REDCap

Research Electronic Data Capture

Best Practices

THE UNIVERSITY OF CHICAGO MEDICINE & BIOLOGICAL SCIENCES

CENTER FOR RESEARCH INFORMATICS

Think about your data

Formally map every piece of data you collect to your analytic plan and/or reporting requirements, and be sure that the data collection will supply everything you need. This will help avoid the two related ills of collecting too much data and missing critical data.

Think about your data in terms of the number of variables you are trying to capture. You should, obviously, try to get as much data as you need to prove or disprove your hypothesis. However, if you set out to get too much data, then you might find yourself swamped in a sea of irrelevant and/or uninformative details. 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.

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 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 Biostatistics Laboratory of the Department of Health Studies.

Describe the expected input data as much as possible

Provide as much information as possible to assist in data entry. Use the Field Label and Field Notes to describe exactly what kind of data you intend to capture in a given data entry field. Do not assume the data entry person knows the expected data, units, or formats for each field. Use Field Notes mainly to supplement the Field Label information, for example use Field Notes to specify the expected format of a validated field or the expected units of measurement.

Keep a codebook

A codebook describes each variable by name according to the type of data – numeric, date/time, character – the units of measurement – grams, feet, micrograms per deciliter – the purpose of collecting it and its relationship to other data. You can use your REDCap project's codebook as a starting point. The codebook is a human-readable, read-only version of the project's Data Dictionary and it is found on your project's Home Page under Quick Tasks.

Use the REDCap identifiers function

There are 18 pieces of information that must be marked as identifiers in a REDCap data dictionary, as per HIPAA policies.

1. Name
2. Fax number
3. Phone number
4. E-mail address
5. Account numbers
6. Social Security number
7. Medical Record number
8. Health Plan number
9. Certificate/license numbers
10. URL
11. IP address
12. Vehicle identifiers
13. Device ID
14. Biometric ID

15. Full face/identifying photo
16. Other unique identifying number, characteristic, or code
17. Postal address (geographic subdivisions smaller than state)
18. Date precision beyond year

Unique Records

The first variable on the first form should be the record identifier (e.g. Participant ID) because it will be used by REDCap as a key variable linking forms for a particular record. The default variable name is "Study_ID".

Capture consent information

Use a form to capture consent information. Capture information such as whether or not the subject was consented, who consented the subject, the date the subject signed the consent form, whether or not the subject was given a signed copy of the consent form and upload the signed consent form.

Reduce the use of free-text fields

Minimize the use of free text fields because these can be difficult to analyze. Use categorical response field types (i.e. dropdown, radio button, checkbox) instead of free text fields (i.e. text box and notes box). The use of multiple choice field types will improve later data analysis. You can augment a categorical response with a text box or notes box to capture additional information.

Do not mix data types

It is possible to mix data types in data entry fields. For example, a researcher might enter a numerical code followed by a comment on the code such as "147 Patient had a cold." Both the code and the comment might be informative but they should be placed in separate data entry fields (with data validation where applicable).

Use data validation whenever possible

When using text box fields, use validation types and set minimum and/or maximum values as much as possible for better data accuracy. In particular, always use date validation for dates. Also, when collecting participant email addresses for surveys, use the email address validation to collect accurate addresses.

For inexact dates, enter day, month, and year separately

Most people will be able to tell you their date of birth. However, very few of them will be able to tell you the date they first noticed symptoms of a particular disease. However, they might be able to tell you the month and year they noticed symptoms. For this type of data, consider entering the month, day and year in separate columns. For example, a patient might not be able to tell you when he first came down with measles as a child. However, he might be able to limit the range of dates to say, March of 1986. You could enter this into a database as:

Day	Month	Year
unknown	3	1986

While the day is missing, the other parts of the date might prove to be useful.

Identify units of measurement

Units for measurements should be clearly identified whenever possible. Avoid abbreviations of units of measurements and never mix different units in one data entry field.

Field Notes

Use Field Notes describing units, formats, etc. whenever appropriate. Do not assume the data entry person knows the expected units or formats. For example, use Field Notes to be sure everyone knows whether height is measured in feet and inches or meters.

Use standard measures and codes

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.

Use the REDCap Shared Library as one resource for standard instruments. For demographics and health status, consider using instruments from national agencies such as CDC (e.g. BRFSS) and AHRQ (e.g. NHANES).

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 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.

Consider recording race and ethnicity according to the current NIH guidelines. The ethnicity categories are "Hispanic or Latino" and "Not Hispanic or Latino." The racial categories are "American Indian or Alaska Native," "Asian," "Native Hawaiian or Other Pacific Islander," "Black or African American," and "White." More detailed information about them can be found at:
http://grants.nih.gov/grants/guide/notice-files/NOT-OD-01-053.html.

Be consistent when assigning numerical codes

For example, if "unknown" is coded as 99 in one response, it should be coded as 99 wherever it appears in the database. Note that generally Yes = 1 and No = 0. The numerical code does not affect the order that choices are displayed in the REDCap data entry form.

Avoid missing values

Does a blank value mean you still need to collect that value, you forgot that value, the value is not available, or the value is not applicable? Missing values might be missing for several different reasons and the reasons they are missing might be relevant to your study. If the researcher's approach to missing values is just to leave a blank in a data set when there is no value, then that researcher unknowingly might be throwing away a chance to collect useful data.

For example, a person might have a missing value for a thyroid scan. It might be missing because the subject still needs to take the test, the subject forgot to take the test, you forgot to record the value, or because the person does not have a thyroid gland. Also consider whether yes or no are really the only choices for a given question. For example a simple yes/no answer may not suffice for questions such as "Have you ever had disease X?" because patients might answer not only yes or no but also "I don't know" or "I was tested for that once but I do not remember the results" or even "I do not want to answer that question."

If data are missing or unknown, you can include reasons in your categorical responses or mark the data as missing and include a text box to record the reason for the missing value.

Group related variables on short forms

Form Names correspond to individual data entry web pages. Forms are groupings of variables within the database. Put variables collected together on the same form to improve data entry workflow. For example, putting demographics together and labs together on separate forms makes data entry more reliable. Keep forms fairly short to minimize risk of data loss (by saving more often when completing a form). If possible, group field types that minimize changing from keyboard to mouse.

Data Entry

Group variables together that follow the data entry work flow, and use field types that minimize changing from keyboard to mouse. For example, you can enter a dropdown field option by typing the first character of the label, allowing you to “tab and type” through the data entry fields, while radio buttons require using the mouse to select an option). Keep forms fairly short to minimize risk of data loss (by saving more often when completing a form) and make it easier to identify data entry errors.