

Data Management Best Practices

Matt J. O’Nuska III
IT Manager, CASAA

Agenda

- Data Management Overview
- Data Management Components
 - Data Acquisition
 - Data Privacy
 - Data Storage
 - Data Entry
 - Data Archival

Data Management

Primary Role

- Ensure data integrity

Secondary Role

- Accelerate timeline from data collection to data analysis and publication
 - Work closely with researchers in every stage of the project lifecycle
 - Educate users
 - Use standard reporting and communication

Data Acquisition

- Design the forms to collect the data specified by the protocol
- Keep questions, prompts and instructions clear and concise
- Use multiple choice, avoid open ended questions if at all possible
- Maintain consistency throughout instruments
- Make the forms available for review at the clinical site prior to approval

Data Privacy

- Privacy protection afforded to research subjects include:
 - Protocol review and approval by an Institutional Review Board (IRB)
 - Right to informed consent
 - Right of the subject to withdraw consent
 - Right to notice of disclosure
 - Confidential collection and submission of data

Data Privacy (continued)

- Best Practices
 - Educate and train all project personnel
 - Minimize identifiers in data collection
 - Protect non-entered data which could impact client confidentiality
 - Ensure privacy during data transfers
 - Design policies and regulations
 - Implement contract contingencies when utilizing external service providers
 - Maintain proper physical and electronic security measures
 - Report any data privacy violations

Data Storage

- Best Practices

- Store all original data collected in secured areas such as rooms or file cabinets with controlled access (e.g., locks)
- Document the procedures for granting access to database servers, establishing system controls and assigning passwords
- Store electronic data in such a way that backups can be made easily and frequently
- Utilize open formats for broad audience compatibility

Data Entry

- Commonly used processes
 - Independent double entry
 - Double entry with blind verification
 - Double entry with interactive verification
 - Single entry with manual review
 - Single entry with no manual review
- Recommended approach
 - Double entry with interactive verification

Data Archival

- Best Practices

- Lock the database to prevent data modification
- Create database design documentation
- Preserve raw data
- Save final dataset in open format

Questions & Answers



Data Management

The primary role of data management is plain and simple: to ensure the highest possible degree of integrity in your research.

A secondary role is to ensure that all methods and practices accelerate the timeline from data collection to data analysis and publication.

Your data management personnel should work in close collaboration with researchers from the study design through the analysis phase to ensure that data are collected correctly from the inception of a research study and to make certain that you are building in the highest degree of data integrity at every stage (i.e., study design, form design, startup, recruitment, follow-up, etc.) of the project lifecycle. Sound data management methods, practices, and principles are necessary only when data integrity is important, which is always.

User Education is fundamental in developing sound data management practices. In any given research, protocols may be complex, and data management systems and procedures may be foreign to study's researchers, coordinators, clinicians, and administrators. Recognizing that data integrity in research begins with people, researchers should participate in various degrees of training, dependent on their role within the research project. Training sessions should cover the protocol specifics, the use of the data entry/management systems, and location of critical documentation. With user education, and training and certification exercises, you can create the foundation that will lead to consistent data flow and efficient operation of your research.

Efficient and enhanced communication among researchers will lead to greater data integrity. In any study lifecycle, many aspects of the research will be refined many times. It is important that critical information regarding protocol enhancements, data instrument refinements, general news, personnel changes, current events and schedules, etc. all be disseminated in a timely manner, ensuring that your researchers invest more time into their research and less time finding the information they need. This is accomplished by carefully assessing each project's requirements and then implementing communications systems that meet those requirements. Reporting and communications systems could utilize a combination of technologies that integrate a variety of automated web-based systems, e-mail notification, and traditional methods which will result in comprehensive, efficient and even redundant channels of communication.

Data Acquisition

There is arguably no more important document than the instrument that is used to acquire the data from the clinical trial with the exception of the protocol, which specifies the conduct of that trial. The quality of the data collected relies first and foremost on the quality of that instrument. No matter how much time and effort go into conducting the trial, if the correct data points were not collected, a meaningful analysis may not be possible. It follows, therefore, that the design, development and quality assurance of such an instrument must be given the utmost attention.

Instrument selection and design should begin as the protocol is being developed to assure that the protocol-specifications regarding data collection are reasonable and achievable. Unfortunately, many instruments are often developed or changed hastily after the protocol has been approved, or worse, after the project has started. When the protocol and forms are designed concurrently, such collaboration will provide the responsible parties with important feedback. Consideration can be given to what data should be collected and how the data will be used to meet the objectives of the study. If a statistical analysis plan exists, it can be used as a guide to what data points are essential. Collection of extraneous data may adversely affect data quality by distracting site personnel efforts from the critical variables. It is especially important to assure that key variables are defined prior to or during instrument development and that they are appropriately captured on the form.

Questions should be made specific and clear enough to assure that complete and comparable data are obtained across the various populations using the instrument. Provide form completion instructions and definitions for items not directly measurable. For example, “Did the subject have hypertension?” should be clarified by the necessary blood pressure range, length of time sustained or necessity of specific intervention for the condition.

At some point most data must be coded prior to analysis or display. Whenever possible, therefore, data should be collected in coded form. Examples of coded formats include multiple choice pick lists and yes/no check boxes. Careful use of coded formats can both provide for multiple responses where needed and track the total number of responses while, at the same time, encourage the individual completing the form to select at least one response. With the possible exception of providing information about safety issues, free text is rarely useful without an extensive coding resource.

Maintain consistency in the order of similar answer choices throughout the instruments. For example, yes/no and the placement of a “none”, “Not Applicable” or “Other” choice within a series of choices should not change. The question or prompt should indicate if multiple choices are mutually exclusive or not. If not, there should be an indication if more than one choice will be accepted or if only one selection should be checked.

Forms should be reviewed at the clinical site by the people responsible for collecting, entering, and analyzing the data. This will help ensure that the above points have all been met.

Data Privacy

Data privacy refers to the standards surrounding the protection of personal data. Personal data can be defined as any information relating to a research subject, which can lead to the identification, either directly or indirectly, of that subject. Some examples of personal identifiable data are patient names, initials, addresses, genetic code, etc.

Here are the best practices you can follow to ensure data privacy.

Educate all personnel who directly or indirectly handle personally identifiable data on company procedures and data privacy concepts. Training sessions should include company policy, regulatory agency policy and applicable local, state, federal and international law. *It is also a good idea to have all personnel who have access to data sign a confidentiality agreement.*

Design data collection instruments with the minimum subject identifiers needed including the design of case report forms, clinical and laboratory databases, data transfer specifications and any other area of data collection that may contain personal information. Collect or use personally identifiable data only when required for specific scientific reasons. Ensure those reasons for use are documented and justified.

The use of these identifiers should be taken into consideration not only in case report form design, but also in scenarios where the processing, transfer, reporting or analysis of data will be completed. In general, a random subject number and gender can be used to resolve any discrepancies that might arise from transcription errors. Although it is the responsibility of the investigator to ensure that subjects have been given a proper informed consent, it may be beneficial to include the question, “Did the subject read, understand and sign the informed consent?” on one of the instruments. This allows data management to process data into the database with confidence that proper consent was acquired. If source documents are to be collected (i.e. lab reports) the sites should be instructed to ensure that all documentation is stripped of personal identifiers and appropriate subject identifiers should be assigned prior to submission to for data entry.

Protect data that is not directly entered into data collection which may impact client confidentiality. *For example, client locator forms*

Implement procedures that occur prior to transfer of data to between sites, departments, subsidiaries and countries that ensure that all considerations about privacy have been considered, addressed and documented.

Promote internal and external accountability through company policy and regulations governing the use of personal information. Implement procedures for using data for an alternate or new purpose other than what was originally intended by the informed consent.

Ensure that all considerations about privacy have been considered, addressed and documented.

Enforce a policy of “NO” access to personal data as a baseline. Evaluate any request for this information, and if it is determined that it is required for specific scientific reasons, and ensure that all considerations about privacy have been considered, addressed and documented.

Make compliance with data privacy regulations a central focus of audits and a contract contingency when using external service providers. (*example, HIPAA*)

Maintain proper physical and electronic security measures. Data should be stored in protective environments relevant to the type of media being stored. Paper case report forms should be stored in an environment with regulated access. Proper precautions should be taken to prevent external access to data such as password and firewall security.

All data, paper or electronic, should be safeguarded with high standards and those policies and procedures should be reviewed on a regular basis to ensure the latest data protection standards are being implemented.

If identified, address any data that is submitted to data management that appears to be collected without consent or authorization being secured.

Data Storage

The secure, efficient and accessible storage of data is central to the success of clinical trials research. Whether data are collected using validated electronic tools or traditional paper forms, data are often transferred many times during a project. These transfers occur between functional groups within an organization as well as regulatory agencies. The potential for data corruption and version control errors during data storage and transfer is significant and must be minimized to assure consistency of results and data quality.

The physical security of original data sources should be maintained carefully. Original paper and electronic documents should be warehoused in secured rooms or file cabinets with controlled access. Database servers can be the primary warehouse of clinical data and should be physically secured with appropriate procedures in place to regulate access.

Direct access to database servers should be restricted to those individuals with the responsibility for monitoring and backing up the system. All other access to database servers should be controlled by logical security and occur across a secure network using appropriate system controls and password access controls. Special considerations must be given to the physical security of computers used in an electronic data collection trial. In cases where data are entered over a live connection to a central database, the physical security of the central server is a primary consideration. If any data are stored locally at the study site (as in the case of a hybrid or “offline” system) before being sent to a central server, the physical security of the system at the source of data entry is more critical. Permission controls and passwords are vital to assure that only authorized personnel have the ability to access study data. Mechanisms should be in place to capture and prevent unauthorized attempts to access a system, and notification should be made to administrator if such attempts take place.

Electronic data should be stored in a central location, such as a network server, and backups should be made daily. The best method, in my opinion, is to use a redundant system. This is where backups are made to two different media sources – this provides an extra layer of protection in case of hardware or media failure. The media should be moved to an off-site location if possible, or at least stored in a secure, fireproof location.

Utilize open formats whenever possible for archival, storage and transport of data (e.g., ASCII, SAS Transport, Portable Document Format (pdf). Adherence to this practice enables access to the data by multiple systems or reviewers, currently and in the future.

Data Entry

The data entry process should address the data quality needs of the trial. Commonly used data entry processes include the following:

Independent double data entry with a third person compare where two people enter the same data independently and a third person resolves any discrepancies between first and second entry;

Double data entry with blind verification where two people enter the data independently and any discrepancies are resolved during second entry;

Double entry with interactive verification where the second entry operator resolves discrepancies between first and second entry and is aware of the first entered values;

Single entry with a manual review

Single entry with no manual review.

I recommend

Double data entry with interactive review.

Correcting data errors is often the most expensive aspect of data entry, sometimes even exceeding the cost of the original data entry. Unfortunately, many of these costs are intangibles or are not easily measured. When errors are detected early in the data entry process the repair cost is minimal – only the time to re-key the correct character or field. After the data has been processed, it is much more expensive to correct errors. This is called effect leakage – the further down a line a problem goes, the more it will cost to fix it. Anomalous values discovered during a commissioned data analysis can bring the work of the data analyst to a standstill. There may be an hour or more of clerical time involved in finding the document and entering the corrected transaction. Failure to detect errors can be even more costly, resulting in erroneous publications and study reports.

Several archival procedures should be followed to assure that the data are preserved in their raw format. Most importantly, the database itself should be locked upon completion of a study. This means that permissions to further modify the data are removed from all except the most critical study personnel.

Database design specifications - Documentation of the table definitions used to build the study database and file structure.

Raw data - The final raw data files preserved within the study database format and all original data transfers in their raw format.

Final data - It is critical to preserve the final data in a standard file format so that it can be easily accessed, reviewed or migrated to another system.

Original study documents – The original and/or scanned images of all original documents. These may be archived separately in a central records facility if necessary.