UCDAVIS HEALTH | Clinical and Translational Science Center

Best Practices for REDCap Database Creation

This document—borrowed and revised from the University of Colorado, Denver—provides general guidelines for the design of REDCap databases. Although the REDCap team will assist you with the design and creation of your database, many of the steps are best performed by the research team. Thus, please review this document before you start working on your database.

When you are ready to learn more about REDCap and get started, please contact the REDCap Team at hs-redcap.support@ucdavis.edu.

Important Concepts

REDCap automatically builds forms and the underlying database from a user-created "data dictionary" that is created in Excel. Thus, the contents and design of the form is controlled by columns in the Excel sheet. Most importantly, REDCap creates the web-based forms and associated database so the user does not need to worry about these technical details—your research team can focus on your research questions and how the data entry forms should be designed to best capture the data that you need. To improve the quality of your data, consider the following concepts when creating your data dictionary.

Longitudinal and Traditional Databases

REDCap makes it easy to create a study with only one event or a longitudinal study with multiple events. For longitudinal studies, you only need to create the form once, and then later specify in REDCap that the form will be used to collect data for multiple data collection periods. For example, you might construct a simple form to capture blood pressure and other vital signs, and then use the matrix in REDCap to assign the form to five different time periods in which data will be collected.

Collect the right data

Before building your REDCap database, think carefully about your research questions to decide which variables you will need to collect. Make sure that you will collect all the data that you need, but avoid collecting data that will not be necessary.

It may be important to consult with a statistician to review your research questions and then discuss the type of data that should be collected to best evaluate them.

Data Entry

When designing the data dictionary, think about ways to help the user efficiently enter data. One important concept is to group variables together that will 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.

It is also important to describe the input data fields so that data entry staff are sure about what they are entering. Use clear Field Labels and Field notes to describe exactly what should be entered.

Database tips

Avoid “free” text fields

Use categorical response options (yes/no, multiple choice) whenever possible. When using text fields, add field validation (format type, valid range) whenever possible for better data accuracy. Minimize use of “free” text fields because these can be difficult to analyze. For example, if you ask people to specify their race and ethnicity within a text field you may get a variety of answers that do not neatly fit into categories that you were expecting.

Use standard measures and codes

If available, consider using existing measures so that your research will be able to be compared to other studies.

Do not mix data types

Although it is possible to mix data types within entry fields, it is usually a good idea to put the information into separate columns. For example, if there is a medical code and a comment (e.g., “428.0 heart failure patient had pneumonia”) put the code and comment in separate fields.

Use validation rules

To improve data quality, use REDCap validation rules. For example, set minimum and maximum values that can be accepted. In addition, use rules to ensure that valid dates are entered.

Reduce the amount of missing data

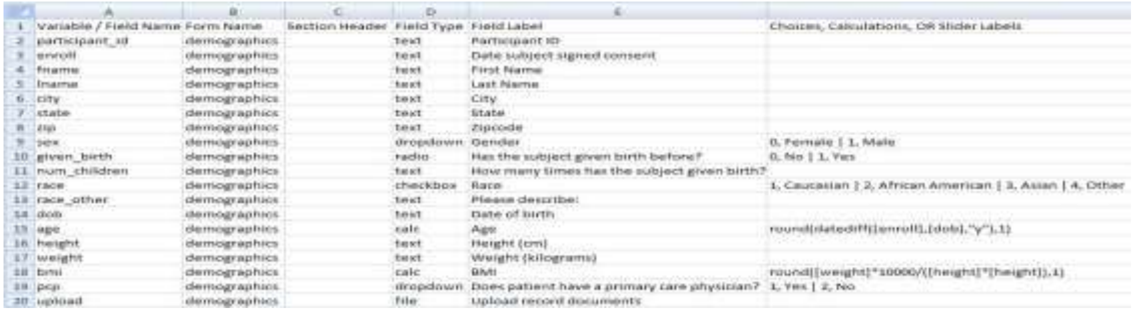
Missing data can substantially reduce your total analysis sample for many statistical analyses. Thus, work hard to avoid missing data when possible. Of course, missing data is a normal part of research so plan for this in your database. First, try to avoid the use of blanks in your database. For example, it is impossible to know if the entry is blank because the entry staff forgot to report it, there was no response from the patient, or the patient has not yet received the outcome. Consider including an option in your dropdowns or checkbox fields to show the reason for missing data. For more complex issues, you can include an option for missing, and then using branching logic, bring up a text field to specify why the data were missing.

Consent Information

Consider creating a form to capture consent information. Important fields include whether or not the subject consented, consent date, who consented the subject, and if the subject was given a signed copy of the consent form.

Creating a Data Dictionary

REDCap allows users to create a database using its “Online Form Editor” or with an Excel spreadsheet. We recommend that you use the Excel method. In general, it is easier to create a data dictionary in Excel and then upload this into REDCap since you do not have to create similar variables individually, but can take advantage of using copy/paste functionality. To assist you in learning how to create a data dictionary in Excel, we will provide some Data Dictionary demonstration files. These files show entries for each required column, as well examples of calculated fields and branching logic. The example, such as the one below, are simple spreadsheets most people are familiar with.



	A	B	C	D	E	
1	Variable / Field Name	Form Name	Section Header	Field Type	Field Label	Choices, Calculations, OR Slider Labels
2	participant_id	demographics		text	Participant ID	
3	enroll	demographics		text	Date subject signed consent	
4	fname	demographics		text	First Name	
5	lname	demographics		text	Last Name	
6	city	demographics		text	City	
7	state	demographics		text	State	
8	zip	demographics		text	Zipcode	
9	sex	demographics		dropdown	Gender	0, Female | 1, Male
10	given_birth	demographics		radio	Has the subject given birth before?	0, No | 1, Yes
11	num_children	demographics		text	How many times has the subject given birth?	
12	race	demographics		checkbox	Race	1, Caucasian | 2, African American | 3, Asian | 4, Other
13	race_other	demographics		text	Please describe:	
14	dob	demographics		text	Date of birth	
15	age	demographics		calc	Age	round(datediff([enroll],[dob],"y"),1)
16	height	demographics		text	Height (cm)	
17	weight	demographics		text	Weight (kilograms)	
18	bmi	demographics		calc	BMI	round(([weight]*10000/([height]*[height])),1)
19	pcp	demographics		dropdown	Does patient have a primary care physician?	1, Yes | 2, No
20	upload	demographics		file	Upload record documents	

Important Note: *Although you will be working in Excel, the file type you work with is actually a ".csv" (comma separated variables) file. When saving your data dictionary, be sure to select the .csv format to upload to REDCap.*

General Best Practices

Note: The term "Column" here refers to Excel spreadsheet data dictionary columns. Also, the terms "field" and "variable", as used here, are essentially interchangeable. Both terms refer to a unique item of data to be collected and analyzed. "Field" is a database term, while "variable" is a data analysis term.

- 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". Demographics is normally the first form, but this is not required. All new projects are provided with a sample Demographics form, but you are free to modify or replace this.
- Use categorical response field types when possible to reduce risk of data entry error (dropdown, radio button, checkbox). If these fields are not feasible, use text fields with validation (date, phone, email, integer, number) whenever possible to reduce the use of free-text fields.
- When using a text field with validation types of number or integer, define range minimum/maximum as much as possible to allow REDCap to perform basic data validation/quality control.
- Put variables collected together on the same form to improve data entry workflow. Putting demographics together and labs together on separate forms makes data entry more reliable.
- Include Field Notes describing units, formats, etc. whenever appropriate. Do not assume the data entry person knows the expected units or formats.

Data Dictionary Spreadsheet Columns

This section describes the function of each column in the data dictionary spreadsheet, and whether or not it is required or optional.

1. Column A - Variable/Field Name (Required)
   - Variable/Field names specify the variable name that will be used in reporting, data export, and data analysis. They are not displayed on the data entry form.
   - Variable names:
     - may contain letters, numbers, and underscores, but no spaces or special characters.
     - cannot start with a number.
     - must be unique, and cannot be repeated within a database, even in different forms.
   - Variable names should be brief; they do not need to be descriptive given that the Variable Label will be applied . A common example is the use of “dob” as a variable name, with the corresponding variable label of “Date of birth.”
   - If you change a variable name in one place, you must change it everywhere it is used (e.g., calculations, branching logic).

2. Column B - Form Name (Required)

   a. Forms are groupings of variables within the database. It’s a good idea to divide your variables into several fairly short forms for ease of data entry, and to provide more opportunities to save data at the end of each form.

   b. Form names must be all lowercase in the Excel spreadsheet, but will be displayed in REDCap with initial capitals. If your form name contains more than one word, connect the words with an underscore, such as “form_name”. The underscore will appear as a space in REDCap.

   c. All variables in a form must be in adjacent rows in the data dictionary. For example, you cannot have a variable in row 6 be in the “demographics_form”, a variable in row 7 be in the “first_visit” form, and then a variable in row 8 back in the “demographics_form”. Variables will appear on the form in the order they appear in the data dictionary.

3. Column C – Section Header (Optional)

   Section Headers are used to visually separate items within a form, primarily to aid data entry. If you are entering data directly into REDCap while interviewing a study participant, you may also want to use Section Headers to display interview script between questions, e.g. to introduce a new topic.

4. Column D - Field Type (Required)

- Specifying the field type determines what types of responses are allowed, and how they will be displayed. Field types include: dropdown, radio button, checkboxes, text box, note box, calculated field, file upload, and section header.
- Categorical field types (dropdown, radio buttons, checkboxes) must also have response options (choices) defined in Column F. Terms used in Column D to define these field types are: *dropdown*, *radio*, *checkboxes*.
- Text field types (text box or note box) should have validation (Column H) whenever possible. If the validation is "integer" or "numeric", you should also include the allowable minimum and maximum values (Columns I & J). Text variable cannot also have choices listed in Column F. Terms used in Column D to define these field types are: *text*, *notes*.
- Calculated variables display the result of a calculation based on responses to previous variables. Data cannot be entered in calculated fields. The term used to define this field type in Column D is: *calc*.
- The file upload field type allows you to attach a document (e.g. consent form) to the record. The maximum file size for any document is 50Mb. The term used to define this field type in Column D is: *file*.

5. Column E – Field Label (Required)

a. A Field Label (or variable label) is a word or phrase that is more descriptive than the variable/field name. It is displayed on the form—instead of the variable name—because it provides more information to the reader.

6. Column F – Choices, Calculations OR Slider Labels (Required)

a. All categorical field types (yes/no, dropdown, radio buttons, checkboxes) must specify response options associating numerical values with labels. For example, Yes=1, No=0.

b. All calculated fields must specify the calculation here. Examples of calculation syntax can be found in the REDCap Help Section and in the demonstration data dictionary.

c. The slider field allows you to label three anchor points: left, middle, and right. An example might be "Strongly Disagree", "Neutral", and "Strongly Agree."

7. Column G – Field Note (Optional)

Optional

Field notes are used to provide information to assist in data entry. Examples are specifying the expected format of a validated field (e.g. phone number), or units (e.g. kg vs. lb).

8. Column H – Text Validation Type OR Show Slider Number (Optional)

   a. Format validation types for text fields are: date, time, integer, number, zipcode, phone, and email. An error message is displayed if an entry does not match expected format.

   b. For slider fields, specify whether to display or hide the value (1-100) selected on the slider.

9. Columns I & J – Text Validation Min/Max (Optional)

   For text validation types of number or integer, minimum and maximum acceptable values may also be specified. An error message, including the acceptable range, is displayed if the entry is out of range.

10. Column K – Identifiers (Optional)

    To be HIPAA-compliant there are 18 pieces of information that must be marked as “identifiers” in a REDCap data dictionary.

    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

11. Column L – Branching Logic (Optional)

Branching logic can be applied to a field to specify whether or not it will be displayed, depending on values in previous fields. For example, a question about pregnancy can be designated to be displayed only if the subject if female. Syntax for branching logic is described in the REDCap Help Section and in the demonstration data dictionary.

12. Column M – Required Field (Optional)

A field can be designated as “required” so that it must be completed before moving on to the next field. An error message is displayed if the field is left blank.

13. Column N – Custom Alignment (Optional)

The location of text boxes or categorical responses (dropdown, radio, checkbox) can be specified as Right/Vertical, Left/Vertical, Right/Horizontal, Left/Horizontal. The default setting, if not specified, is Right/Vertical.

14. Column O – Question Number (surveys only)

Optional

REDCap can be set to auto-number questions on a survey. However, if you want a custom numbering scheme, you can specify each question number here.