

Translational Informatics

Managing Research Data

Good data management from the start can streamline your data collection and analysis. It can also help you repurpose your data for additional studies and share your data with collaborators.

A poorly executed data management plan, however, can bog down your efforts and may even impede you from completing your intended analysis. The CCTSI Translational Informatics Core has developed a number of resources to help you use well-established data management practices for collecting and storing research data.

Video Resources on Data Management

- [Managing Your Research Data](#)
- [Why not use Excel for data management?](#)
- [Care and Feeding of Your Grant – Part 1](#)
- [Care and Feeding of Your Grant – Part 2](#)

Last Things First: Let Your Analytic Plan Shape Data Management

Make sure that you have formally mapped every piece of data you collect to your analytic plan, safety monitoring or reporting requirements, and make sure that the data collection will supply everything you need. By "formally mapped", we mean each piece of data has an identified role or use. This will help you avoid two related issues:

- **Collecting Too Much Data:** Although it is tempting to collect as much data as possible (who knows what you might find!), more is rarely better (beware of what you find!). Rather than producing intriguing new discoveries, you will find that extraneous data requires extra work to extract and manage, and ultimately produces spurious, undesirable associations. Consider also whether adding data leads to tests for additional associations that require you to adjust for multiple comparisons – pushing the p-value for your core hypothesis above 0.05.
- **Missing Critical Data:** It is heartbreaking to invest your efforts as well as those of your staff and your subjects in collecting a well-validated seven-item assessment instrument at three time points, only to find that item #4 was inadvertently left off the third time point, and the response scale on item #6 changed from the first time point to the second. Yet another reason to avoid collecting too much data is that trying to manage too many data elements makes it easy to overlook tiny critical errors in the data you care most about. As you add more data, the overall quality of all data, including your core variables, goes down.

It is a good idea to map your data and analytic plan both "forward" (mapping concepts/aims to data) and "backward" (mapping data to concepts/use).

- **"Forward" data mapping example:**

Construct	Measured by	Source/Instrument	Frequency
Primary Outcome	Hemoglobin A1c	Health Trac; VDW	PC Visit
Secondary Outcomes	SBP, LDL	Health Trac; EMR	PC Visit
Predictors	Dx of Cancer	Tumor Registry	Initial
	Dx of Depression	Health Trac; EMR	Initial
Moderators	Morbidity	Quan index; ICD-9	Yearly
	Continuity of Care	PCP vists	Yearly
	Rx Adherence	Pharmacy	Yearly
Covariate	BMI	EMR	PC Visit

- **"Backward" data mapping example:**

Questionnaire	Measurement	Role
Demographic	Age, Gender	Descriptive, Covariate
	Race, Diagnosis	Descriptive
Baseline	Social Support Scale	Moderator
Assessment	SPSI-R	Primary Outcome, Mediator
	BDI, POMS	Secondary Outcomes

It is critical to consult with your statistical consultants early—well before you implement your data collection plan. Statistical consultants can provide “a fresh set of eyes,” identifying data definitions or collection problems that may have been overlooked by the investigator. More importantly, statistical consultants can propose alternative approaches that can vastly improve the power and quality of your analysis. Statistical consultation is available through the [Colorado Biostatistics Consortium \(CBC\)](#).

Keep Well-Documented Data Dictionaries (Codebooks)

Make sure you create a formal data dictionary for each of your datasets, and update them immediately when any changes are made. At the very least, the dictionary should describe the following **for each field in the dataset**:

- **Field Name (variable name):** Although some database management systems constrain the length of field names, don't shorten field names unnecessarily. ICLR may be a cute name for the "eye color" field, but you and your colleagues will be much more likely to remember what EYECOLOR represents when you return to the dataset months or years later
- **Description:** A label or descriptor that clearly identifies the contents of the field
- **Type:** Will the data be numeric, character, a date, or something else? If numeric, is it an integer or a real number? If a date, what date format (MMDDYYYY, Julian, etc.)?
- **Key "metadata" (data about data)** are particularly critical. When possible, define:
 - **The coding scheme:** Such as the metric used (e.g. micrograms) or the values represented (e.g. 1=sluggish reflex, 2=normal reflex, 3=brisk reflex).
 - **A standard representation of the measurement:** Such as the relevant LOINC or SNOMED code (see "Use Standard Instruments and Representations" below).
 - **How the measurement was obtained:** How the measurement was obtained, such as the assay used, or the algorithm for a derived value such as creatinine clearance.
 - **Upper and lower bounds:** This will facilitate range checks that can be used both at data entry and data validation to ensure data integrity (e.g. making sure you don't have 8-year old mothers, or 45-year old infants in your data set).
 - **How missing data are represented:** If you choose a special value (as is commonly done in social science research, e.g. 97=subject declined to answer, 98=subject did not know) use it consistently throughout (not 97 for one field and 997 for another). **Never use zero to indicate missing data.** Zero is especially likely to be misinterpreted as an actual value rather than as missing data.

Example of a data dictionary:

Data Dictionary for COMIRB protocol #_____.			
NB: Missing numeric data are coded -999			
Values below detection limits are coded -666			
Missing categorical data are coded 6=UNK 7=Refused			
Variable name	units	Format (length)	Description & Additional info (assay/machine/algorithm)
ID		Integer (5)	Unique participant identifier
age	years	integer (2)	Age at date of consent
sex		1=male 2=female	
race		1=White 2=Black 3=Asian 4=Pacific Islander 5=Mixed Race 6=UNK 7=Refused	
ethnicity		1=Hispanic 0=Non-Hispanic 6=UNK 7=Refused	
HTX		0=no 1=yes 6 =UNK	Hypertension indicator
BMD_hip	g/cm ²	f5.2	Hologic, total hip
BMD_troch	g/cm ²	f5.2	Hologic, trochanter subregion
BMD_LS	g/cm ²	f5.2	Hologic, L2-L4
E2	pmol/L	f5.2	E ₂ (estradiol), Diagnostic Systems Laboratories (DSL)
T	nmol/L	f5.2	Total Testosterone, Beckman Coulter
DHEA	μmol/L	f5.2	DHEA, Diagnostics Systems Lab (DSL)
DHEAS	μg/dL	f5.2	DHEAS, Diagnostic Products Corp (DPC)
ITT		1=yes 0=no	Intent to treat indicator (Participant was randomized and given study drug)
Compliant		1=yes 0=no	Compliance (>80% of pills used)

Your data dictionary will be a crucial resource not only as you conduct your current research project, but also if you later wish to share your data with collaborators or other researchers.

Use Standard Instruments and Representations

Don't develop measures from scratch if you can use existing ones. Using standard measures will allow your findings to be compared meaningfully with those of others, and will allow you to reuse your datasets later. Of course, it is vital that these instruments not be modified, or they will no longer be "validated" or comparable. For demographics and health status, consider using instruments from national agencies such as CDC (e.g. [BRFSS](#)) and AHRQ (e.g. [NHANES](#)). If you need help selecting appropriate behavior measures to include in your study, in the near future, investigators will be able to review and select common behavioral and social science measures from the [Grid-Enabled Measures \(GEM\) database](#) in caBIG (the cancer bioinformatics grid). If you must develop new metrics based on questionnaires or scales, consider consulting with a psychometrician to ensure these metrics are reasonably well validated.

Investigators are also increasingly recognizing the benefits of incorporating standard representations of data such as laboratory values (like hemoglobin A1c or glucose), diseases (such as sarcoidosis), and symptoms and findings (such as shortness of breath). Incorporating the following standards in your datasets will greatly improve their reusability in the future, making it easier for you to collaborate with others and allowing you to contribute your data to local and national repositories:

- For laboratory values, [LOINC](#)® (Logical Observation Identifiers Names and Codes) is preferred (e.g. hemoglobin A1c = 17855-8)
- For diseases, symptoms and findings, [SNOMED-CT](#) is preferred (e.g. sarcoidosis = 24369008, shortness of breath = 267036007)

Excel Is Never the Right Answer: Choosing a Database Structure

While there are a number of other database management systems to choose from, the use of spreadsheets such as Excel for data entry and storage is never a good idea! Protect yourself by keeping original primary data in a robust database, which can be exported to a spreadsheet or statistical package for analysis without corrupting the underlying data. **Among the many problems with Excel for data entry and storage are:**

- It is much too easy to corrupt data in Excel. If you make the common error of sorting on a single column and forgetting to undo the change before saving the dataset, your dataset is now hopelessly corrupted and unrecoverable.
- Excel doesn't provide facilities for storage of metadata
- Range checking/data validation is possible but cumbersome
- Keeping all the data on a single spreadsheet encourages PHI to be mixed with non-PHI, which can create privacy and security concerns

While MS Access does solve some of the data entry and storage problems inherent to Excel, it does not meet HIPAA standards for security, including standards related to authorization, authentication and audit controls. Refer to URL: <http://www.ucdenver.edu/academics/research/AboutUs/regcomp/hipaa/> for additional information.

Instead of using Excel or Access, all translational researchers are strongly encouraged to use REDCap (Research Electronic Data Capture) for data entry and storage. Information about using Redcap is available at <http://redcapinfo.ucdenver.edu>. These services are free (underwritten) for translational projects that have COMIRB/IRB approval for any CCTSI investigator at any CCTSI affiliated institution.

Why use REDCAP?

- REDCap provides the ability for you or your delegate to implement clinical report forms and surveys without the need for a programmer.
- Data stored on REDCap are assured of being compliant with COMIRB and HIPAA standards for security.
- By using REDCap, you are protected from data loss, because (1) data are backed up automatically twice daily and (2) changes to data are logged, creating an audit trail indicating which data were changed, by whom, and when. With this information, unwanted changes can be undone if necessary.
- Because REDCap is Web-accessible, co-investigators from other sites can access the data without having to join a virtual private network.
- REDCap provides an expanding set of tools for analysis and visualization of data.
- Local assistance on the UC Denver Anschutz Medical Campus is readily available from REDCap@ucdenver.edu.

REDCap does have some shortcomings, however, and there may be instances in which use of an alternative database management system may be indicated:

- REDCap does not currently have all of the functionality needed for trials whose data will be submitted to the FDA for a new drug or medical device application (i.e., it is not 21 CFR Part 11 enabled). Of course, Excel and Access are even less compliant with FDA studies.
- Because REDCap allows you to easily create data collection forms, it offers limited flexibility in form design. (By contrast, Access allows greater flexibility; however—beware of the temptation to fiddle endlessly with the appearance of Access data collection forms.)
- Data entry into REDCap is currently possible only when a live Web connection to is available.

Note that REDCap is a data collection tool. To generate reports from data that have been collected in REDCap, one must first download the data to an analysis package. REDCap makes such downloads very easy and can produce analysis files for a variety of packages, including SAS, SPSS, R, and Excel.

More-detailed comparison of Excel, Access, and REDCap:

Feature	Excel	Access	REDCap
Summary	Excel is a convenient spreadsheet tool, used for storing, organizing, manipulating, and visualizing data. However, it lacks features important to support data collection and management in health research, making it an inappropriate tool.	Access is a sophisticated data management tool. However, it is not HIPAA-compliant, and database development may require programming assistance. For projects with complex data collection requirements, Access may be an acceptable alternative to REDCap.	REDCap is a secure, web-based, application designed to capture and store health research data. It is HIPAA-compliant, compatible with several statistics packages, and easy to use. REDCap is designed to export data for analysis and reporting, and therefore does not include these capabilities.
HIPAA Compliance			
Secure location - PHI data reside on a secure server	Only if the file resides on a secure server, not on a PC	Only if the file resides on a secure server, not on a PC	Yes. REDCap resides on the university's secure server and is accessed via the internet
Encrypted - data files accessed without proper authentication cannot be read	Manual encryption of file possible	Manual encryption of file possible	Auto-encrypted
Authentication - logins and passwords	No	No	Yes
Authorization - role-based security	Access to data can be set by file owner with installation of MS Information Rights Manager Plugin for Office 2007	Access to data can be set by file owner with installation of MS Information Rights Manager Plugin for Office 2007	Yes. User accounts are controlled by the REDCap Administrator; access to individual databases is controlled by the owner of the database
Audit Trail - created by event logging	No	No	REDCap logs every database change, creating a comprehensive audit trail
Managing Data			
Creation of metadata	No	Yes	Yes
Data validation	No	Yes	Yes
Attaching documents	No	Yes	Yes
Vulnerabilities	Easy to corrupt data by accidentally sorting a column		Limited flexibility in form design
Layout in form view for easy data entry	No	Yes	Yes

Multi-User Collaboration			
Multiple user access to data	No	Record-level locking	Yes, but no locking; last save overwrites previous save
Data Storage			
Flat vs. relational data	Flat	Relational	Flat
Unique identifiers	No	Yes	Yes
File size limitations	255 columns	Very Large	Unknown
Data Import and Export			
Importing data	CSV	ODBC compliant	Excel
Exporting data	CSV	ODBC compliant	CSV, SAS, SPSS, Stata, R (import tool creates syntax files)
Use of Forms and Reports			
Create and use forms	No	Custom	Standard
Export forms	N/A	PDF	PDF
Query data			
Sort and filter data	Moderate	Strong	Very Simple
Create different views and complex queries	None	Strong	None
Data Analysis & Visualization			
Graphing data	Strong	None	None
Calculations	Strong	Moderate	Limited
Convenience			
Remote access	No	No	Yes-web-based
Offline data collection	Yes	Yes	Yes
Portable device compatible	No	No	Yes
Programming assistance required	No-generally easy to set up, familiar to most people	Setup is complex and may require programming assistance	No-very easy to setup and use
Layout in form view for easy data entry	No	Yes	Yes
Other			
Secure file transmission	No	No	Yes, using Send-It feature
Assistance available	No	No	Yes
Cost	Free	Free, but may require programming fees	Underwritten for COMIRB-approved studies. Otherwise monthly fees apply

If you

would like to discuss whether REDCap is suitable for your project with us, please contact us at REDCap@ucdenver.edu.

Storage: Finding a Safe Place for Your Data to Abide

In considering where data abide, keep in mind two types of data, both of which must be stored in a manner that protects the privacy and security of your subjects:

- The original data you have collected, which must be protected from corruption and stored in a robust, audited, and recoverable system.
- Analytic datasets you derived from your original data, which can be manipulated at will using tools like SAS, SPSS, and Excel.

Given the need to protect your original data, it is clear that the hard drive of a desktop computer (much less the hard drive of a laptop or a thumb drive) is completely unsuitable for storage of original data, nor do these comply with existing UCD policies that require encrypted disk drives and thumb drives. If you do not use REDCap, make sure that your data are stored on a secured UC Denver server (generally maintained by your Division or Department). The data on secured servers are backed up regularly, firewalled and password protected, and subject to an audit trail which can identify who has accessed your data. For questions on data storage options, please contact Thomas.Yaeger@ucdenver.edu.

While corruption and recoverability are not as significant issues for analytic datasets, privacy and security remain major considerations. There are severe penalties for privacy breaches, in some cases requiring the University to report breaches to news organizations. **Unencrypted personal health information (PHI) must never be stored outside of a secured server—you as an investigator are personally liable if you store unencrypted PHI on a laptop or thumb drive that is stolen.** Try to perform all your analyses on the secured server if you can—in the event of a suspected data breach audit trails can determine whether a reportable breach has actually occurred. A virtual private network (VPN) connection can allow you to access these data remotely. If you must store analytic data elsewhere, make sure that (1) the data are stripped of all PHI, or (2) the data are encrypted or (3) both. Your department or division should be able to provide you with drive encryption software.