Guidelines for Collecting and Sharing Data

Biostatistics, Epidemiology, and Research Design (BERD) Methods Core

1) Purpose

These guidelines have been created to promote research integrity, protect patient privacy, and to make the data transfer process more efficient. Any research involving human subjects must follow Health Insurance Portability and Accountability Act (HIPAA), Duke Institutional Review Board (IRB) regulations, and all applicable regulatory guidelines.

2) IRB Approval

For human research studies, before allowing access to the data for anyone (including data management and the quantitative methodologist) you will need to add him or her to 'key personnel' in the IRB study protocol. For more information, visit https://irb.duhs.duke.edu/.

For animal studies, the study should obtain approval by the proper governing body at Duke.

3) Data Collection

All data should be stored in a database program that allows for proper reproducible research, data integrity, and data security/protection. REDCap is a widely used data collection program and available for free at Duke. We recommend that investigators involve the quantitative methodologist in the discussion on data collection and database design.

a) Advantages of Using REDCap at Duke

- DOCR's instance of REDCap has been systematically tested and validated.
- School of Medicine supports infrastructure costs associated with DOCR's instance of REDCap, including dedicated servers, daily backups, application updates, and software validation.
- DOCR initial consultations are free. DOCR's data managers and analysts can help you find the best data collection solution for your project.
- Contracting with DOCR is often less expensive than hiring your own staff because you do not have to find, hire, and train staff, or worry about staff turnover.
- DOCR provides training classes in the use of REDCap, available in the LMS system.
- Data can be exported to the biostatistician's preferred format.

Contact DOCR at redcap-docr@duke.edu to discuss the use of REDCap.

b) Data Collection in Excel

The use of Excel for research data collection/management should be avoided if possible (why we avoid Excel). If you need to use Excel, please follow the guidelines below to minimize errors and ensure data quality. If using REDCap, most of these items will be ensured through database data validation processes.

1. Every patient/subject should have a unique identifier.
2. Avoid all use of commas in data fields. This includes both text and numeric fields. For example, use 1298 rather than 1,298.

3. Do not include line breaks within a cell.

4. Keep column/variable names under 32 characters, while keeping each one unique. Do not start a column/variable name with a number or symbol. Column/variable names should not include spaces or special characters.

5. Dedicate the top row only for the column/variable name; do not repeat rows of column/variable names.
6. If there are several groups of patients, use a separate column to identify group membership for each patient. Do not indicate any distinguishing patient characteristic with highlighted cells. Instead incorporate a separate column to indicate the characteristic. The following spreadsheet contains group information in the column labeled 'Arm'. In this case the patients were randomized to one of two treatment arms: experimental (Exper) or control (Control).

DukeID	**Arm**	**Value**
A0001	Exper	5
B0002	Control	4
C0004	Exper	2
D0005	Control	6
E0006	Control	8

7. Use the same format for all variables in a column. If a variable is to be analyzed as numeric then all entries in that column must be numeric. Any characters or symbols including '<', '>', '=', '*', '?' etc. are not permitted.

8. For missing data, leave the cell empty to indicate a missing value; do not use 'N/A'.

9. For character variables, be consistent with the letter case and exact cell content. For example, yes, Yes, and YES are all considered different responses. Spaces are considered characters; 2 spaces between characters are different than 1 space.

10. For variables with the same response options (such as yes/no), use consistent coding. Do not code one variable as '1=yes, 0=no' and another variable as '1=no, 0=yes'.

11. Do not include blank rows or columns.

12. Do not hide rows or columns of data instead of deleting them as they will still be imported into the statistical software.

13. Do not include summary data in the data file.

14. Do not include comments or footnotes in the data file. Comments or explanations of variable names, study design, data collection, any irregularities that occurred during the study or data collection are encouraged, but they should be listed in a separate document.

15. Data with repeated measures can be collected in long format or wide format (see examples below). For long format, each patient has a row of data for each time measurement and all observations from the same patient are indicated with the unique identifier. For wide format, each patient has one row of data and repeated columns of measurements.

Long format

DukeID	Time	Marker
3	1	34
3	2	23
3	3	45
4	2	35
5	1	27
5	3	76

Wide format

DukeID	Marker1	Marker2	Marker3
3	34	23	45
4		35	
5	27		76

16. If there are corrections to the data, it is the responsibility of the investigator to provide the statistician with an updated file as soon as possible. Please include an explanation why data correction was warranted.

c) Example data in Excel

Incorrect data collection example: This spreadsheet breaks many of the rules above and would require a lot of time for the quantitative methodologist to clean the data.

Data for ARC trial						
	Treatment	1st Date	Age of Subject	Patient's Gender	Height at baseline	*blood pressure
	1	Oct 25th, 2019	44	m	67"	120/82
	2	7/5/2018	62	Female	5'10"	>150/90
	3	28-Feb-2018	30-34	female	182cm	135 over 85
	4	6-Apr-18	22.5	male	74	normal
	5	9/12/2017	69	F	5.5	160/110
	Control					
	1	6/22/2018	65+	femlae	5ft3	130/60 130/70
	2	December 26, 2019	73	Male	unknown	140/75
	3	8/5	49	M	66	N/A
	4	4/12/2019	60 1/2	MAle	~61	SBP: 110 DBP: 60
	5	10/16/18	??	f	6'3	80/120
Average					68	
	*collected at baseline					

Good data collection example: This spreadsheet requires very little data cleaning, so the quantitative methodologist will be able to get to the analysis more quickly.

ID	Arm	First_Date	Age	Gender	Height	SBP	DBP
1	0	10/25/2019	44	M	67	120	82
2	0	7/5/2018	62	F	70	150	90
3	0	2/28/2018	32	F	72	135	85
4	0	4/6/2018	22	M	74	120	80

5	0	9/12/2017	69	F	65	160	110
6	1	6/22/2018	66	F	63	130	60
7	1	12/26/2019	73	M		140	75
8	1	8/5/2017	49	M	66		
9	1	4/12/2019	60	M	61	110	60
10	1	10/16/2018		F	75	120	80

4) Sharing/Sending Data

- Information about all the permitted data storage resources is on ISO's website: https://security.duke.edu/policies/duke-services-and-data-classification.
- Duke Box (NOT Drop Box) is the preferred way to share data. It is only for the transfer of data and should not be used for long-term data storage.
- Sharing research data via email, even with send secure is discouraged and should be avoided.
- Any questions pertaining to data sharing or storage can be directed to the DHTS Information Security Officer (ISO) at security@duke.edu or the research practice manager.
- If data are not shared or stored in an approved manner, a violation of the protocol must be reported to the IRB in a timely manner.