# An Overview of Data Management in Human Subjects Research

Andrew G Miller, David Lipscomb, and Chi Hornik

Introduction
Data Collection Tools
Data Security and Data Storage
File Management
Government Regulations and Research Misconduct
Data Management Plan
Preparing Data for Analysis
Data Analysis
Reference Management
Manuscript Files
Summary

**Research studies generate data in various forms. Data can be quantitative or qualitative. Research involving human subjects requires protection of data to ensure privacy. Various regulations and local policies need to be followed to ensure data security. Data management plans are critical for effective data stewardship and include details plan on data collection, management, storage, and formatting. This paper will review data collection tools, data security strategies, file management, data storage, government regulations, prepping data for analysis, reference management, and file management.** *Key words: research; data management; data management plan; data analysis.* [Respir Care 2024;69(2):256–262. © 2024 Daedalus Enterprises]

## Introduction

All research studies include data that need to be collected, securely stored, and properly formatted for statistical analysis.[1] Data can be quantitative or qualitative. Quantitative data include continuous, ordinal, or categorical variables. Qualitative data are more difficult to analyze and result from subject interviews, comments on surveys, or observations of behavior. Proper management of data is integral to producing high-quality and reproducible research.[2] This paper will provide an overview of data management, data security, regulations protecting research data, effective data management strategies, data management plans, and software that can be used for research.

## Data Collection Tools

Data management encompasses building the data collection tool or case report form (CRF), secure storage of data, quality assurance, quality control, and formatting for statistical analysis.[1,2] Data collection tools include paper forms to be filled out at the bedside; electronic forms filled out using a computer, tablet, or smartphone; software that automatically collects data from the electronic medical record; or qualitative forms to collect free-text data.[3] Data can be sourced directly by researchers through medical records, recorded at the bedside by bedside staff, through surveys, or downloaded from devices. In most cases, the data should not solely exist within the data collection tool but also be available in another location, such as an electronic medical record for lab values, blood gas results, or ventilator settings. Whereas electronic data collection tools are used primarily, some studies may utilize paper forms, usually as part of a CRF.

Paper forms have traditionally been utilized and have certain advantages such as rapid data collection, the ability to bring them to the bedside, limited design expertise required, low cost, and simplicity.[4] Paper forms can be useful when electronic data capture (EDC) is difficult or infeasible. The downsides of paper forms are significant and

Table 1. Paper Versus Electronic Data Collection Tools

| | Paper | Electronic |
|---|---|---|
| Data security | Requires secure space for storage | Improved data security |
| | Increased likelihood of loss | Risk of loss of data if server is corrupted |
| | | Vulnerable to hacking or lost devices |
| Illegible handwriting | Common | Eliminated |
| Data formatting | Formatting issues are common | Properly built EDC tools can reduce or eliminate errors |
| | Staff requires additional training | Fewer resources |
| | More resources to confirm data elements | |
| Resource utilization | Can be faster | Requires internet access and server maintenance |
| | Less training for staff | Less time spent cleaning improperly formatted data |
| | | Easier to identify missing data |
| Statistical analysis | Requires transfer to electronic format | Can be exported into statistical program format |
| | | Reduces formatting errors |
| Automated calculations | Not possible | Can be built into tool |
| Mandatory data fields | Can be omitted | Cannot be omitted |
| Video, photography, or audio files | Not possible | Can be uploaded and stored online |

EDC = electronic data capture

include issues with illegible handwriting, incomplete data, difficulty changing once approved, lost forms, data security, and storage. The use of paper forms also results in redundant data collection as all data will need to be entered into an EDC tool, thus potentially wasting resources while increasing the risk of errors during data entry. Several studies have noted that switching to electronic CRFs (eCRFs) from paper forms reduced errors and decreased the amount of time required to complete each CRF.[5,6] Whereas institutional policies vary, many centers require all data from studies enrolling children research projects need to be kept for 3–5 y after the study is completed or until the youngest subject is 21 years of age, whichever is longer, and paper forms will need a secure storage place. There are medical storage facilities that institutions must pay for secure storage of these data. This can be costly for sites that have a

Mr Miller is affiliated with Division of Pediatric Critical Care Medicine, Duke University Medical Center, Durham, North Carolina; and Respiratory Care Services, Duke University Medical Center, Durham, North Carolina. Mr Lipscomb is affiliated with The Emmes Company, Rockville, Maryland. Dr Hornik is affiliated with Department of Pediatrics, Duke University Medical Center, Durham, North Carolina; and Duke Clinical Research Institute, Durham, North Carolina.

Correspondence: Andrew Miller MSc RRT RRT-ACCS RRT-NPS FAARC, 2301 Erwin Road, Durham, NC 27710. E-mail: Andrew.g.miller@duke.edu.

high number of enrollments or a large number of paper forms for each enrollment. Lastly, paper forms require entry into electronic databases, which introduces the potential for transcription errors and increases the likelihood of missing data. To minimize missing or inaccurate data values, transcription into electronic database should occur prior to subject discharge. A site source log that records when and by whom data are entered can also help reduce time spent gathering data by reducing redundant activities and allowing for easier auditing.

Electronic databases have emerged as the primary tool for data collection and management. eCRFs eliminate or reduce many of the errors associated with paper forms. A comparison of paper versus electronic forms is included in Table 1. Data security is dependent upon using protected databases, such as Research Electronic Data Capture (REDCap), which is 21 Code of Federal Regulations (CFR) Part 11 ready.[7] The details of 21 CFR Part 11 can be accessed at https://www.ecfr.gov/current/title-21/chapter-I/subchapter-A/part-11 (Accessed July 17, 2023). Even when using a secure database, multiple levels of security and appropriate procedures must be implemented to prevent breaches or other data integrity problems. Individual users will have limited access based on their role, with only the principal investigator (PI) and select designees having full access to the data. Each project needs appropriate procedures, documentation, and to ensure all team members have the appropriate qualifications and training. Access to secure servers used to store data needs to be limited to only those listed as key personnel on the protocol approved by the institutional review board (IRB).

Advantages of eCRFs include ease of data entry, standardization of data elements within the database, the ability

to use smartphones or other portable electronic devices, and proper formatting of variables.[8] If built correctly, proper formatting of variables can reduce or eliminate free texting–related mistakes. Examples of variables, in which proper formatting can reduce mistakes, are date/times, number formatting, and multiple-choice options in lieu of free-text sections. eCRFs can be shared with different centers, which can allow for common data elements across studies, which will make it easier to compare results and ensure similar variables are collected. There is also potential for studies to be included in preplanned or patient-level meta-analysis. Lastly, eCRFs are more environmentally friendly as they have a minimal carbon footprint, require less physical space, allow for real-time quality control, and data trends can be analyzed to allow for corrections or protocol updates.[8]

Disadvantages include the need for expertise to build the eCRF, time required to test and validate the tool, the need for readily available computers with internet access, and data security depending on the program used.[8] eCRFs are often used for surveys and multi-center studies. In some studies, eCRFs can be auto populated from the electronic medical record to reduce the workload and costs associated with data collection.[9,10] Availability of automated data collection varies from center to center as these systems require extensive testing, validation, auditing, and proper security measures before they can be used. Data can also be collected from many devices used on patients, including ventilators, monitors, and CPAP machines. In most cases, these data are used for clinical practice and then extracted for research. A summary of paper versus electronic data collection is included in Table 1.

## Data Security and Data Storage

Protected health information (PHI), defined as any potentially identifying data collected as part of a research study, requires protection. Maintaining confidentiality is critical when accessing PHI or other sensitive data.[11] Study team members should limit their access to what they need to perform the study. All data collected should be securely stored. Paper forms, including consent forms, are required to be kept in a secure location. Only the PI or designee should have access to the forms, and they should be kept in a locked cabinet.

Electronic data should be stored within a secure EDC system to comply with 21 CFR Part 11, Health Insurance Portability and Accountability Act (HIPAA), Federal Information Security Modernization Act (FISMA), and General Data Protection Regulation (GDPR). HIPAA was passed in 1996 and is intended to protect patient's health care–related data. Health care providers, including researchers, should only access the minimal data required for their job function, cannot disclose any PHI, and must diligently protect patient data.[11] FISMA was passed in

2002, updated in 2014, and applies to studies funded by federal or state governments. It provides standards and guidelines to management of information risk (https://csrc.nist.gov/projects/risk-management/fisma-background. *Accessed July 17, 2023*). GDPR is the European Union's data privacy law that prevents sharing of personal data without consent. All health care data are covered by the GDPR but are also required to be kept confidential.[12] GDPRs are quite complex, vary from country to country, and the details are beyond the scope of this article.

If possible, data should be de-identified as soon as possible. De-identification means removing any PHI that could identify the subject, including dates of interventions. After de-identification, a key or log is usually kept by researchers to link the records in the event additional data need to be collected or an audit is performed. The log or key needs to be secured and password protected. Access to the database should be limited depending on individual study roles. Usually, those responsible for data entry have limited access compared to the database manager, statisticians, and PI. Data should never be stored on personal devices, such as personal computers, unencrypted thumb drives, or external hard drives. With proper encryption and approval from the IRB, encrypted hard drives or secure cloud-based servers can be utilized.

An example of a widely used, secure electronic EDC system is REDCap.[7] It is a web-based program that can be used for any type of data collection. It complies with 21 CFR Part 11, HIPAA, FISMA, and GDPR. Access is limited to those with an account and generally limited to those who work at individual institutions, although some centers may grant access to external personnel in certain circumstances. For multi-center studies, the primary site or data coordinating center will control the EDC system, and other sites will have a secure link to enter data but will not usually have rights to make changes to the forms or export data. All team members with access to the database must be listed as key personnel on the IRB. Each subject is entered separately, and there is no risk of inadvertently altering other subjects' data during data entry. It also allows the database manager(s) to query for missing data elements and allows timely acquisition of the missing data. It also includes a log so the database manager(s) can see who entered specific data elements, which allows for quality audits and follow-up of missing data. Different levels of access can be granted, and most team members will not be able to export data. Once data on individual subjects are complete, records can be locked so no further changes can be made. REDCap also has a feature that allows for de-identification of data sets prior to export and can automatically format the exported file for several commonly used statistical programs. Lastly, since REDCap is 21 CFR Part 11 ready, it may be used for electronic consent, which eliminates the need for storage of paper forms.

Another common tool for EDC is Microsoft Excel (Microsoft, Redmond, Washington). Whereas many medical devices can export data into large Excel files, it should be used cautiously for clinical research. REDCap allows for uploading of data from Excel. Challenges with Excel are formatting, data entry, missing data, and accidental deletion of entire rows or columns. Those with substantial expertise in using Excel can build databases to minimize these risks; however, other programs generally provide better security with a shorter learning curve. When many records are in a single Excel database, transposition of a record could occur, resulting in all the entered data being one column or row off. This could take hours to fix or in some cases result in the inability to complete the project. Improperly formatted data entry can also result in substantial time to clean and reformat when data are analyzed.

Additional programs used for data collection include internet cloud-based programs like Microsoft Forms, Microsoft Access, or Google Forms (Google, Mountain View, California). Regardless of the software chosen, it needs to meet the regulations for protecting research data and approved by the IRB. There are many other commercially available programs in addition to REDCap. Regardless of the EDC tool chosen, it should be easy to use for those entering data, export data in a useable format, and meet all the appropriate regulatory requirements.

## File Management

File management is critical for a successful project. A project can be derailed if files are not managed correctly. For example, files should be named so that another researcher can find and identify individual documents. Developing a master word document with file names can also be used to help keep track of files, including what is included in each file. A data dictionary should also be made to define individual variables, measurements, and instruments used. If the protocol changes or measurements change, this can be updated throughout the project.

For projects including PHI or sensitive data, a folder on a secure server should be used for all data containing PHI or identifying information. Shared folders with study documents, including references, proposals, grant applications, IRB documents, manuscript drafts, and statistical outputs, can also be used. It is generally a good practice to keep all study documents in one place. Multiple subfolders may be necessary. Depending on your institution, you may need to request access to certain secure servers, and access will usually be limited to those listed as key personnel on the IRB protocol. It is a good strategy to save each revision of documents in a format that can be understood by another researcher. The final document can be saved as final or as a .pdf so it can no longer be edited. Examples of file naming methods are provided in Table 2.

## Government Regulations and Research Misconduct

Government regulations regarding research data are complex and can be difficult to navigate.[11,12] Each institution should have standards in place to ensure compliance with all government and local regulations. A complete description of federal and state regulation is beyond the scope of this article. Research misconduct is a significant departure from accepted practices and includes fabrication, falsification, or plagiarism (https://ori.hhs.gov/content/chapter-2-research-misconduct-federal-policies. *Accessed July 17, 2023*). Good data management is critical in case a researcher is accused of fraud or data fabrication. It could derail or destroy someone's career if they are unable to reproduce their data, even if there was no malfeasance on the part of the researcher. Some journals may ask for data to be shared so an independent statistician can evaluate the results, especially when the study has major implications or reviewers have concerns about the veracity of the data. Privacy breaches result in significant civil liability, with over $130 million in total fines being levied and even criminal cases pursued in rare circumstances (https://www.hhs.gov/hipaa/for-professionals/compliance-enforcement/data/enforcement-highlights/index.html. *Accessed July 17, 2023*). Whereas not all of these were related to research data, there have been substantial fines levied for lost research data or unsecured data. If a breach or data loss is experienced, it should be reported immediately to your institution's IRB and other appropriate departments.

## Data Management Plan

It is essential to have a clear data management plan before starting the study. Ideally, a data scientist, statistician, or someone with experience in data management should review the plan. Every study requires a data management plan, approved by the IRB.[2,13] This includes how the data will be collected, where data will be stored, who has access to the data, a quality assurance plan, when and if the data will be de-identified, and when or if the data will be destroyed. All data management plans need to be approved by the IRB and meet all internal institutional standards. Some centers may have a secure, centralized server or data storage facility in which data can be stored. When possible, researchers should take advantage of these programs as they have appropriate safeguards in place, including server maintenance and other preservation services. The sponsor or funding agency may have additional requirements that must be met, including the sharing of de-identified data sets generated from public funds or who owns the data after the study is complete.

Multi-center studies also require data use agreements so the data can be shared within a single database or combined for analysis. These databases usually include de-identified

Table 2.    Examples of File Names

| | Examples |
|---|---|
| File name - Vx | PICU Asthma Protocol Research Proposal - V1 |
| Vx - file name - date | V1 - PICU Asthma Protocol - Research Proposal - 6-14-23 |
| Date - file name - Vx | 6-14-23 - PICU Asthma Protocol - Research Proposal V1 |
| Year/mo/date - file name Vx | 2023/6/30 - PICU Asthma Data Dictionary V1 |
| Specific file types, use the same naming system as above | |
|     Data dictionary | Store in a word document, update as needed |
|     Data sets, spreadsheets | Name so they can be accessed and found in a search |
| | Label raw versus cleansed |
|     Digital media, pictures, videos | Secured on a server with adequate space |
| | Consider backups in a different location |
|     Other media, analog recordings | Kept in a safe, dry, and secure location |
| | Attempt to transfer to a digital media if possible |
|     Statistical outputs | Export within program |
| | Export raw results into a word document in case the original statistical program is no longer in use |
|     Manuscript drafts | Save within a distinct folder, name each version, and save feedback from other authors |
| | Label revisions clearly so they can be easily found |
|     Tables and figures | Save separate files as most journals require tables and figures to be uploaded separately |
| | Prior to submission, keeping tables and figures in a single file for circulation among team members can be helpful |
|     Lab notebooks | Clearly label the outside of a paper lab notebook |
| | Scan into a digital format and store on a secure server |
| | Electronic versions should follow the same rules as any other file |

V = version.

data, and any issues with accuracy or missing data need to be resolved by the disclosing site. Usually, one center serves as the data coordinating center, and all data are submitted to a single site through a secure online system. Clinical trials and rigorous observational studies include robust auditing procedures, including remote and on-site monitoring visits to verify source documentation of data, before they are finalized.

Even with a sound data management plan, unexpected challenges will occur in any study; and changes may need to be made to the CRF, data collection tool, or measurements made. In most studies, it is a good idea to thoroughly test and then review the first few CRFs to identify and fix any problems early. Fixing an error in the CRF may not be feasible in large studies, especially if it is not discovered until the data are being analyzed or after a large number of subjects have been enrolled. Once data collection is complete, most studies require audits and data cleansing to identify any issues with the data. Lastly, once the manuscript is being drafted, the data need to be stored in a secure location; and multiple copies should be available in case of audit, verification of results, or the data are needed for future explorations. For example, data could be stored within REDCap or another EDC tool, the raw data in a spreadsheet on a secure server, and a third copy on a secure, encrypted hard drive. Tips for creating a data management plan are summarized in Table 3.

## Preparing Data for Analysis

Once data collection is complete, data need to be prepared for analysis. The first step is to identify any missing data and attempt to acquire it. Next is to flag any unusual data points and double check they are correct. Depending on the type of study, this can be followed by a random audit of data for different subjects. It is reasonable to audit 5–10% of subjects to ensure data elements are accurate. If there are certain data points that were inconsistently collected, it may be necessary to audit all the records for those variables. After completing quality audits, collecting missing data, and fixing any errors, the data are ready for the last steps before analysis.

The final step is to clean, format, and transform the data so they can be analyzed by your statistical program. This can involve calculating variables from raw data (such as $P_{aO_2}/F_{IO_2}$, oxygenation index, or time between 2 dates), creating categories from continuous variables, and any formatting needed. Formatting includes ensuring categorical

Table 3. Tips Creating a Data Management Plan

**Before You Start**
  Have a data management plan from the beginning, include in initial proposal
  Involve a data scientist or statistician early
  Get organized; create folders and file naming structures to easily find documents
  Determine the sponsor's (if applicable) and local requirements
  Identify the data to be collected
  Include plan and appropriate storage for any retained biospecimens
  Define how the data will be organized
  Have a plan for quality assurance and auditing
  Develop a storage and preservation plan; consider the 321 data storage rule (3 copies of the data, at least 2 types of media, and one stored at a
    separate site)
  Data should be findable, accessible, interoperable, and reusable (FAIR)
  Define data policies, including clear strategies for PHI
  Describe how the data will be disseminated, planned papers, target journals
  Assign roles and responsibilities, who will manage the database, perform audits, and who is ultimately responsible for the data
  Prepare a budget, ensure adequate resources are available, including time for data management
**After You Start Collecting Data**
  Conduct regular audits for quality
  Provide refreshers with team members to ensure compliance with regulations or identify mistakes during data collection
  Refine data collection tool or CRF based on feedback from team
  Record any changes made
  Keep files organized and secure
  Create a clear system for managing different versions of data sets, manuscript drafts, and other important data
**Once Data Collection Is Finished**
  Perform data audit; fill in missing data or clarify mistakes where possible
  Once audit is complete, lock data set
  Save multiple copies of data set in a secure location
  Consider saving a raw data set and cleansed data set
  Data must be kept on site and archived per institutional policies
  Clean and format data for use by statistician
  Save different versions of analysis
  Each multivariable model run should be saved
  Properly store or dispose of any retained biospecimens
**After Data Analysis Is Complete**
  Create clear file names for manuscript drafts
  Consider making separate files for tables and figure
  Save final analysis within same folder as tables and figures
**Once Primary Manuscript and All Secondary Analysis Are Published**
  Keep all files in accordance with regulations and institutional best practices
  Save copies of raw and cleansed data
  Ensure data and analysis are accessible in case of audit, verification of results, potential for inclusion in an individual patient level meta-analysis,
    or compare future work to current study

PHI = protected health information
CRF = case report form
Adapted from references 2 and 13.

responses are identical; a common source of frustration is when a combination of n/a, not applicable, none, or other designations are used to indicate missing data. Most statistical programs require identical responses, and it can take a significant amount of time to ensure they all are identical. It is common to detect additional errors or missing data in this phase. If possible, data should be de-identified during this stage. For example, dates used to calculate time variables can be deleted once the calculations are finished. Once this process is complete, data are ready for analysis.

## Data Analysis

Each round of data analysis should be recorded and saved in an appropriately named file. For complex studies with multiple complicated statistical models, each file

should be updated with the version and date. Files should be saved within the format used by the statistical program, and the final analyses should also be saved in a word document. Similar to the overall project data dictionary, studies involving complex analysis should also consider a document summarizing the various analysis performed. The results should be transcribed from the statistical output into separate tables for submission. A comprehensive review of data analysis is beyond the scope of this article.

## Reference Management

An important part of any research project is keeping track of the relevant prior research on the topic.[14] Due to random file names of downloaded papers, it can be a major challenge to keep track of portable document formats (PDFs) of individual studies. Similar to other files, a good naming system can help authors keep track of the references for their study. Generally, the name of each PDF should be easy to find in a search and kept within the main study folder or a subfolder if there is a large number of references. If authors prefer to print out PDFs or use printed journals, keeping copies of the papers in a physical folder that can easily be accessed is important. In contrast to subject data, security is not a concern with references, and they can be stored anywhere that is convenient for the authors.

There are several commercially available programs that can help with reference management. Commonly used programs include EndNote, Mendeley, and Zotero. All 3 programs can be used to curate reference lists for individual papers, organize papers by topic, and insert references into documents. PDFs can be stored within each program so the full text can easily be accessed. EndNote and Zotero also have RESPIRATORY CARE–specific formatting that can be downloaded. Using these programs can save substantial amount of time formatting and updating references. Certain programs can also be linked to your institution's library so full text articles can be acquired quickly and easily. Lastly, individual libraries can be accessed remotely and shared with other members of the research team.

## Manuscript Files

Once the data are analyzed, the manuscript can be written. Each draft of the manuscript should be clearly labeled, with the first draft being considered V1. Various labeling methods for revisions can be used, and usually coauthors will label their version with their initials at the end of the file. To help with writing the results section, it is reasonable to make a separate document that includes all the tables and figures. This way the study results can be kept side by side

when writing the manuscript to keep the author from having to navigate a single document. It is also helpful for coauthors to be able to refer to 2 documents at the same time. RESPIRATORY CARE requires tables and figures to be uploaded separately. They can be saved as individual files and labeled appropriately during submission.

## Summary

Data management is integral to a good research project. Having a clear plan before the project starts is important to ensure data security, integrity, and formatted. Selecting the right tools and complying with government and local regulations are critical to protect research data.

## REFERENCES

1. Surkis A, Read K. Research data management. J Med Libr Assoc 2015;103(3):154-156.
2. Kanza S, Knight NJ. Behind every great research project is great data management. BMC Res Notes 2022;15(1):20.
3. Saczynski JS, McManus DD, Goldberg RJ. Commonly used data-collection approaches in clinical research. Am J Med 2013;126(11):946-950.
4. Wilcox AB, Gallagher KD, Boden-Albala B, Bakken SR. Research data collection methods: from paper to tablet computers. Med Care 2012;50 Suppl(Suppl):S68-73.
5. Thriemer K, Ley B, Ame SM, Puri MK, Hashim R, Chang NY, et al. Replacing paper data collection forms with electronic data entry in the field: findings from a study of community-acquired bloodstream infections in Pemba, Zanzibar. BMC Res Notes 2012;5:113.
6. Fleischmann R, Decker AM, Kraft A, Mai K, Schmidt S. Mobile electronic versus paper case report forms in clinical trials: a randomized controlled trial. BMC Med Res Methodol 2017;17(1):153.
7. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42(2):377-381.
8. Bellary S, Krishnankutty B, Latha MS. Basics of case report form designing in clinical research. Perspect Clin Res 2014;5(4):159-166.
9. Cheng AC, Banasiewicz MK, Johnson JD, Sulieman L, Kennedy N, Delacqua F, et al. Evaluating automated electronic case report form data entry from electronic health records. J Clin Transl Sci 2023;7(1):e29.
10. Zong N, Wen A, Stone DJ, Sharma DK, Wang C, Yu Y, et al. Developing an FHIR-based computational pipeline for automatic population of case report forms for colorectal cancer clinical trials using electronic health records. JCO Clin Cancer Inform 2020;4:201-209.
11. Rose RV, Kumar A, Kass JS. Protecting privacy: Health Insurance Portability and Accountability Act of 1996, Twenty-First Century Cures Act, and social media. Neurol Clin 2023;41(3):513-522.
12. Chico V. The impact of the General Data Protection Regulation on health research. Br Med Bull 2018;128(1):109-118.
13. Michener WK. Ten simple rules for creating a good data management plan. PLoS Comput Biol 2015;11(10):e1004525.
14. Goodfellow LT. An overview of how to search and write a medical literature review. Respir Care 2023 [Epub ahead of print].