or define.pdf is probably the most important documentation of the datasets. It will clearly provide descriptions of the data contents, derivation and usage of the data. Manual creation of this document can be very labor intensive and error prone. Since most information necessary is from the datasets themselves, an automation of this task seems reasonable. There are some macros available for this purpose, this paper has highlighted the areas where particular attention must be made when generating the define.pdf.

For the e-submission, following the FDA or CDISC guidelines based on the industry standards is paramount. Statistical review aids that allows FDA reviewer to work easier and faster is also important. Development of document templates based on previous submission examples may be helpful to the new project teams creating the define.pdf and statistical review aids.

REFERENCES

FDA (2001). Example NDA Submission (08MAR2001), http://www.fda.gov/cder/guidance/NDA_Example.htm
FDA (2003). Electronic Regulatory Submissions and Review, http://www.fda.gov/cder/regulatory/ersr/default.htm
CDISC (2004). SDTM/SDS V3.1 (25JUN2004), http://www.cdisc.org/models/sds/v3.1/index.html
CDISC (2004). ADaM statistical model V1.0 (DEC2004), http://www.cdisc.org/models/adam/V1.0/index.html
CDISC (2005). Case Report Tabulation Data Definition Specification (CRT-DDS, define.xml V1.0, 10FEB2005), http://www.cdisc.org/models/def/v1.0/index.html
Dave Christiansen and Stephen Wilson (2004). Submission of analysis datasets and documentation: scientific and regulatory perspectives. PharmSUG2004, FC04

ACKNOWLEDGMENTS

The author would like to thank Mary McKenna (Sanofi-Aventis Pharma. Inc.) for her review and comments.

CONTACT INFORMATION

Hang Pang
Sanofi-Aventis Pharma. Inc.
BX2-300E, 200 Crossing Boulevard
Bridgewater, NJ 08807-0890
Work Phone: 908-231-4187
Email: hang.pang@sanofi-aventis.com