

Reviewing External Vendor Data Transfer Specifications from Programming Standpoint

Mahesh Minnakanti, Vita Data Sciences, Waltham, USA

ABSTRACT

Data transfer specifications (DTS), drafted by the data management team, carries information on how the non-CRF related raw data should be transferred for integration with clinical study database. The specifications are finalized after due feedback from the biostatistics, vendors, and the programming team. With the advent of decentralized trials and increase in device data, DTS hold a potential to not only define a dataset but also a potential to delay the go live of a clinical trial if the trial transfers are not carefully tested in time. In this paper, we will discuss the general contents of DTS and what aspects of DTS are to be reviewed from programming standpoint with few examples. This paper will also discuss about certain approaches on pre-programming activities to map data not only as per CDISC SDTM standards but also from analysis standpoint to make data mapping easier on receipt of live raw data.

INTRODUCTION

Significant amounts of data are collected that supports clinical study design end points originating from a variety of sources. Clinical data collection is often comprised of two databases which are electronic data capture (EDC) system and external vendor database. Electronic Data Capture (EDC) systems is commonly made up of electronic case report forms (eCRFs) by Clinical Data Management (CDM) team. External vendor database is maintained by vendors who support study-specific data needs to perform specialized testing and maintenance of separate databases often termed as external vendor databases to maintain data such as laboratory, electrocardiogram [ECG], electronic diary [ediary], magnetic resonance imaging [MRI], etc.

To make EDC and external vendor data integration efficient and streamlined, most vendors will draft and circulate a data transfer specifications/agreement (simply know as DTS/DTA). In this paper we will discuss on the considerations and expectations of the DTA from the statistical programming standpoint which will help in the process of mapping the external vendor data to the submission ready datasets which are referred as SDTMs

BACKGROUND

Data collection is often defined by the protocol design and study end points. A Clinical Data Management (CDM) team is responsible for the build of the clinical database, used to capture clinical evaluations, adverse events, concomitant medications, etc. Additional vendors also referred as external vendors may also be involved based on the study data collection needs.

Clinical data management (CDM) team make sures the primary focus of database design in relation to vendor data is to capture enough data to sufficiently reconcile the clinical database against the vendor data and avoid capturing redundant information in the clinical database. At minimum EDC systems are built to have eCRF forms that confirms as to whether or not the assessment/test/sample collection (along with the subcategories as needed e.g., for lab data) was performed along with the collection date and time as needed.

Each vendor has their own database processing systems. They have their own limitations on how they can process and present the data to the study team. Over the years vendor database processing systems have evolved to process and present the data close to submission ready. But challenges still exist while mapping the vendor data based on the study design and based on the vendor selection too. As an example, safety laboratory testing may utilize the laboratory associated with each site (commonly referred to as "local laboratories"), or a central (core) laboratory. A specific eCRF may be designed to manually enter the laboratory test results to facilitate a data export to a single file; or if each local laboratory has export capability, each of the exported files, potentially each with different file formats, will need to be processed. In either case, harmonizing data from local laboratories, including mapping to common test names, and converting to common units is very time consuming and potentially error prone. In contrast, when a central laboratory is contracted, samples are collected at each of the investigational sites, shipped to the central laboratory, and processed in the same manner. The central laboratory therefore promotes consistent testing and analysis methods, which yields a single data export containing common test names and units.

To have a defined process and make EDC and external vendor data integration efficient and streamlined most vendors will draft and circulate a data transfer specifications/agreement (simply know as DTS/DTA) document for review. Clinical data management (CDM) team will initially review the document and circulate it to the appropriate study teams for further review before the DTA is finalized and agreed between the study team and the vendors.

Development of Data Transfer Specifications:

DTA's are usually encouraged to be developed and finalized before the EDC go live. There are multiple steps and layers of reviews involved in the process of developing and finalizing the DTA before the external vendor database go live. There are multiple goals in the process of developing the DTA. First and foremost is to make integration efficient between EDC and external vendor database. Along with the integration there are some other major considerations such as reporting necessary variables and having clear data blinding plans based on the protocol design and data needs. Also, to have the variable and value level meta-data listed clearly for efficient mapping of external data to the submission ready datasets.

Along with the CDM and vendor data management team it is encouraged that DTA's are reviewed by study Biostatistician and Statistical programming team. Biostatistician review the DTA to make sure all the study needed assessments/tests are being reported and proper blinding processes are followed if the study has any blinding requirement that are part of the vendor data.

The submission database is presented as clean package of datasets with associated define documentation. However, behind the scenes, the creation of that neat and tidy submission package is often complicated by data collection inconsistencies and data reconciliation issues that cause inefficiencies to programmers. Spending time in the initial stages of the study database built and getting inputs from programming team while developing the DTA would help the clinical trial to ensure consistent data values among various databases and constructing checks to reconcile common data points to improve the data quality and more efficient programming.

CONTENTS AND REVIEW OF DATA TRANSFER SPECIFICATION FROM A PROGRAMMING STANDPOINT:

Each vendor will have its own database and processes for receiving and inputting data values. It is important for the programming team to make sure that they review and provide inputs as needed on how the external vendor data is collected and reported to make the SDTM mapping process efficient.

Prior to the creation of any database, the definition of a few key variable values should be considered. Based on the trial design, more variables may be needed for unique identification, but for most trials, subject and visit identifiers are the minimum requirement. We will discuss more in this paper on the contents and why is it important for a programming team to review and provide feedback from programming standpoint before the DTA is finalized.

Import Scope:

Import scope specifies if the scheduled transfers are incremental (newly reported data compared to the past reported data) or Cumulative (new data along with the already reported data from past visits). In most of the cases it will be cumulative other than if the protocol has any special needs. Knowing this will help us to build programs to have data cuts based on the specific study milestone needed.

SCHEDULE OF DELIVERY:

This section specifies the frequency of the data transfers from the vendor. It is important that the transfer schedules are planned based on the study deliverable needs which are defined during the protocol design. For a programmer to know the data transfer frequency will always help to plan the programming timelines in accordance to study timelines.

MEANS OF TRANSFER:

This section specifies the means that will be used for the data transfers between the vendor and the study team. Mostly transfers are done through a secure file transfer portal referred as SFTP. Its always good that programmers have access to the SFTP in order to have access to the latest raw data posted by the vendor.

DATA FILE FORMAT:

Depending on the vendor database capability they can provide data to study team in a specific file format. Examples are comma separated (CSV), excel (XLSX), export file (XPT), SAS dataset (SAS7BDAT). SAS, XPT, CSV should be preferred format respectively to avoid any special character or data import issues.

FILE NAMING CONVENTION:

This section specifies the data file naming convention that vendor delivers to the study team. If there are multiple files from the vendor each file should be named appropriately based on the data. Data file name should help to identify if its test or production data. If the file is in SAS dataset format as a statistical programmer, we need to make sure that the SAS file name shouldn't start with a character and file name length less than 32 characters.

Example for test filename convention: vendorname_labs_YYYYMMDD_HHMMSS_test_n.SAS7BDAT,
vendorname_labs_YYYYMMDD_HHMMSS_n.SAS7BDAT

BLINDING AND UNBLINDING PROCESS:

This section is carefully reviewed by the CDM team and biostatistician to make sure appropriate processes are set in place to maintain the data blinding based on the study design. Plasma concentration values, for instance, may expose which study drug a subject has received. Establishing a good process such as, blinded data files are reviewed by the unblinded team before delivering it to the blinded programming team. Programming team will be first in the line to work closely with the data, so it is good to know what data is blinded in the study and know study milestones that need the unblinded data to plan the programming activities and timelines.

FILE STRUCTURE:

This section lists out the variable and value level meta data that are part of the data transfer. There will be multiple individual file section for all the listed data files in the DTA. Prior to the creation of any database, the definition of a few key variable values should be considered. Based on the trial design, more variables may be needed for unique identification, but for most trials, subject and visit identifiers are the minimum requirement. Ensuring the subject identifier is of the same format in all supporting databases is critical.

Visit and data time point consistency across the databases. In the below example 1, visit mapping is straight forward across the vendors 1 and 2 with minor naming convention changes.

Study A – Vendor 1		Study A – Vendor 2	
Clinical Database	Vendor Database	Clinical Database	Vendor Database
Screening	Visit 1	Screening	Screening
Week 2	Visit 2	Week 2	Week 2
Week 4	Visit 3	Week 4	Week 4

But in the below example 2, study A has multiple cohorts and each cohort have its own visit schedule. Vendor 1 is differentiating between cohort 1 and 2 which makes visit mapping straight forward. But Vendor 2 is not differentiating between the cohorts. Programming team should make suggestions on how the visits formats should be in order to map them to the appropriate cohort visits.

Study A – Vendor 1		Study A – Vendor 2	
Clinical Database	Vendor Database	Clinical Database	Vendor Database
Cohort 1 Screening Day -28	C1 Visit 1	Screening	Screening
Cohort 2 Screening Day -29	C2 Visit 1	Week 2	Week 2
Cohort 1 Week 2	C1 Visit 2	Week 4	Week 4
Cohort 2 Week 4	C2 Visit 2	Week 6	Week 6

DTA may also include anticipated result values and test names. Carefully review these values if provided, and if appropriate, inquire as to whether custom output values are available, such as test name values mapped to Clinical Data Interchange Standards Consortium (CDISC) controlled terminology. Like the example of visit name values described earlier, if data values will be submitted following standard terminology, implement standard terminology in the database as much as possible, otherwise time will be spent during the programming process researching and mapping values. Apart from the common key variable considerations approach of review varies based on the external vendor databases (e.g., laboratory, electronic diary [ediary], electrocardiogram [ECG], magnetic resonance imaging [MRI], pharmacokinetic [PK], Questionnaire etc.).

For example, let's consider the lab DTA where, making sure the mapping related variables such as categories, and specimen details are presented. and analysis related variables such as alert flag for continuous results and clinically significant for categorical results along with low and high range indicators. For **ECG** data where triplicates are collected make sure to review and confirm if the DTA has a variable to identify the triplicates.

Reviewing and understanding the variable types, description, and their formats along with field lengths is important. It is important to review and understand the comments field section in the file structure that contains the examples of data values and all the possible values if no appendix for value level is presented.

Below is the example of a file structure for the falls data:

Col#	Field Name	Item Description	Item Format	Field Length	Comments
1	STUDYID	Study Identifier	CHAR	12	Default: STUDYID
2	SITE	Site number	Numeric	5	The site number is specified by a 5-digit #
3	SUBJID	Subject Identifier	Numeric	3	Subject identifier is specified by a 3-digit #
4	RECTYPE	Type of record	CHAR	5	Code list: • Compliance • Fall
5	DATE	Date of recorded fall and daily compliance of medication and fall	DD-MMM-YYYY	11	e.g., 05-May-2010
6	INPUT	Source of input	CHAR	6	Code List: Sensor : (from sensor device) Diary : (reported by patient)
7	Time	Time	Hh:mm:ss (24-hour clock)	8	e.g. 13:25:02

If RECTYPE==Compliance, time is set to 00:01 (dummy time)
If RECTYPE==Fall and INPUT==Diary, this denotes the time when the user reported the fall during fall diary
If RECTYPE==Fall and INPUT==Sensor, this denotes the time when the fall was detected by the sensor.

FEW CONSIDERATIONS FOR VENDOR DATA ON PRE-PROGRAMMING ACTIVITIES BASED ON DTA:

To make vendor mapping easy many companies have come up with different processes such as maintaining the vendor and domain-based CT mapping catalog that will make mapping process simpler. Every time a new vendor is added the mapping catalog for the given data is updated to include those vendor and data specific tests.

Once the DTA is finalized we can have format catalogs build for both test and visit level mapping. This will also help to check the production transfers against the DTA. Any inconsistencies between DTA and data are reported as data issues and no work around programming should be done to map this data.

Having proper checks to avoid any potential unblinding if the vendor data has any blinded data. Also, if the eCRF indicates data should be present in the vendor database, check that corresponding data exist in the vendor database. If data exist in the vendor database, check the eCRF confirms data should be present in the vendor database.

CONCLUSIONS:

Programming inputs and review of DTA at the development stage brings more efficiency and programming awareness. It is good to have finalized DTA before the database go live along with the CRF. It gives enough time for the programming team to have the submission datasets ready if there is any planned immediate safety data review. In addition, development of proactive reconciliation checks results in cleaner data, and ultimately more efficient statistical programming. Selection of vendor plays a major role. As now a days some of the vendors maintains robust database systems that are close to the submission standards which makes programming easy.

REFERENCES

<https://www.lexjansen.com/pharmasug/2017/DA/PharmaSUG-2017-DA05.pdf#:~:text=Clinical%20trial%20data%20collection%20is%20often%20comprised%20of,perform%20specialized%20testing%20and%20maintenance%20of%20separate%20databases.>

CONTACT INFORMATION

Author Name: Mahesh Minnakanti

Company: Vita Data Sciences

Address: 281 Winter St,

City / Postcode: Waltham, MA 02451

Email: minnakantimahesh@gmail.com