

REVIEW ARTICLE

External Data Transfers

Data collected from external sources can be essential to the quality of a clinical trial. This chapter reviews some of the types of external data that may be utilized within a clinical trial and discusses the best practices for handling such data. Processing steps for the validation, editing, and verification of external data are examined, and the importance of key variables is emphasized. Discussions are included concerning file and record formats, transmission of data, procedures for database updates, and archiving of external data.

Keywords: Clinical Data Management; Data Transfer Specifications; Test Data Transfer; Non-EDC Data

Introduction

Often during the conduct of a clinical trial, much data external to the case report forms (CRFs) will be collected. If not included in the primary safety or efficacy parameters, these data can be used for subject screening, routine safety and quality-of-life monitoring, or trend analysis. To speed up this process and minimize the use of different analyzing methodologies and equipment, it is common for sponsors to refer to the use of centralized vendors. Such vendors provide electronic transfer of computerized data into the sponsor's database, thereby offering quick results, standardized testing, and reference and calibration values applied to data collected across study sites with the potential to eliminate transcription errors and key entry of data. This chapter focuses on the structure and handling of external data most often required in clinical trials.

Scope

What follows is the data management perspective of the challenges involved in incorporating any external data into a clinical database while assuring that the quality, integrity, confidentiality, and plausibility of the clinical information is maintained. Further, processing steps that affect the data quality are identified, and a solution framework proposed.

Since regulatory guidance exists and data interchange standards have already been proposed, this chapter will reference on a smaller scale (but not attempt to fully cover) the subjects of providing data for regulatory submissions, clinical data interchange standards (FDA,¹ CDISC²), and validation of computer programs (FDA,³ ACDM/PSI⁴).

For information specific to the handling of laboratory data, see the chapter of *Good Clinical Data Management Practices* entitled "Laboratory Data Handling."

Minimum Standards

- Establish the procedures for collecting, transferring, loading, validating, and editing external data through sponsor and vendor collaboration.
- Identify and involve vendors as early in the process as possible.

- Identify key individuals for communication and follow through.
- Provide written specifications for loading external data into the sponsor's database. In advance of loading the data, identify and agree upon mandatory fields or critical variables.
- Maintain a documentation trail.
- Ensure that parties involved have written standard operating procedures and documentation to support that the SOPs have been followed.
- Establish written procedures for safeguarding the blind when primary efficacy data are collected externally.
- Apply quality control procedures to each stage of data handling to ensure that all data are reliable and have been processed correctly.

Best Practices

- Audit external data providers on a regular basis as part of your vendor audit practice (see also the Vendor Selection and Management chapter).
- Enforce a formal data clarification process for handling data discrepancies and data updates.
- Validate all programs and systems used for processing clinical trial data in a clinical research environment (see also the Database Validation, Programming, and Standards chapter).
- Provide vendor-specific training. A clear understanding of what is expected by both sides is critical for quality and efficient conduct of the clinical research.

Types of External Data

External data can originate from different sources, but it is a common practice for a centralized vendor to specialize and produce one or more major data types. Examples of data types include:

- Laboratory and PK/PD Data
- Device Data (ECG, Flowmetry, Vital Signs, Images, and other)
- Electronic Patient Diaries

It is significantly important to identify and describe the variables that must be included in any data transfer, regardless of where the data originate or the information contained within the data transfer. The purpose of these variables is to merge the external data to the sponsor's clinical database; safeguard the blind; and ensure that data belonging to a particular protocol, investigator, and subject cannot be loaded to a subject enrolled into a different protocol or to an incorrect visit. Working with the end goal in mind, one can observe that these data may constitute an integral part of the dataset domains proposed by FDA/CDISC for submission:^{1,2}

Dataset	Description
DEMO	Demographics and subject characteristics
DISPOSIT	Disposition
EXPOSURE	Drug exposure
AE	Adverse events
CONMEDS	Concomitant medications
CHEM	Labs – chemistry
HEMAT	Labs – hematology
URINE	Labs – urinalysis
ECG	Electrocardiogram
VITAL	Vital signs
PE	Physical examination
MEDHIST	Past medical history

Refer to CDISC for additional information.

External Data Processing Steps Affecting the Data Quality

The following areas may adversely affect the integration of external data and should be accounted for during database setup:

- Definition of key variable and mandatory fields
- Data editing and verification procedures
- Record formatting and file formats (e.g. SAS®, ASCII)
- Data transmission
- Database updates
- Data storage and archiving

Key Variables

To ensure that sufficient information is available to identify and process data at the sponsor's site, it is imperative that key variables (those data that uniquely describe each sample record) be carefully selected. Without such variables, it proves difficult (if not impossible) to match patient, sample, and visit with the result records accurately.

While these variables are intended to uniquely identify and clarify subject visit records, incomplete data collection or presentation of errors in either primary or secondary key variables can result in inadequate information. Therefore, completeness in the choice of variables collected and transferred offers a way to increase the accuracy and overall quality of the process. Primary

(protocol subject identifiers) and secondary (additional subject and unique vendor identifiers) key variables can include the following:

Primary Key Variables (Protocol subject identifiers)	Secondary Key Variables (Additional subject and vendor identifiers)
Sponsor Name/ID	Subject's Gender
Study/Protocol ID (any combination of project and protocol)	Subject's Date of Birth
Site/Investigator ID	Subject's Initials
Subject Identifier (Subject Number, Screening Number or number assigned by the CRF used)	Transmission Date/Time
Clinical Event ID (Visit Number)	Date associated with the Subject visit
Sample ID (vendor or device specific sample identifier or a subject visit)	Sequence Number (when more than one observation per record exists)

Data acquisition forms, whether conventional or electronic (i.e., CRF, e-CRF), should be designed to facilitate the full and accurate reporting of key information at the study site.

Parties involved in the process should identify in writing and agree in advance upon key variables or fields for loading external data into the sponsor's database. They should also avoid duplication of information. For example, if subject initials and date of birth are already in the database from the CRF and are not selected as primary keys, these variables should not be transferred on the external file. The key variables and value ranges should be specified in advance so that they can be incorporated in range-checking programs.

When any of the efficacy parameters are collected in the external data, particular attention should be paid to safeguard the blind. For example, bone density indicators in an osteoporosis trial may be collected with a study's lab data and could be blinded to the physicians and clinical personnel at the sponsor's site. In case of full double-blind or full triple-blind trial, these data must only be disclosed to parties not directly involved in the trial or data safety monitoring committee. A written procedure must exist describing how this data will be handled and to whom it can be disclosed before the clinical database lock. In a similar scenario, subjects may be excluded from the efficacy analysis for loss of baseline data if any of the pre-treatment blind results are incidentally revealed to personnel directly involved in handling the subject.

Data Editing and Verification Procedures

For quality and timely processing of data, errors must be eliminated at the source or as close to the source as possible. To facilitate this goal, sponsors and vendors must work together to develop editing and verification procedures. These procedures should include:

- Provisions for treatment of partial data
- Checking for duplicate demographic details and results (real or near real time where possible)
- Range of subject numbers allocated for the study, investigator, or both
- Range of treatment codes allocated per study, investigator, or both

The sponsor and vendor should identify key individuals for communication and follow-through. A representative from clinical data management should be included. It is recommended that the sponsor provide a range of subject and treatment codes for each protocol before external data are received for integration. The allocated ranges should be included in data validation routines and any discrepancies handled as part of a formal discrepancy management mechanism. Very often, a centralized vendor (ECG, laboratory organization) with quick results turnaround time will be able to identify and resolve data discrepancies before any other clinical information is entered into the database or even reviewed.

The vendor should perform duplicate record checks as subject visit data is received. Duplicates should be resolved following a formal data clarification process with the investigative site.

Whenever possible, the sponsor should provide the vendor with a complete listing of subjects' demographic details or IVRS demographic data for an independent reconciliation of the sponsor database and remote database during the study conduct or before database lock.

The vendor and sponsor should agree upon procedures for assuring that the sponsor receives complete data. If partial records are included in a data delivery, they should be indicated as such. The vendor should provide procedural verification and assurance that a hard copy of the results is identical to the electronically transferred results. Any changes to the system or the programs used to create either of the reports must be tested and documented accordingly. If data are transformed during processing, a comparison of the original data and observations to the processed data should always be possible.

If applicable, the vendor should provide a complete list of reference values and their effective dates at the onset of the study. Procedures to minimize the possibility of changes during the course of the study must be implemented.

Definition and details of the process for resolution of discrepancies between external and CRF data should be established as part of the study setup. The process should address the issues of both sponsor and vendor or third-party participant.

Record Formatting and File Formats

Quality and efficient integration of data demands up-front consensus between the sponsor and vendor with respect to record and file format. Areas for initial discussion include the size of data fields, clarification of numeric versus character fields, decimal granularity, use of characters such as ">" and "<," quotation marks, commas, and other

special characters. Special consideration should be paid to handling of null or missing data.

Depending upon the characteristics of the database management systems and expertise at the sponsor and vendor sites, there may be a wide variety of acceptable record, field, and file formats. Thus, both parties must negotiate in writing a mutually acceptable and detailed record format structure.

Areas to consider include the following:

- The sponsor should provide in writing a complete list of reportable variables in the order required. If data is requested in a SAS dataset, the output of the CONTENTS procedure should be provided as part of the specification. For ASCII files, the column positions or delimiter, record heading, and field justification should be specified.
- Character and numeric fields should be differentiated. Field formats should be specified, in advance, as numeric or character. Reporting of results that can be either character or numeric should be minimized.
- Sponsor requirements on date and time reporting should be negotiated and specified in writing; examples include DATE9., YYYYMMDD or TIME5., HH:MM (24 hr).
- Procedures should explicitly describe the handling of greater-than (>) or less-than (<) signs. Absolute values should be used where possible or to separate the numeric and character portion of the observation into two fields.
- If applicable, comments regarding the condition of the sample or its non-availability should be reported in a field that is separate from the results.
- The test data in the agreed upon format should be available in a file to be used during database set-up and validation at the receiving Sponsor or designee. Successful generation, transmittal, receipt, loading, and screening of the test data validate the data transmittal process.

Data management professionals should evaluate and leverage the experience of some of the existing and emerging vendor independent standards for data interchange between clinical systems, including HL7,⁵ ACDM's Standards for Electronic Transfer of Laboratory Data,⁶ and CDISC.²

Data Transmission

Problems encountered with transmission of data from vendor to sponsor will result in data being lost or incorrectly loaded. To facilitate the transmission process in all cases, complete naming conventions and labeling information must be established. Any data transferred between the vendor and sponsor must contain sufficient information to be uniquely linked to the source of the data and corresponding project and protocol. Origin, date created, date sent, number of records, and a version-controlled file naming convention should be followed.

Public encryption mechanisms such as PGP® (*Pretty Good Privacy*®) are recommended for use when transferring

data via the Internet. Thus, the data transfer process will ensure compliance with the regulatory guidelines and provide authenticity and confidentiality protection. Not all countries allow the use of strong encryption software. In such cases, consider the use of password-protected files such as ZIP archives or dial-up FTP transfer. Both processes will verify the integrity of the file being transferred and provide feedback in case of file corruption.

Procedures for Database Updates

The processes by which updates to subjects' records are made are among the most vulnerable for generation of errors. Special consideration should be paid if the edit affects any of the primary key variables, and thus propagates multiple records (see also the Data Processing chapter).

Errors generated by the data-cleaning process in the sponsor's database should be communicated back to the vendor for follow up and resolution through a formal data-clarification process. To update a record when the original records are either incomplete or contain erroneous data, the vendor frequently will send a second transmission. Updates can be sent either as a full or partial transmission depending upon the capabilities of the systems in place. It is essential that the vendor and sponsor establish procedures that define how retransmissions are identified and handled throughout the study.

Strategies to consider include the following:

- During study set up, provide the vendor with a list of in-house data checks, supporting documentation, and sample subject-number allocations.
- Use correction flags. When possible, two separate types of flags should be used to distinguish an initial record from a correction or addition.
- Corrections to key variables should be identified and flagged. Updates to key variables should be sent as full records (i.e., including result variables) and should be flagged at a record level.
- Only current results should be reported.
- Maintain an audit trail. The source systems should be designed to permit data changes in such a way that data changes are documented and that there is no deletion of entered data.⁷

If applicable, vendors should provide the investigator site and sponsor with updated hard-copy information in addition to electronic updates.

File Storage and Archiving

Ultimate responsibility for the quality and integrity of the trial data always resides with the sponsor.⁸ Thus, the sponsor should specify in the contract a definitive time period beyond the initial transmission of information during which the records will be maintained by the vendor for access by the sponsor and regulatory agencies. It is desirable that vendors maintain active copies of data files during the study stages that require unconstrained accessibility.

After these stages, the vendor should maintain an archived version for the remainder of the retention period. When all reports have been finalized and the sponsor's database has been locked, a study should no longer require access to the records except for auditing purposes during the record- retention period.

For additional information, see the Data Storage chapter, the Database Closure chapter, and the FDA's *Guidance for Industry: Computer Systems Used in Clinical Trials*.³

Recommended Standard Operating Procedures

SOPs should be established for, but not limited to, the following:

Sponsor (CRO)	External Data Provider (Vendor)
External Data Loading and Validation	Data Extraction and Validation
Query Generation and Vendor (remote) Database updates	Data Transfer and Discrepancy Handling
Vendor Auditing	Database Updates
Database lock procedures	Database Archiving and Security
Study-specific procedures (including the handling of extra/unscheduled data)	Study-specific procedures (including the handling of extra/unscheduled data)

Competing Interests

The author has no competing interests to declare.

References

1. **US Food and Drug Administration.** *Guidance for Industry: Providing Regulatory Submissions in Electronic Format, General Considerations (IT2) and NDAs (IT3)*. Rockville, MD: US Department of Health and Human Services; 2003.
2. **Christiansen D, Kubick W.** *Submission Metadata Model: Version 2.0*. Austin, TX: Clinical Data Interchange Standards Consortium; 2001. Available at: <http://www.cdisc.org/standards/SubmissionMetadataModelV2.pdf>. Accessed February 20, 2007.
3. **US Food and Drug Administration.** *Guidance for Industry: Computerized Systems Used in Clinical Trials*. Rockville, MD: US Department of Health and Human Services; 1999. Chapter VI, sections C and D.
4. **Association for Clinical Data Management and Statisticians in the Pharmaceutical Industry.** *Computer System Validation in Clinical Research, a Practical Guide*. Macclesfield, UK: Author; 1997.
5. **HL7 Working Group.** *Health Level Seven Standard Version 2.5: An Application Protocol for Electronic Data Exchange in Healthcare Environments*. New York: American National Standards Institute; 2003.
6. **Association for Clinical Data Management.** *Standards for Electronic Transfer of Laboratory Data*. St. Albans, UK: Author; 1997.
7. *Code of Federal Regulations, Title 21, Volume 1, Part 11: Electronic Records; Electronic Signatures*. Washington, DC: US Government Printing Office; 1998.

8. **International Conference on Harmonisation.** *Guideline for Good Clinical Practice, E6, Section 5.2.* Geneva, Switzerland: Author; 1996.

Recommendations for improving quality and reducing the hidden costs. *Drug Information Journal.* 1993; 27: 933–942.

Further Reading

Bloom JC. Drug Information Association (DIA). Clinical Trial Symposium, January 12–13, 1998.

Bornstein S. CDM task list. *Data Basics.* Fall 1999.

Goodman DB. Standard for ECGs in clinical trials. *Drug Information Journal.* 1996; 30: 237–242.

Peairs D, Ginsberg A, Gillibrand D, Malcolm R. New technologies enhance central laboratory data collection, transfer, and analysis. *Applied Clinical Trials.* 1996; 5.

Pharmaceutical Manufacturers Association, Clinical Data Management Group Task Force. Electronic exchange of laboratory data:

Chapter Revision History

Publication Date	Comments
September 2000	Initial publication.
May 2007	Revised for style, grammar, and clarity. Substance of chapter content unchanged.
October 2009	Name of chapter changed from “Lab and Other External Data” to “External Data Transfers” and a reference to the “Laboratory Data Handling” chapter added to the Scope of this chapter.

How to cite this article: External Data Transfers. *Journal of the Society for Clinical Data Management.* 2023; 1(1): 14, pp.1–5. DOI: <https://doi.org/10.47912/jscdm.334>

Submitted: 01 December 2023

Accepted: 01 December 2023

Published: 22 December 2023

Copyright: © 2023 SCDM publishes JSCDM content in an open access manner under a Attribution-Non-Commercial-ShareAlike (CC BY-NC-SA) license. This license lets others remix, adapt, and build upon the work non-commercially, as long as they credit SCDM and the author and license their new creations under the identical terms. See <https://creativecommons.org/licenses/by-nc-sa/4.0/>.



Journal of the Society for Clinical Data Management is a peer-reviewed open access journal published by Society for Clinical Data Management.

OPEN ACCESS