Clinical Data Standards – An eSubmissions Perspective

Shy Kumar, Datafarm, Inc., Marlboro, MA
Gajanan Bhat, Boston Scientific Corporation, Natick, MA

ABSTRACT

Sometimes we might lose sight of the fact that the data itself is our most important product in the clinical sections of an eNDA or eBLA. A marketing licensing application approval process is highly dependent on the data submitted with the application - its content, integrity, structure and documentation.

The structure of the data (submission data) is important to those reviewers who will actually use the data. The structure of the database is usually different for data collection (the operational database) and data analysis and reporting (the submission database). The operational database records information directly from the study CRFs (case report forms.) The submission database will contain derived variables and is structured for easy use with statistical software. Standards for the documentation of the data are found in FDA guidelines. The Clinical Data Interchange Standards Consortium (CDISC), an industry standards group, recommends additional standards.

In this paper we will discuss the issues, advantages and disadvantages associated with the submission data structure and standards. The process of preparing submission data will be demonstrated.

INTRODUCTION

The process of discovering and developing therapeutic biotechnology products is a lengthy, and labor-intensive one, with extensive regulatory and review guidelines. As shown in Figure 1, it takes an average of 15 years and more than $500 million to bring a new product from the research stage to market.

Figure 1

During this process, the agency review and approval is highly dependent on the data within the submission, making it extremely important. The more confidence the reviewers have in the data (which reflects what actually happened to the patients) 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. This helps the reviewer to better understand the specific details of studies having complex analytical algorithms.

The most important end-user or customer of the clinical data is the agency reviewer. It is every sponsor's goal to satisfy this end-user by providing necessary information while complying with the agency requirements.

CLINICAL DATA

Together, Clinical and Statistical programming group (along with Data Management) is responsible for preparing the clinical data submitted to the agency. Generally, this group contributes, over 70%, to the content of a submission, which ends up at the agency as part of the new drug application

When we started working on this paper we realized that the message we would like to convey could be start-to-end process. But, as you know the clinical data goes through various steps and processes before it becomes submission data. Especially, depending on the company size and resources these steps and processes can vary significantly.

The change is inevitable. Whether we like it or not, we have to comply with ever-changing regulations and agency guidelines.

We hear a lot about new initiative "Implementing CDISC Standards". That is good, but is your standard purely based on the CDISC recommendations? (some?) Or, are they still in draft stage? If yes, please think twice before you implement them.

So, we would like to make an attempt to walk you backwards as the final goal of all entities are the same but the processes and operational requirements could still be different.

The following are the steps that we think are very important in the process. We are most interested in:
1. Deliverables
   o Expectations
   o Requirements/Guidance
2. Process
   o Analysis
   o Verification
   o Data Collection
3. Study
   o CRF
   o Design

Further, the concept of Operational Data (OD) and Submission Data (SD) (thanks to CDISC) is very important since, as mentioned earlier, the operational requirements and processes require different standards.

DELIVERABLES

EXPECTATION

With the publication of the FDA Guidance for Industry in 1999, the expectations of agency for how we deliver the clinical data for review and approval have been made clear.

The data organization varies from indication to indication. Prior to the submission, you should discuss with the review division the datasets that should be provided, the data elements that should be included in each dataset, and the organization of the data within the file. The type of individual datasets that are generally provided to support each study and integrated summaries are included in the following list.
- Demographics
- Inclusion criteria
- Exclusion criteria

- Concomitant medication
- Medical history
- Drug exposure
- Disposition
- Efficacy results
- Human pharmacology & bioavailability/bioequivalence data
- Microbiology data
- Adverse Events
- Lab – chemistry
- Lab – hematology
- Lab – urinalysis
- ECG
- Vital signs
- Physical examination

REQUIREMENTS/GUIDELINES

The advent of the electronic submission and the corresponding data standards have had a significant impact on everything we do related to clinical data. We now have new deliverables: Case Report Tabulation datasets (SAS xpt files), Annotated Case Report Forms (blankcrf.pdf), CRT dataset documentation (define.pdf) and related Table of Contents (crttoc.pdf & datatoc.pdf) files. We also have different deliverable media. The CRT datasets are delivered as SAS transport files and the dataset documentation is delivered in bookmarked and hyperlinked PDF files. Figure 2 shows a typical directory and folder structure submitted for the CRT section.
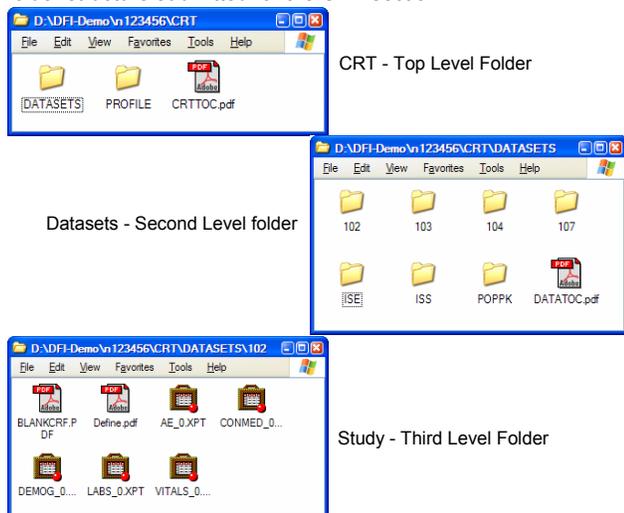

CRT - Top Level Folder

Datasets - Second Level folder

Study - Third Level Folder

Figure 2

In this discussion the requirements focus on deliverables only. The way we process and manage clinical data is not part of these requirements. The agency requirements… "CRTs are item 11 on page 2 of FDA form 356h. You should provide CRTs in datasets allowing the reviewers to use their own software for analysis. Each dataset is a single file and, in general, includes a combination of raw and derived data. Just as you provide each CRF domain (e.g., demographics, vital signs, adverse events) as a table in a paper submission, in an electronic submission, each CRF domain should be provided as a single dataset. In addition, datasets suitable for reproducing and confirming analyses may also be needed."

And they (FDA) add… "Prior to the submission, you should discuss with the review division the datasets to be provided and the data elements that should be included in each dataset. In addition to electronic datasets, study data collected for individual patients, organized by time, can be provided in PDF files. We call this collection of data a patient profile, and it serves as an adjunct to the electronic datasets. Patient profiles are not meant to be a replacement for electronic datasets."

PROCESS

In order to expedite the drug approval process without compromising the approval standards, and also with the advantage of the current information technology, FDA requires sponsors to provide the regulatory NDA submission in electronic format. This provides many advantages such as global standards in submission formats, increased efficiency in sponsors' R&D efforts, faster and more accurate reviews and approval. However, this requires many deviations from the current submission practices, and calls for new standards in formats, data models, and metadata.

In general, in an electronic submission, the process for delivering agency requirements is still the same. However, some of the elements of the final deliverables have been changed. Let's look back and identify the areas that may require change.

PAPER SUBMISSION

Even though the information technology played a major role in the areas of clinical trial, the core part of the process still remains unchanged. In fact, whether we are working in a paper-based or computer-based system, the steps involved in the clinical trial process still remains the same.

Some of the steps that are more relevant to this discussion are:
- Finalize Protocol
- Finalize the CRF
- Design Database
- Data Entry
- Data Verification
- Data Analysis
- Report/Tabulations generation

Since paper printout was the final goal, then the focus was mainly on what went on paper. Some of the items that we say need to be standardized in this process really didn't matter and no one bothered about it. For example:
- Variable Name – The names used were inconsistent within a study and across studies. This was not a problem because it never appeared on the paper.
- Variable Label – The focus was on the label that appeared on the final output. Whether this label was inserted permanently in the dataset or inserted during the execution of the program didn't really matter.
- Dataset Name – Again there was no need for any standard naming convention. For example: for the Demographics dataset, some called it DEMO and others called it DEMOG.
- Datasets Label – Again, the focus was on the label that appeared on the final output. Not many people bothered to store the dataset label within the SAS dataset.

The standards focused on the limitations of SAS software such as length of the variable name and label, length of the actual variable, etc.

This system worked as long as paper was the final output.

ELECTRONIC SUBMISSION

Over the last five years the information technology, especially the Internet, has played a major role in the Life Science industry. It has helped the industry to achieve their goals in speeding-up the drug development process and shortening the time to market products. Of course, the information, which took days and weeks to reach the destination during the paper transaction, was readily accessible in the electronic format. In addition, the electronic clinical trials, electronic data capture (EDC) and finally the electronic submission came into existence. Regardless, some of the steps used in paper submissions still apply. They include:
- Finalize Protocol
- Finalize the CRF (eCRF)
- Design Database

- Data Entry (EDC)
- Data Verification
- Data Analysis
- Report/Tabulations generation – Electronic

So, really there is no significant change in the steps we follow, but the process became more efficient and information is easily accessible.

WHY CHANGE?

The system we use is regulated and changes are inevitable in such an environment. Sometimes change is initiated due to internal business needs and sometimes they are enforced by regulations.

CLINICAL DATA STANDARDS

The standards in this discussion again focus on the deliverables. Looking back to the agency requirements and guidance, it is obvious that one should ensure that while implementing standards, these items are addressed.

The agency has clearly outlined the requirements, in form of guidelines, to sponsors, what they need and how it should be delivered. It is your (sponsors) responsibility to create and provide what agency wants.

Since the changes are inevitable, the standards you put in place should be flexible (i.e. they should not be focused on one aspect of the process). If we look into deliverables, we need to provide to the agency the following for a submission:

- **SAS Transport (XPT) files**
  In SAS, SAS XPORT (transport) files are created by PROC XCOPY in Version 5 of SAS software and by the XPORT engine in Version 6 and higher of SAS Software.
- **Data Definition File (DEFINE.PDF)**
  The data definition file mainly contains metadata describing the data structure, data model, definitions of contents, sources, etc.
- **Annotated Case Report Form (BLANKCRF.PDF)**
  The annotated CRF is a blank CRF form that includes the treatment assignment and maps each blank on the CRF to the corresponding element in the database. The annotated CRF should provide the variable names and coding. Each page and each blank of the CRF should be represented. You should write "not entered in database" in all sections where this applies. The annotated CRF should be provided as a PDF file for each study.
- **Table of Contents (CRTTOC.PDF & DATATOC.PDF)**
  Typically, they are created by the regulatory group - one for the section and the other for the datasets.
- **SAS Programs & Documentations (REQUIRED BY SOME DIVISIONS AT THE AGENCY)**
  Some divisions require all the SAS programs used along with their documentation.

Currently, most of this work is being performed, as a post process, just before the submission, thus requiring an enormous time and personnel. A single minor modification starts the chain reaction and the entire process is affected. The whole concept of electronic submission that is supposed to speed the process is actually consuming more time.

WHY STANDARDS?

"Standard" - Something, such as a practice or a product, that is widely recognized or employed, especially because of its excellence.

The standards are the means for us to be efficient and effective in our day-to-day job. Ultimately, the goal is to develop and deploy the drugs speedily and efficiently.

POINTS TO CONSIDER

If you review the requirements, it is very clear that, by following some standards we can eliminate a lot of post publishing process and reduce the submission preparation time. Also, remember that the regulatory publishing group is not familiar with data and you are the best person to handle this task.

Let's look at the requirements in detail and see what we need to create as deliverables, and how we can create them.

SAS Transport Files

Immediate Need: The requirement here is to make sure that the datasets are converted to v5 Transport files using PROC COPY procedure.

Future: XML format may (not sure) replace the SAS Transport file requirement. Please ensure your new standards and systems do have the flexibility to add modules when necessary.

Define.pdf

Immediate Need: Metadata information at dataset level and variable level. Some of it comes from dataset itself and some need to be stored outside the dataset. Majority of the information is available, collected and stored somewhere by someone during the process.

Future: The metadata will continue to be a requirement regardless of the format of delivery.

Blankcrf.pdf

Immediate Need: Required as described in guidelines.

Future: Authors assume that this will continue to be a requirement regardless of the data delivery format.

Table of Contents

Immediate Need: Typically handled by regulatory publishing group.

Future: With eCTD (electronic Common Technical Document) the TOCs will be replaced by XML backbone. Once eCTD is implemented and accepted the TOCs will no longer be a requirement.

SAS Programs & Documentations

Immediate Need: Optional and required by some divisions within FDA.

Future: May be required in the future.

BEFORE YOU IMPLEMENT

Before you move forward with your standardization process, consider both the short- and the long-term needs. Since, it is impossible to come up with a system that satisfies all the requirements, we think that the ODM and SDM based standards (described below) are really the way to go. CDISC has already done lot of work in this area and there is a lot of input from industry, vendors, as well as from the FDA.

CLINICAL DATA INTERCHANGE STANDARDS CONSORTIUM (CDISC)

CDISC has been playing a major role in this area and has already developed several standards. The two most important standards, more relevant to this discussion are:

1. OPERATIONAL DATA MODEL (ODM)

Since the requirements for data management vary across systems and across companies, one should have the flexibility to handle the data the way they want. Historically there has been no established industry standard for the context, representation and interchange of the data collected during the course of clinical trials. The standards in these areas contribute significantly to the cost and time associated with drug development by eliminating barriers to important information-handling processes and activities.

2. SUBMISSION DATA MODEL

The CDISC Submission Data Model has focused on the use of effective metadata as the most practical way of establishing meaningful standards applicable to electronic data submitted for FDA review. Metadata is defined as "data about the data"; in other words, metadata includes description of the content, context, structure, and/or purpose of a database. It is important to recognize that the metadata provided by the model is intended to be the minimum required to meet the need of FDA users, and is not intended to fully meet all of the needs of the sponsor's data management, statistics, or other internal groups

The primary goal of the SDM is to provide regulatory reviewers with a clear understanding of the datasets provided in a submission by communicating clear descriptions of the structure, purpose, attributes, and contents of each dataset and dataset variable.

MUST HAVES IN YOUR STANDARDS

Regardless of the standard and the model you decide to implement, there are several items you should consider, as must haves, in order to meet current and future requirements for an electronic submission. Again, we will look one-by-one into these requirements.

SAS Transport Files:

Make sure your system uses PROC COPY procedure and creates v5 transport file. This is going to be a requirement for a while. Since a majority of you is migrating to SAS v8, you have to verify and ensure that the length of variable name, label, and many other enhancements in v8 comply with v5 requirements. This is an important issue and should be handled appropriately.

Once XML or other format of delivery comes into existence your system should have the capability to accommodate it. We believe in modular approach that can be implemented without interrupting the entire system.

Data Definitions File (DEFINE.PDF):

By now you know very well what goes into this file. Here is the list and points to consider while building your metadata repository.

Dataset	Current	Recommended
Name	Stored within the dataset/Not Standardized	Consider standard naming convention
Label/Description	Manually entered at the end	Store within the dataset
Keys	Varies. Some sort, others don't. Also, the v5 transport file does not retain this information.	Make sure to sort the dataset and store the *sortby* info in dataset. Also, it is important to store this information in your metadata repository.
Purpose	Entered at the end	Store in your repository
Comments	Manually entered at the end	Store in your repository

Regardless of the number of operational datasets you use, it is recommended that you follow the submission data model (SDM) naming conventions for all the datasets. Your system should have the capability to generate submission ready datasets using the operational datasets. Also, your submission data or CRT

data should be the same data your statistical group uses for all the analyses and reporting.

Building a metadata repository will help a great deal, as this information will be reused across projects and can be a good source for standardization.

Variable	Current	Recommended
Name	Stored within the dataset/Not Standardized	Consider standard naming convention
Label	Sometimes not stored in the dataset.	Store within the dataset
Type	Stored in the dataset	Consider SDM recommendations
Length	Stored in the dataset	Consider SDM recommendations
Formats	Varies. Some companies are holding back from using formats and creating decoded variable right next to coded variable.	If you do use formats please limit it to one or two. Use of formats has its own advantages. Your system should be flexible enough to accommodate both.
Origin	Not stored in Dataset. Manually entered at the end	Store in your repository
Role	Not stored in Dataset. Manually entered at the end	Store in your repository
Comments	Not stored in Dataset. Manually entered at the end	Store in your repository

Again use SDM recommendations while building your repository. The information not stored in the dataset should be stored in the repository during the dataset creation.

Annotated CRF (BLANKCRF.PDF):

It will be nice if we could create one annotated CRF for both internal and submissions purpose. That way we always have one document to deal with. One could create two annotated CRFs one for operational data and another for submission data. This is OK if the metadata is stored and managed appropriately. (see Figure 3)
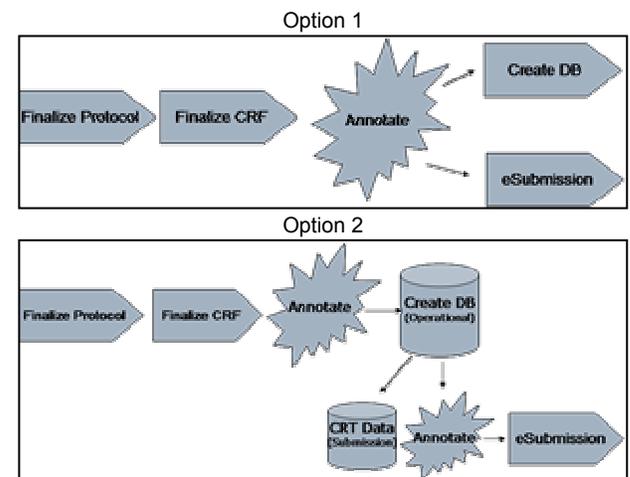

Figure 3

*Table of Contents (*CRTTOC.PDF **& **DATATOC.PDF*):*
It is desirable to deliver the CRT section in its entirety to the regulatory group. Again, this is not an important element in this discussion and it is included for the purpose of completeness. Some publishing systems have the capability to create this item automatically. eCTD does not require creating any PDF table of contents however, the xml backbone attributes require the metadata that you have stored in your repository.

SAS Programs & Documentations
Submitting this item is not a requirement, however, it is recommended to consider it in your planning and submission effort. This ensures that you will have the capability to deliver it, if it becomes a requirement, in the future.

We recommend creating the list of all programs and macros used in the CRT development including CRT programs and analysis programs. Of course you could store more information for your day to day use. The following are the recommended list of items to be considered for submission:

- **Dataset creation programs**
  - Name
  - Description
  - Purpose
  - Specifications
- **Table/Listing/Figure creation programs**
  - Name
  - Type (Table/Listing/Figure)
  - Number (Table/Listing/Figure number)
  - Description
  - Purpose (In-text/End-text/Appendix)
  - Specifications
- **Macro Programs**
  - Name (macro name)
  - Purpose
  - Location (SAS program in which macro call is stored)

CONCLUSION
Of course, to be in compliance with agency regulations for handling clinical data, besides submission requirements and guidelines, we need to follow additional guidelines and requirements. The focus here is how do we deliver data in the way agency wants without having to spend too much time at the end of the process. The tendency has always been to dump the responsibility on others saying, "This is submission work and belongs to Publishing group".

Ultimately someone has to do this and it is better if the right person handles this job. Implementation of standards and procedures definitely help individuals who perform this task. Also, building a central repository with metadata information not only helps different entities within the group, but also paves the way to better understand and reuse information while complying with standards you put in place.

Consider eCTD requirements in your effort. This is yet another important stepping-stone in global regulatory submission for new drug application.

Regardless of the size and complexity of the system the information we need for data submission can be easily retrieved and deployed to the agency when necessary if the information is appropriately created and stored. We wish you best of luck in your efforts. Please remember "**Patient is waiting**", your goal is to develop drugs **not** software.

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

Shy Kumar
Datafarm, Inc.
221 Boston Post Road, Suite 480
Marlboro, MA 01752
Work Phone: 508-624-6454
Fax: 508-624-0848
Email: shy@datafarminc.com
Web: www.datafarminc.com

Gajanan Bhat
Boston Scientific Corporation
1 Boston Scientific Place
Natick, MA 01760
Phone: 508.650.8000 x5254
E-mail: Gajanan.Bhat@bsci.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

REFERENCES
1. Guidance for Industry - Providing Regulatory Submissions in Electronic Format — NDAs (January 1999)
2. Regulatory Submissions in Electronic Format — General Considerations (January 1999).
3. Operational Data Model (ODM) Version 1.1 Final (April 2002)
4. Submissions Data Model (SDM) Version 2.0 (November 2001)