PhUSE 2007

Paper RA11

Programmers Need to Know about Electronic Submissions

Simon Dundas, Pfizer Ltd, Sandwich, UK

ABSTRACT

It is essential that programmers and system developers working on the generation of study tables understand how their work impacts the compilation of an Electronic Submission (ESUB) to FDA. Datasets are a major part of an NDA or supplemental filing - as important as study tables. The datasets and the associated documentation must be accurate and user friendly. Processes for study table production should equally be aimed at producing suitable ESUB components as it is inefficient and difficult to retrospectively adjust datasets and documentation.

Programmers need to be aware of the current FDA guidances and be involved with discussions as to what is to be provided in each Electronic Submission (ESUB). They should imagine what it's like for the FDA review team, with tight timelines, trying to use the datasets to reproduce analyses or perform new ones.

INTRODUCTION

FDA used to receive an NDA in the form of one or more van loads of paper so inevitably there has been a move towards a paperless submission. They also like to do their own analyses so ask for user friendly datasets and associated documentation. Whereas the EMEA focus more on the submitted summary tables and reports, FDA places more emphasis on doing for themselves the analyses they believe are needed to demonstrate the drugs safety and efficacy. For these reasons **the datasets and ESUB components warrant as much attention as tables**. The aim of this paper is to inform programmers and system developers working on the generation of study tables as to how their work impacts the compilation of an Electronic Submission to FDA. Being better informed they will be able to save time setting up an ESUB, improve quality, aid usage by FDA and reduce the number of regulatory queries. All parties including the programmers will benefit.

The main purpose of the ESUB structure is to

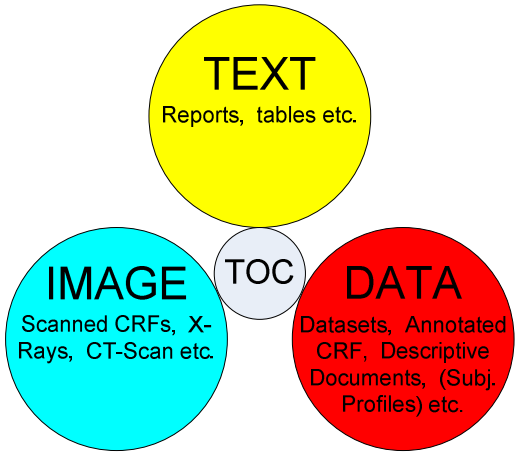
- Move towards a paperless submission
- Aid data searching. With the large amounts of data submitted, datasets can permit faster and more complex searching than using study listings.
- Standardise what FDA get from each customer
- Enable FDA to quickly understand the data structure and be able to do their own analyses.

At present, apart from a few cases, the guidance tends to focus more on the content of dataset documentation rather than dataset structure but as CDISC (Clinical data Interchange Standards Consortium) standards are implemented the latter is being addressed.

A background to what goes into an ESUB and the regulations is provided below in order to understand the subsequent sections on the the importance of the programmer's role.

OVERVIEW OF ESUB STRUCTURE

An ESUB typically has three main components linked into the main Table Of Contents as represented by the following diagram:



Study programmers generate the study tables part of the text component but play a major role in the data component . It should be noted that other agencies can ask for datasets but most do not. They may have their own guidance's on structure and content. The FDA guidances are very helpful but there is flexibility so the exact components should be agreed with FDA in advance of the submission. In this way the submission will meet their expectations and time won't be wasted producing unnecessary items or doing rework.

SCOPE OF THE DATA COMPONENT

The deliverables can vary from submission to submission and should be agreed with FDA but this is a guide:

1. Datasets provided as SAS® version 5 Transport Files can include:
    - All data* for Pivotal phase 2/3 studies
    - PK data for phase 1 studies – concentration and parameter
    - Population PK to support reports - may include NONMEM files
    - Pooled ISS/ISE data
    - PKPD to support reports
    - Carcinogenicity data from toxicology studies
    - Pharmacogenomic data
    - Any Datasets prescribed by FDA**
2. Blank Case Report Form annotated with corresponding dataset variables. Only raw variables given to FDA within datasets are annotated (a lack of submissions can indicate some data is missing from the submission)
3. Dataset Definition Document in PDF format containing details dataset / variable information
4. Subject Profiles for certain studies
5. Less common but these items may be requested:
    - SAS programs
    - Program flow charts
    - Annotated tables – indicating the source variables

* 'all data' means all raw data collected in the study and any derived and dictionary variables that helped with the reporting and so could be helpful to FDA. Derived variables enable repetition of submitted analyses and checking of accuracy of coding. It is also helpful to the sponsor as FDA are more likely to be able to understand and reproduce the analysis.
** FDA reviewers may provide a detailed specification for additional analysis datasets.

AVAILABLE INFORMATION

- FDA Guidance for Industry: Providing Regulatory Submissions in Electronic Format - General Considerations / NDA's. This has the basic rules for creating the ESUB data components – SAS transport files, dataset definition document , annotated Case Report Form and reference to Subject Profiles. This is available at http://www.fda.gov/cder/regulatory/ersr/default.htm

- FDA Guidance for Industry Antiviral Product Development — Conducting and Submitting Virology Studies to the Agency (June 2006.) The attachment has very detailed description of an FDA specific dataset for virology and available at
  http://www.fda.gov/OHRMS/DOCKETS/98fr/05d-0183-gdl0002-02-hiv-resistance.pdf

- 21CFR Part 11 - Covers electronic systems and the clients responsibility to ensure they are dependable and so governs the production of datasets and any system used to produce the ESUB.

- CDISC (Clinical Data Interchange Standards Consortium) Contains details of desired dataset structures agreed by the industry as a whole , including FDA. Develops standards for case report tabulations and analysis datasets.
  www.cdisc.org

- ICH eCTD (Common Technical Document) Covers how all the ESUB components fit together so as to standardise the format of the ESUB across agencies.
  http://www.fda.gov/cder/regulatory/ersr/ectd.htm

- FDA may add new guidance documents, delete or update existing ones so their web site should be scanned prior to working on an ESUB

Guidances can be very detailed. Here is an extract from the HIV resistance data specification mentioned above:

> *'Sponsors are encouraged to use the following sample format for submitting HIV resistance data. One dataset combines patient data, endpoint data, genotypic data, and phenotypic data. There are a number of ways datasets can be subdivided (i.e., by clinical study, baseline isolates, or virologic failure isolates) and this should be discussed with the division before submission. For each study, we recommend constructing datasets as SAS transport files containing the following information:*
>
> - *One record (row) per patient per isolate (e.g., baseline, failure, and other time points).*
> - *Data in columns (with suggested column headings shown below) on all isolates.*
> - *Genotypic data should be provided on the corresponding record for each patient isolate for baseline isolates of all patients in treatment-experienced studies and the endpoint isolates of virologic failures and discontinuations in all studies. In treatment-naïve studies, a baseline sample should be collected and stored from all patients for future phenotypic and genotypic analysis of virologic failures.*
> - *Phenotypic data should be provided on the corresponding record for each patient isolate for baseline isolates and the endpoint isolates of virologic failures and discontinuations. In treatment-experienced studies, it is recommended that baseline phenotypic data be obtained for all patients.*

The specific criteria for defining virologic failures should be discussed with the division and may include multiple primary and secondary protocol endpoints. The endpoints for clinical virologic and resistance outcome analyses should be consistent

From this you can see how knowledge up-front can save huge effort as the dataset could be designed with FDA in mind to start with rather than just at ESUB time.

HOW DO FDA USE THE DATA?

There are two factions of the FDA that are interested in clinical trials data, the archive group and the review divisions. The rules for the format of an ESUB are laid down by the archive division. Not only do they want reviewers to be given a standard set of deliverables but also to enable the data to be archived in a standard format to aid data mining or pooling in the future – should it be required.

The FDA have training for their reviewers in how to use the datasets (actually SAS transport files) from an ESUB. This is available for all to see at
http://www.fda.gov/cder/regulatory/ersr/#Reviewer%20Training%20Manuals

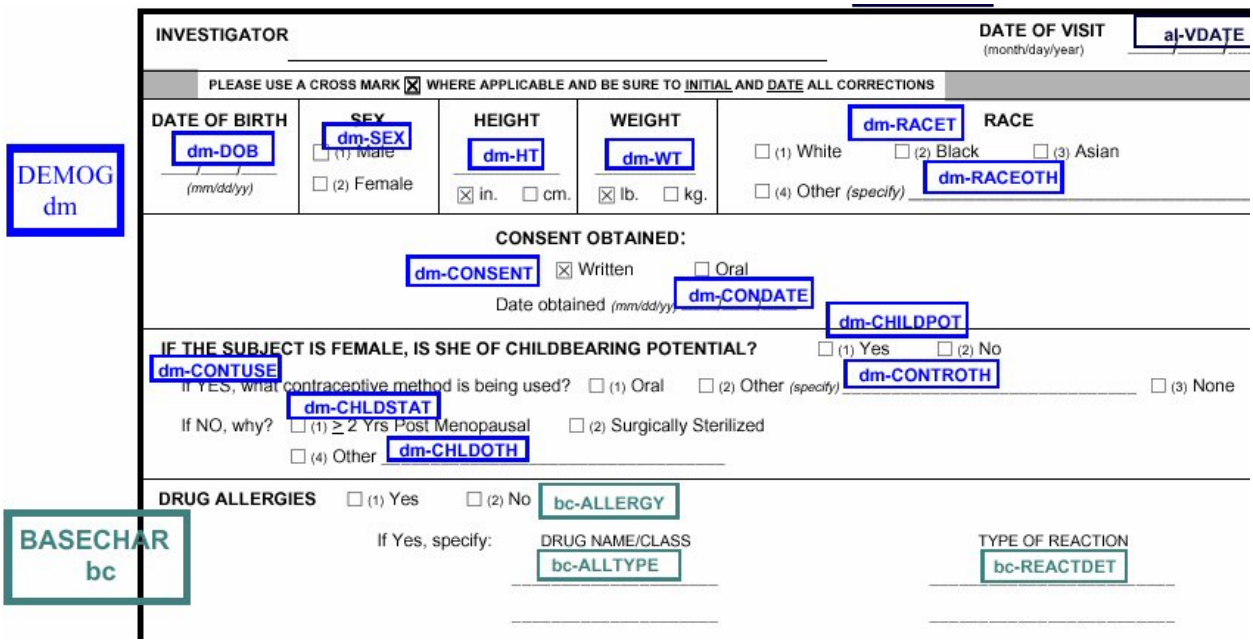
The training incorporates the use of the SAS System Viewer, Stat/Transfer, and JMP to view and manipulate the data and perform basic analysis. Additional introduction to JMP courses are also available that discuss analysis of adverse events, exposure, efficacy, lab, and demographic data. SAS JMP can quickly produce Excel spreadsheets.

FDA reviewers have statisticians to do some analysis but may not have much other programming support so may

elect for the sponsor to perform specified analyses. Analysis may seek to independently perform end-point and exploratory analysis. The sponsor will not know the detail, and so needs to produce user friendly datasets with clear documentation to enable the user to select appropriate variables and understand the data points.

WHAT DO THE ESUB COMPONENTS LOOK LIKE?

This is part of an Annotated Case Report Form which indicaties the source of the raw dataset variables:



FDA can also tell from this if any CRF data is not being provided so it is important the ESUB team checks this.

And here is part of a Dataset Definition Document to indicate the table of contents and variable details:

Dataset Definition Document for Study A9991001

Dataset Name	Description	Location
ADVERSE	Adverse Events. One observation per event.	ADVERSE.XPT
BASEOTH	Other Baseline Characteristics: Smoking and Alcohol Consumption. One observation per subject	BASEOTH.XPT
DEMOG	Demographics. One observation per subject	DEMOG.XPT
EVAL	Subject Evaluation. One observation per subject	EVAL.XPT
LAB1	Laboratory Data. Observation per subject/test/visit/time point /sample. Part 1	LAB1.XPT
LAB2	Laboratory Data. Observation per subject/test/visit/time point /sample. Part 2	LAB2.XPT

Variables for Dataset ADVERSE

Variable	Label	Type	Format	Codes	Comments
AEACTC	Course of action with this AE (code)	NUM			
AEACTT	Course of action with this AE (text)	CHAR			CRF page 42
AEACTXT	Other course of action	CHAR			CRF page 42

	with this AE (text)				
AEFDAY	Start day of AE	Num			Adverse event start date – date of first dose +1
SAE	Flag indicating if event is serious		YESNO	1=Yes 2=No	

The SAS transport files are version 5 so there is a limit of 40 characters for the SAS label. The comment field can be used to expand on the label and give further detail to aid understanding of the variables use and derivation. It is possible to reference other documents being supplied such as the Statistical Analysis Plan or else add an appendix.

THE KEY ROLE PLAYED BY THE CLINICAL PROGRAMMER

The programmer will have a large impact on the usability of the submitted datasets and documentation. Inadequate datasets and documentation could annoy the reviewer, slow down the approval process, lead to queries, cast doubt on the integrity of the filing and even result In a refusal to file.

The key items they can impact are :

- Determining which datasets to submit
- Determining the dataset structure
- Choosing variables to keep
- Creating dataset documentation suitable for extraction into the ESUB.
- User friendliness/donsistency
- Planning for efficiency

DETERMINING WHICH DATASETS TO SUBMIT

To submit all the datasets generated in the study reporting process could well be confusing and may contain out of date information and so lead to queries. Since FDA like to have all the raw data collected in the study and additional variables that could help with the data analysis then there is likely to be a need for a mixture of raw and value added (VA) datasets. If a VA dataset contains all the raw data then it will not be necessary to supply to raw dataset as well.

DETERMINING THE DATASET STRUCTURE

Vertical versus horizontal? Currently it is up to the sponsor but if CDISC is adopted then a vertical structure may be imposed. If for each coded variable there is a similarly named variable containing the decode then the user has a choice as to which to use and there is the added benefit that it is unnecessary to supply extra decode information.

SAS labels added at source can save time. The order of variables and having sensible variable lengths can help. Dataset names should indicate content. Variable names should be no more than 8 characters, reflect content and be standardised to indicate code/decode pairs. Datasets over 100MB need to be split in an ESUB so ideally names should allow for additional digits to be added indicating the split.

Here is an extract of part of the lab data specification from the CDISC Study Data Tabulation Model to indicate the detail involved. Codelists have or will be defined and so to submit an ESUB in CDISC format mapping would be required unless it is adopted as the company standard.

CDISC SDTM Implementation Guide (Version 3.1.2)

6.3.3 LABORATORY TEST RESULTS — LB

lb.xpt, Labs — Findings, Version 3.1.2, July 25, 2007. One record per lab test per time point per visit per subject, Tabulation

Variable Name	Variable Label	Type	Controlled Terms, Codelist or Format	Role	CDISC Notes	Core
STUDYID	Study Identifier	Char		Identifier	Unique identifier for a study.	Req
DOMAIN	Domain Abbreviation	Char	LB	Identifier	Two-character abbreviation for the domain.	Req
USUBJID	Unique Subject Identifier	Char		Identifier	Identifier used to uniquely identify a subject across all studies for all applications or submissions involving the product.	Req
LBSEQ	Sequence Number	Num		Identifier	Sequence Number given to ensure uniqueness of subject records within a domain. May be any valid number.	Req
LBGRPID	Group ID	Char		Identifier	Used to tie together a block of related records in a single domain for a subject.	Perm
LBREFID	Specimen ID	Char		Identifier	Internal or external specimen identifier. Example: Specimen ID.	Perm
LBSPID	Sponsor-Defined Identifier	Char		Identifier	Sponsor-defined reference number. Perhaps pre-printed on the CRF as an explicit line identifier or defined in the sponsor's operational database. Example: Line number on the Lab page.	Perm
LBTESTCD	Lab Test or Examination Short Name	Char	(LBTESTCD)	Topic	Short name of the measurement, test, or examination described in LBTEST. It can be used as a column name when converting a dataset from a vertical to a horizontal format. The value in LBTESTCD cannot be longer than 8 characters, nor can it start with a number (e.g.'1TEST'). LBTESTCD cannot contain characters other than letters, numbers, or underscores. Examples: ALT, LDH.	Req

VARIABLES TO KEEP

To mimimise rework of datasets for submitting in an ESUB and aid clarity the programmer should keep only raw variables and key derived variables. Intermediate variables, those no longer used or resulting from clumsy merges should be dropped. Ensure all raw data is retained in at least one dataset to save having to add it back at ESUB time.

CREATING DATASET DOCUMENTATION SUITABLE FOR EXTRACTION INTO THE ESUB.

This is where the programmer has the most impact on the efficient production of a quality ESUB

Having enough detail in a programming plan and in the appropriate language can save having to revisit the matter later. There should be a high level description of the dataset so the purpose and preferably the structure are clear. This can then be lifted into the TOC part of the definition document . Variable content , derivation and formats should be clear, without company jargon and uncommon abbreviations. More is mentioned in the next section.

USER FRIENDLINESS/CONSISTENCY

The ESUB data will be easier to use of there is consistency within study datasets, between studies and between studies and integrated analyses. It may be obvious that variables like patient number and collection date should be in the same format to enable merging of datasets but when multiple programmers are involved then variable names may not remain unique and derivations/format may not remain consistant. Here are some issues to be aware of - Imagine you are the reviewer facing this:

- Same variable name across datasets but different labels or derivations
- Same variable but different format or code values
- Different variables with the same SAS label
- Blank variables – is it by design or an error
- Variable order – subject and visit variables first aid viewing.
- Truncated SAS labels – version 5 transport files allow 40 characters only
- Coded variables without a decode for all terms
- Variable descriptions that refer to variables either not supplied in that dataset or not supplied in any dataset. If the descriptions are copied from source code they may indicate a lack of professionalism

```
data Tom;
set Jerry;
if var2 = 'Yes' then myvar='delete as believe of no value';
```

What are datasets Tom and Jerry about?
What is var2?
Is data being ignored?
If the programs are given to FDA would they understand them?

So supplying derivations in sas code format needs care. A wordy description could be supplied

'Flag to indicate whether the subjects with an abnormal lab value'

May not mean anything to the reviewer without details of the lab variables involved, the lab tests of interest and abnormaliy criteria being applied.

A combination of text and code can prove more user friendly.

PLANNING FOR EFFICIENCY

It is far easier to prepare ESUB components for a study while it is fresh in the minds of the team rather than years down the line when team members may have forgotten the detail or moved on to new pastures. Programmers should be involved in discussions on ESUB deliverables as they understand the data. SAS macros may have to be submitted so should be born in mind when creating them.

Discuss the ESUB deliverables with the project team and then get this agreed with FDA as early as possible. It can vary depending on the therapeutic area, study, and FDA reviewers. Early agreement will help keep ESUB activities off the critical path.

FUTURE

Watch out for CDISC developments and search the FDA web site for changes to ESUB requirements. Currently they have developed very detailed specifications for case report tabulation (raw) and analysis datasets. Variable names, SAS labels, code files and derivations are provided. The Adam team (Analysis Data Model) aim is to design analysis datasets that are one statistical procedure away from producing the statistical results.

New guidance on submitting data may be brought in such as the recent one for HIV resistance data.

CONCLUSION

It should now be apparent that the ESUB datasets and documentation are as important to the success of a filing to FDA as the study tables. There is guidance documentation available which programmers should be familiar with and look out for future changes to the specifications.

The programmers work clearly greatly impacts the content and acceptability of the package. Working up front on appropriate dataset design, documentation, system design and processes is to the advantage of the sponsor's programming team as well as the reviewer. It will reduce the chance of issues cropping up at ESUB production or FDA review - issues that will inevitably come back to the programming team.

The main message to take home is:

The reporting system and process should be designed and operated with ESUB in mind. Imagine you are the reviewer without all your experience of the algorithms and data structures and limited time to complete the review. Would the programming plan and datasets be adequate?

Agree deliverables and start planning as early as you can.

ESUB apart, much of this is about good programming practice that aids code development and maintenance!

CONTACT INFORMATION

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

Simon Dundas
Pfizer Ltd (IPC 191)
Ramsgate Road
Sandwich CT13 9NJ
Work Phone: 01304 646267
Email:Simon.Dundas@Pfizer.com