PharmaSUG 2017 - Paper DA05

Clinical and Vendor Database Harmony; Can't we all just get along?

Brian Armstrong, QST Consultations, Ltd.
Renée Kerwin, QST Consultations, Ltd.

ABSTRACT

Clinical trial data collection is often comprised of multiple databases, in a hub and spoke network with the clinical database as the hub and external vendor databases as spokes (e.g. laboratory, electronic diary, electrocardiogram, magnetic resonance imaging, pharmacokinetic, etc.). 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 at the front end of a clinical trial to ensure consistent data values among various databases and constructing checks to reconcile common data points will promote quality data and more efficient programming.

INTRODUCTION

Clinical trial data collection is often comprised of multiple databases, in a hub and spoke network with the clinical database as the hub and external vendor databases as spokes (e.g. laboratory, electronic diary [ediary], electrocardiogram [ECG], magnetic resonance imaging [MRI], pharmacokinetic [PK], etc.). The clinical database is commonly created using an electronic data capture (EDC) system made up of electronic case report forms (eCRFs). Depending on the trial indication and the data required to support safety and efficacy, vendors may be contracted to perform specialized testing and maintenance of separate databases.

While creating submission data sets from multiple sources, programmers are frequently forced to map data values and often encounter data inconsistencies that require research and resolution. Efficient integration of separate databases benefits from common design elements and data reconciliation. Considering the hub and spoke design of these databases, information from each spoke needs to link to the hub. However, since the clinical database and vendor databases are created and maintained by independent companies, a data reconciliation process is recommended to cross-check common data values.

STUDY PREPARATION

During protocol development and trial design, study endpoints are defined, which in turn defines data collection required for the study. A contracted data management group is commonly responsible for the build of the clinical database, used to capture clinical evaluations, adverse events, concomitant medications, etc. Additional vendors may also be contracted as needed to support study-specific data needs.

With the addition of vendors comes the challenge of determining how data external to the clinical database will be collected, stored, transferred and integrated. 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. The use of local laboratories generally increases data management cost. 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.

Prior to the creation of ANY database, the definition of a few key variable values should be considered. Given the trial design, more variables may be needed for unique identification, but for most trials, subject and visit identifiers are the minimum requirement.

SUBJECT IDENTIFIER FORMAT

Ensuring the subject identifier is of the same format in all supporting databases is critical. Indeed, defining a consistent subject identifier may seem obvious, however, when a clinical trial is designed and contracts with several vendors are being solidified, work is divided and conquered and such detail may be overlooked. It is advised to determine a subject identifier format that is unique and compatible with each database. However, some studies are designed with more than one subject identifier, such as a screening identifier and an on-study subject identifier. If separate screening and on-study subject numbers are necessary, ensure the numbering scheme of each is unique and that each identifier is captured in the clinical study database for reconciliation purposes.

Below are suggestions for subject numbers:

- As a general rule for all sequential numbering schemes, avoid the use of leading 0's, as some system exports may not preserve leading 0's (i.e., number "001" and "0001" will be exported as "1").
- Consider the use of a subject identifier that incorporates the investigational site number and subject number. For example, if investigational sites are numbered 101-1xx and subjects are numbered 101-1xx within each investigational site, subjects could be numbered 101-101, 102-101, and 101-111, or 101101, 102101, and 101111.

In addition to the subject number, the subject's initials and subject's date of birth are commonly captured. However, some countries prohibit the capture of these two data fields due to protection of privacy as they are considered personal identifying information. At trial design, determine whether subject initials and/or date of birth are needed and whether it is reasonable to obtain for all subjects. If the information will ultimately not be available for every subject, do not include initials/date of birth as identifiers in order to design each database to report consistent information for all subjects.

TIME POINT CONSISTENCY

Promoting time point nomenclature consistency between the clinical database and each applicable vendor data set will result in efficiencies during data reconciliation as well as programming final data sets.

Table 1 includes visit names for two example studies, Study A and Study B.

Study A – Visit Names		**Study B – Visit Names**	
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
Week 6	Visit 4	Week 6	Week 6
Week 8	Visit 5	Week 8	Week 8
Week 10	Visit 6	Week 10	Week 10
Week 12/ET	ET	Week 12/ET	Week 12/ET
Unscheduled	UNS	Unscheduled	Unscheduled

Table 1. Visit Names Example

The visit conventions for Study A differ between the clinical and vendor databases. In order to merge the data sets, one of the data set visit names must be mapped to the match the other. In Study B the visit names are the same between the clinical and vendor database, thus requiring zero extra effort to merge data sets. Mapping visits in Study A is not difficult. That being said, this is just one variable from one data set, and it does take time. Scale this type of mapping across multiple databases and multiple variables, and you have the recipe for unnecessary time and money spent.

A common discussion relating to visit naming conventions revolves around whether retests are considered "repeats" or "unscheduled" and whether to include an "early termination" visit.

In the study visit flow chart pictured above, green represents visits with scheduled laboratory testing and blue represents visits without scheduled laboratory testing. A subject attends the Week 4 visit and has blood samples drawn for laboratory testing. After receipt of the sample, the central laboratory indicates that the sample was not sufficient for testing. The subject returns for the Week 6 visit and another blood sample is drawn. Question – Is this "Repeat" or "Unscheduled"? Answer – Both! The laboratory sample was taken as a repeat to the Week 4 insufficient sample and was collected at a visit that was not scheduled for laboratory sampling. Unscheduled may represent samples taken at the time of an Unscheduled Visit (e.g. the subject returned for a visit between the Week 4 and 6 visits); on the other hand, "Unscheduled" may also represent testing performed at a visit different than the study plan. In the end, naming these samples "Retests" or "Unscheduled" is personal preference, but the naming conventions among all databases should be the same.

CLINICAL DATABASE

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 set. Since the goal is to use the vendor data in the final data sets, capturing redundant information in the clinical database is not beneficial, as it generates more work for data entry staff, data management staff, and the statistical programming staff to reconcile. At a minimum, confirmation as to whether or not the assessment/test/sample collection was performed should be entered in the clinical database.

When using a central laboratory for safety laboratory testing, gather information from the vendor or contract research organization regarding the investigational site procedure of collecting samples and how the samples are used for testing. For example, if serum chemistry, hematology and urinalysis testing are to be performed, understand whether one tube of blood will be drawn for serum chemistry, one tube will be drawn for hematology and one container will be collected for urinalysis. If that is the case, design the eCRF to capture each applicable sample date/time. Avoid designing ambiguous clinical database questions that are compound and do not provide enough detail to adequately reconcile the data. See below for an example of incomplete clinical database information and one possible ramification to the corresponding laboratory vendor database:

Subject Number: **101001**
Visit: **Visit 4**
Date of Visit: **01Jan2017**

Were laboratory samples collected?: **Yes**

SUBJID	VISIT	LBCOLDTC	LBCAT	LBTEST
101001	VISIT 4	2017-01-01T08:38:00	CHEMISTRY	ALT
101001	VISIT 4	2017-01-01T08:38:00	CHEMISTRY	AST
101001	VISIT 4	2017-01-01T08:38:00	HEMATOLOGY	RBC
101001	VISIT 4	2017-01-01T08:38:00	HEMATOLOGY	WBC
101001	UNSCHEDULED	2017-01-02T09:23:00	URINALYSIS	PROTEIN
101001	UNSCHEDULED	2017-01-02T09:23:00	URINALYSIS	PH

The clinical database collects the date of visit and response to the general question of "Were laboratory samples collected?". During reconciliation, data management notices missing urinalysis results for 01Jan2017. Due to the lack of precise data collection for each sample, a query must be issued to inquire whether a urine sample was collected on 01Jan2017. The response to the query is "Subject unable to provide urine sample on 01Jan2017. Subject returned on 02Jan2017 and urine sample was collected." As demonstrated by the example above, poor database design may increase the number of queries posted by data management requiring response by the investigational site. Additionally, due to the increased volume in queries, some of which are deemed unnecessary, queries that represent true data issues may receive inadequate attention.

In contrast to the example above, by collecting confirmatory data for each sample, the following clinical database design captures the date of collection of each expected sample as well as the reason a sample was not collected, thus eliminating the need for further queries and reconciliation.

Subject Number: **101001**
Visit: **Visit 4**
Date of Visit: **01Jan2017**

Were laboratory samples collected?: **Yes**

Chemistry Sample Date:	**01Jan2017**
Reason Not Done:	
Hematology Sample Date:	**01Jan2017**
Reason Not Done:	
Urinalysis Sample Date:	**Not Done**
Reason Not Done:	Subject could not provide urine sample.

Just as collecting too few data points is problematic, capturing too much information in the clinical database can also create reconciliation issues. The example below depicts a study utilizing ECG testing from a central laboratory where the clinical database is capturing the same results as the vendor data set.

Subject Number: **101001**
Visit: **Visit 2**
Date of Visit: **14Nov2016**

Was ECG performed?: **Yes**

ECG Date: **14Nov2016**
ECG Time (24 hour clock): **13:08**

Overall Interpretation: **Normal**

Heart Rate (beats/min):	**75**
RR Interval (msec):	**806**
PR Interval (msec):	**156**
QRS Interval (msec):	**91**
QTcB Interval (msec):	**431**

SUBJID	VISIT	EGDTC	EGTEST	EGORRES
101001	VISIT 2	2016-11-14T13:08:25	Interpretation	NORMAL
101001	VISIT 2	2016-11-14T13:08:25	ECG Mean Heart Rate	75
101001	VISIT 2	2016-11-14T13:08:25	RR Interval, Aggregate	806
101001	VISIT 2	2016-11-14T13:08:25	PR Interval, Aggregate	165
101001	VISIT 2	2016-11-14T13:08:25	QRS Duration, Aggregate	91
101001	VISIT 2	2016-11-14T13:08:25	QTCB Interval, Aggregate	431

As can be imagined, manually-entered data are subject to human error. The PR Interval, as highlighted above, was received in the vendor data transfer as 165, however, the result was incorrectly entered into the clinical database as 156. A query must now be issued to confirm which value is correct, which causes data management, the site, and contract research organization to spend additional time during the reconciliation process, thus driving up the overall cost of study management.

Although avoidance of duplicate data capture is preferred, certain studies may require such entry. For instance, a study with strict or complicated inclusion/exclusion criteria may benefit from entry of certain tests results as confirmatory checks of subject eligibility prior to randomization.

In addition to capturing data for vendor reconciliation purposes, the clinical database is often designed to capture information to augment or further categorize results from vendors, such as collection of the investigator's interpretation (e.g. clinical significance determination for abnormal test results) or adjudication information from an independent committee (e.g. Clinical Endpoint Committee adjudication). When additional information is captured within the clinical database, it will most likely be required to associate the information to a record or set of records within the external vendor data. Therefore, enough information must be captured within the eCRF to reliably link to the external vendor data. For instance, one needs to consider whether multiple records per day are possible, which will require additional entry of time. Moreover, exact test names, as received from the vendor, should be made available within drop-down or radio-button lists and include an "Other, specify" option to accommodate unanticipated test names. As a special note, if capturing additional information for laboratory tests, include both absolute and percentage differential test name options in the eCRF. One or the other (i.e. instead of both or neither) could be flagged as abnormal.

VENDOR DATABASE

Each vendor will have its own database and processes for receiving and inputting data values. As described above, when developing the clinical database, knowledge of the vendor database is useful in ensuring the two databases are designed with common values and to avoid capture of duplicate information. Most vendors will draft and circulate a data transfer specifications document for review. This document will likely describe file format and transfer frequency, but may also include additional detail that requires careful thought and consideration, such as whether resulting data from an external vendor may be potentially unblinding. Plasma concentration values, for instance, may expose which study drug a subject has received.

The data transfer specifications document 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. Similar to 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.

In most cases, the data transfer specifications document will provide the meta-data of the vendor database and as well as describe the timing of data transfers. Such information is essential when constructing checks used in data reconciliation.

DATA RECONCILIATION

Proper design of the clinical database in coordination with the vendor database specifications values will promote the data reconciliation process, but does not necessarily circumvent the process. The following are principal concepts for data reconciliation:

- Check both ways
- Schedule of data reconciliation checks
- Query process

CHECK BOTH WAYS

The most important and fundamental concept to data reconciliation is to perform checks both ways (i.e., cross-check the consistency of values considering each database as containing correct data).

1. Clinical Database → Vendor Database
2. Vendor Database → Clinical Database

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.

Consider an example of a study performing ECGs. In this case, the ECGs are collected and stored in machines furnished to each of the investigational sites by the central ECG vendor. The investigational sites must perform an upload/sync process to transmit the ECG records to the ECG vendor's data repository, and in turn be part of the data transfer from the vendor. Imagine a reconciliation check that is unidirectional, such that when the vendor data export is received, the check attempts to find a "home" for each external vendor record. As long as the data received is linked to an appropriate visit, all is well. Seems reasonable, right? WRONG. Unnoticed will be ECGs that were completed, not uploaded to the central repository, and are indicated as performed in the eCRF. If data reconciliation checks are constructed to check both ways, a flag will be raised that ECGs are indicated as being performed, however, there are no results present in vendor data transfer.

DATA RECONCILIATION SCHEDULE

A schedule for data reconciliation activities should be determined and documented appropriately (i.e., within a data management plan). Understanding the life cycle of data values in both clinical and vendor databases will aid in determining the most appropriate timing to perform data reconciliation. The clinical database is a live database such that data entry and source data verification may take place at anytime. Hence, a data export from the clinical database is a point-in-time copy of data, which may be out-of-date by the time the data export is saved to one's working environment. The same is true of the vendor database. It is a live database with a delivery schedule of data transfers, which may be weekly, monthly, quarterly, etc. Until the end of the study (when all data entry is complete in both databases), discrepancies WILL exist due to timing of data exports from each database.

For the clinical database, the life cycle of a data value will follow a similar process to that below:

Data are collected from the subject during the study visit and investigational site staff enter data in the clinical database. Source data verification is then completed (typically by a contract research organization). If discrepancies are identified, queries are posted and the investigational site staff reviews

and corrects data as appropriate. Source data verification is again completed and the data value is considered clean.

Each vendor database will have its own life cycle for data. For some vendors (e.g. safety laboratory, PK, mycology), samples require shipment, processing, resulting, and posting to the database. Other vendors are able to obtain data electronically via uploads (e.g. ECG, ediary).

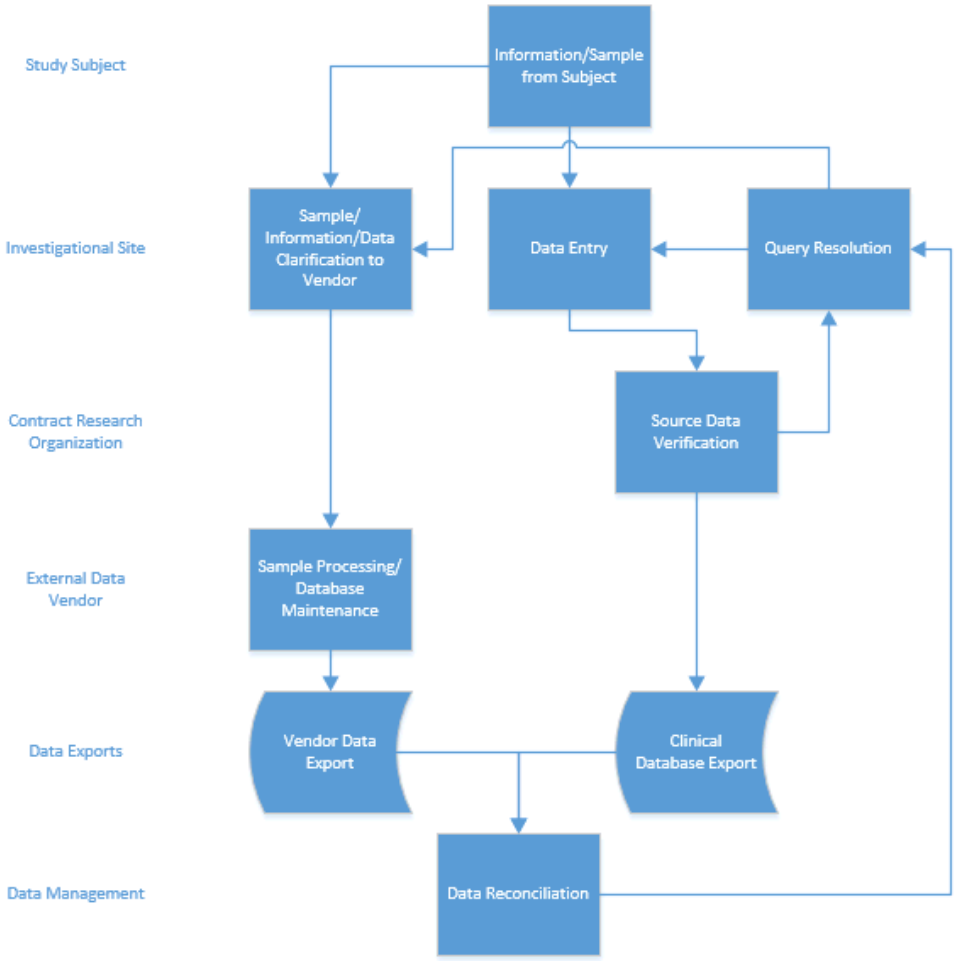
Given the life cycle of data values for the clinical database and vendor database, a reconciliation process that cross-checks data from each without consideration of timing and completeness will be inefficient. Time will be spent reviewing discrepancies due to export timing, rather than true discrepancies. From the clinical database, if possible, restrict the reconciliation process to source data verified or "monitored" records, as these records have been entered and independently reviewed. Have a discussion regarding the expected timing of data entry by investigational site staff and time to results within the vendor database. Consider adding a calendar time element to the reconciliation process, such as to review reconciliation issues for visit and/or vendor dates that had occurred at least on or more than two weeks earlier than the date of the source data exports. For example, if the vendor data transfer is dated 16Dec2016 and the clinical database export is from 17Dec2016, consider reconciliation issues from either source data verified records that are from 02Dec2016 and earlier.

QUERY PROCESS

An established query process is an essential component to clinical/vendor database reconciliation efficiency. Presumably, issues identified during data reconciliation are after data entry and source data verification. Discrepancies that are not caught with data reconciliation will either be included in the submission data sets or cause issues for statistical programming, which may generate inopportune queries during database lock processes and finalization of submission data sets. For these reasons, it is best practice to clean the data prior to statistical programming.

The specifics of the query process and information flow will need to be vetted given the parties involved. Once a discrepancy is identified by data management, who is the best party to have the information available to research and understand appropriate action, the investigational site or the vendor? For example, when safety laboratory testing is performed, a blood sample is collected from the study subject and sent to the central laboratory with a requisition form that includes sample identifying information. Consider a sample drawn in the first part of January, where the date of the sample was mistakenly noted as 2016 instead of 2017. The central laboratory will not have the appropriate information to know this was in fact an error, and will include the incorrect date in the database and in the data transfer. In this case, it seems reasonable that the query be first sent to the investigational site.

As the clinical database is the hub, most (if not all) information to be reconciled should be sourced to the clinical database. Since the investigational site owns all information input into the clinical database, the query process typically begins by sending queries to the investigational site. See the flow chart below for an example workflow tracing the complete life cycle of data captured from both a vendor as well as within the clinical database:



As seen above, the investigational site will review each query and supporting data against source documents. The investigational site will then determine whether a correction is needed in the clinical database, the vendor database, or whether all data are correct as entered. Clean data reconciliation reports are desired at the time of database lock, which indicates data values common to multiple databases match.

CONCLUSION

In conclusion, promoting consistent data capture across source databases will gain efficiency in programming of submission data sets. In addition, development of proactive reconciliation checks result in cleaner data, more efficient database lock timelines, and ultimately more efficient statistical programming time. Key points to remember are to design databases to capture NECESSARY information in CONSISTENT format and perform TIMELY database RECONCILIATION.

CONTACT INFORMATION

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

Brian Armstrong
QST Consultations, Ltd
(616) 892-3706
barmstrong@qstconsultations.com

Renée Kerwin
QST Consultations, Ltd
(616) 892-3713
rkerwin@qstconsultations.com