Guidelines for Collecting Data via Excel Templates

QHS

Quantitative

HEALTH SCIENCES

"We aim to make your studies significant"

Table of Contents

1.0 Introduction

This document describes how to optimally develop a data base in MS Excel. With these guidelines

- the data collection process may be streamlined,
- accuracy and consistency of the data may be increased,
- higher data quality may be achieved, and
- data analysis facilitated.

This document is broken down into three main areas; philosophy of developing a database, general layout and data collection considerations. A Checklist is included in Appendix A. Appendix B gives a list of other documents you may find helpful. They are available on the QHS website http://www.mcw.edu/Quantitative-Health-Sciences/Resources.htm.

2.0 Philosophy

2.1 Development of framework for database

When developing a database there are several general considerations:

- What is the purpose of the data collection?
    - What are the aims?
- What will be the source(s) of the data?
    - Data entry?
    - Data uploads?
- What variables will be needed?
    - How are the variables related?
        - Do they naturally divide into "forms"?
        - Are there repeated measurements?
    - What will be the format and values?
- How will security be maintained?
    - How and when will data be unidentified?

2.2 Data Dictionary

The data dictionary describes the variables and values that variables can take. It is best to create your data dictionary first and then set up your template. The data dictionary can be circulated to other investigators to ensure there is agreement on the data entered. This may save time for the person(s) abstracting data from charts or entering the data into the spreadsheet. It may be helpful for your statistician working with your data as well, especially if the variable names are not as meaningful to the statistician. Some examples are included below:

- For all variables "-9" indicates "missing" and "-8" means "not applicable."
- ID: 4-digit number.
- Sex/gender: M=Male, F=Female

It's possible to set up the data dictionary to be used for data validation as well. See section 4.3.

2.3 Maintenance of database

When maintaining a database there are several factors for consideration:

- Who is responsible for maintaining the database?
- Is there a procedure for modifications?
    - Will there be one correct database?
    - How will an audit trail of corrections be maintained?
    - Will any changes be kept in a database with a new name?

2.4 Monitoring the contents

Reports and checks can improve data quality. In developing a monitoring plan the following should be decided:

- How often will summary reports be generated
  - Do not wait until the end of the project to generate summary reports.
  - Reports should be generated early and often in order to identify problems in the data entry or collection.
- What will reports be for?
  - Tracking enrolment?
  - Reasons for inclusion/exclusion?
  - Subject dropout?
  - Death?
- The schedule for tracking reports?

It is always worthwhile consulting at the beginning with your statistician and data manager expert on these details.

3.0 Setting up a general template

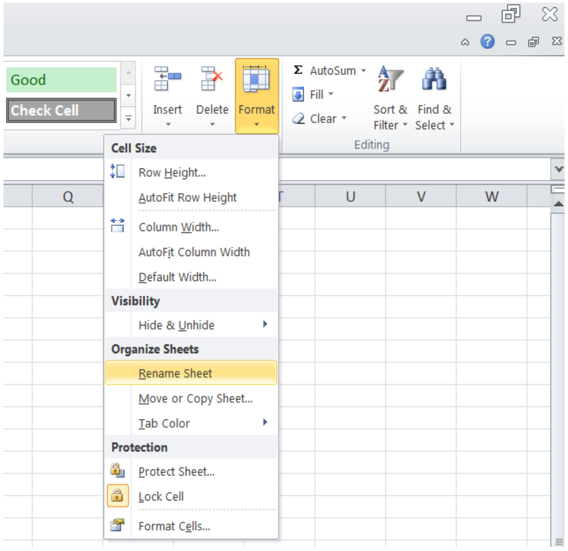
3.1 Merged cells only for non-data collection

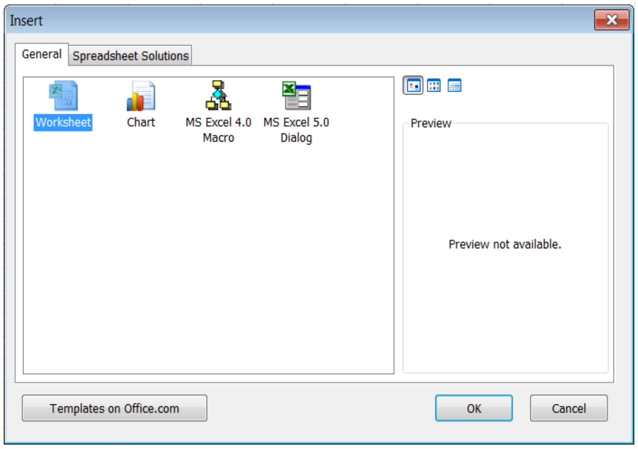
Merged cells should only be used for formatting purposes (i.e. within a non-data collection section of the template). For example, merging the cells within the table header would be fine, but having merged cells in the active area of the form where data is being collected would not be fine. During the analysis phase any cells that have been merged that lie within the active data collection area of the form will need to be unmerged, which has the potential for data loss.

3.2 Naming and inserting worksheets

If there are different sections to a template, it may be easier to have each section in its own tab of the Excel file. Each tab in the Excel file should have a clear and concise name that describes what it is. Keep in mind that Excel has a 31 character limit on the size of the name.

To rename a tab in an Excel file click on the “Format” drop-down menu located in the main ribbon. Then select “Sheet”, and then “Rename”. To add a new worksheet right click on any of the tab names, followed by clicking on “Insert” in the pop up menu, then select “Worksheet”. Excel will add another worksheet to the book.

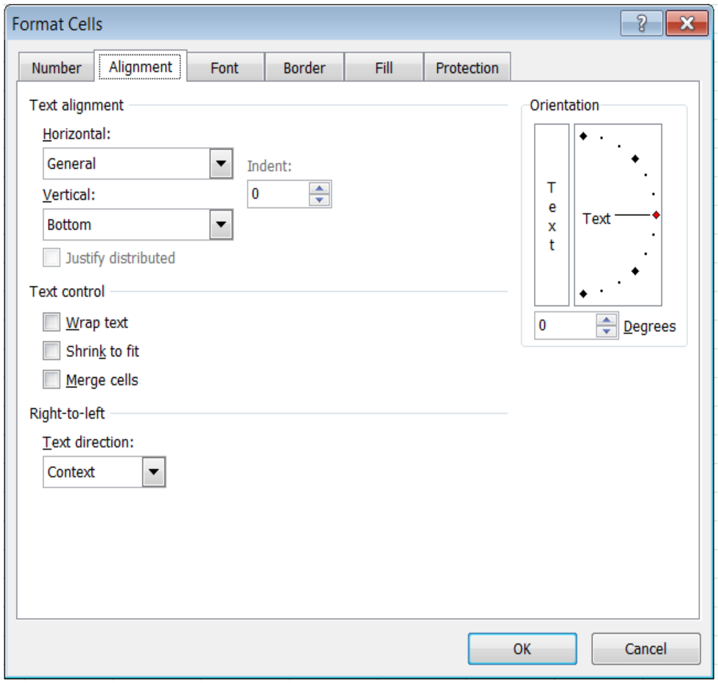




Renaming Excel Sheet   Insert Worksheet

3.3 Need to use header rows

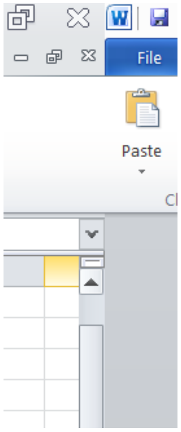
All templates must have a header row(s) clearly identifying what data will be stored in each column of the Excel file. It is alright to change the alignment of the text in the cell to conserve space. To change the alignment right click the cell and select “Format Cells” and then click on the “Alignment” tab of the pop up form. The “orientation” will change the alignment of the text while the “Horizontal” and “Vertical” text boxes will change the positioning of the text in the cell.



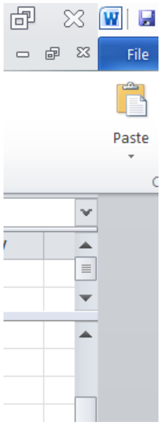
Horizontal: General
Indent: 0
Vertical: Bottom
□ Justify distributed
□ Wrap text
□ Shrink to fit
□ Merge cells
Text direction: Context
0 Degrees

Text alignment

For ease of use during data entry the header row(s) can be split into its own pane allowing users to scroll through the spreadsheet and still be able to match up data cells with the header information. To split a heading row click the split button located on the upper right hand of the worksheet and drag to the desired location.



Split button



Split button dragged

3.4 Avoidance of use of color

The use of color to highlight the importance of text or specific sections should be avoided.

3.5 Reviewing the final layout

After the template has been set up final visual template checks should be made to make sure that all necessary variables have been captured in the template. It may be helpful to view the template in Print Preview. To view a document in Print Preview click on the “File” drop-down menu and then click on “Print”.

4.0 Data collection form considerations

Using the data dictionary the following practice should be observed.

4.1 Formats in Excel

- Excel automatically formats data.
- A common problem occurs when numeric fields or date fields are formatted as a character.
  - This often happens when the first row contains missing data.
  - Check the formatting and specify the appropriate formats. Some recommended formats are included below:
    - Dates DD/MM/YYYY
    - Times H:M:S PM

4.2 Variables

A variable is a characteristic of a unit being observed that may assume more than one of a set of values to which a numerical measure or a category from a classification can be assigned (e.g. income, age, weight, etc., and "occupation", "industry", "disease", etc.).

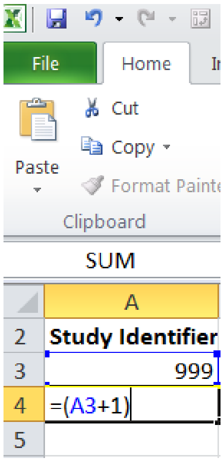
- The variables should be aligned with a header column in the template and only hold one value. See section 2.2 for how to make sure to do this.
  - Any time multiple pieces of information make up what we think of as one piece of information (i.e. blood pressure) it needs to separated out into its distinct pieces and stored separately in different variables.
  - An example of a common variable that appears to be one unit of data is date/time. A date/time should really be stored in two separate variables, one for date and one for time.

See section 4.2.2 for more details on how variables hold data.

4.2.1 Identifiers

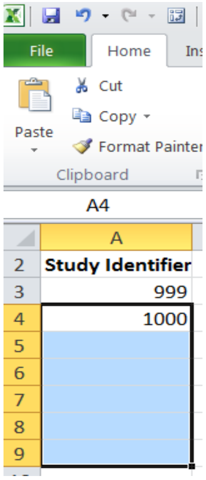
Each individual record needs to have a unique identifier that is specific to just that record.
Identifiable data such as MRN numbers should not be used as the identifier in the final dataset.

To create a basic study identifier in Excel that can be used for each new record that is added, type the following formula into the cell following the first entered identifier value, and then copy and paste the formula. More elaborate identifiers can be created, contact QHS for assistance.



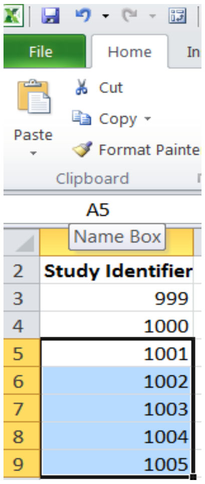
	A
2	Study Identifier
3	999
4	=(A3+1)
5	

Excel formula



	A
2	Study Identifier
3	999
4	1000
5	
6	
7	
8	
9	

Copy & paste formula



	A
2	Study Identifier
3	999
4	1000
5	1001
6	1002
7	1003
8	1004
9	1005

Results of the copy & paste

4.2.2 Categories

There are two data storage categories for variables and they function differently during the data analysis phase of the study. There are two data storage categories for variables and they function differently during the data analysis phase of the study.

- The Mutually Exclusive category describes variables where there is only one value being collected (i.e. “Select one of the following” types of questions). Examples:
    - Yes or No questions
    - True or False questions
    - Select one response from multiple values (i.e. 1 = low, 2 = Medium, 3 = High)
- The Non-mutually exclusive category describes the “Select all that apply” type of questions.
    - For example, a patient’s tumor was detected via any or all of the following: physician exam; radiographic imaging; or laboratory evaluation. In this case, three variables need to be created with 1=Yes or 0=No to indicate if the given method was used.

Study ID	Q4.Tumor_detected_**pe**	Q4. Tumor_detected_**ri**	Q4. Tumor_detected_**lab**
999	0	1	1
1000	1	1	0
1001	1	1	1

- For questions where “Other” is a choice, whether in a mutually exclusive or non-mutually exclusive format, the variable should not be an open ended text box. Instead have two variables, one for “Other” and one where a specific response can be typed in.

4.3 Data validation within the form

To assist with data entry and to decrease common data issues (different spellings or abbreviations, upper/lower case usage, values that are out of range, etc.), data validation within the template should be used. Data validation is used to control the type of data or values that users can enter into a cell. It can be used to allow only a certain range of dates, limit choices by having the data entry person select from a list or allowing only numbers to be entered. This step assures data consistency and reduces time needed for data cleaning during the analysis phase of the project. There is a lot that can be included in a template and only a few basic validation configurations are covered below. Check the Excel help file or contact QHS for more information.

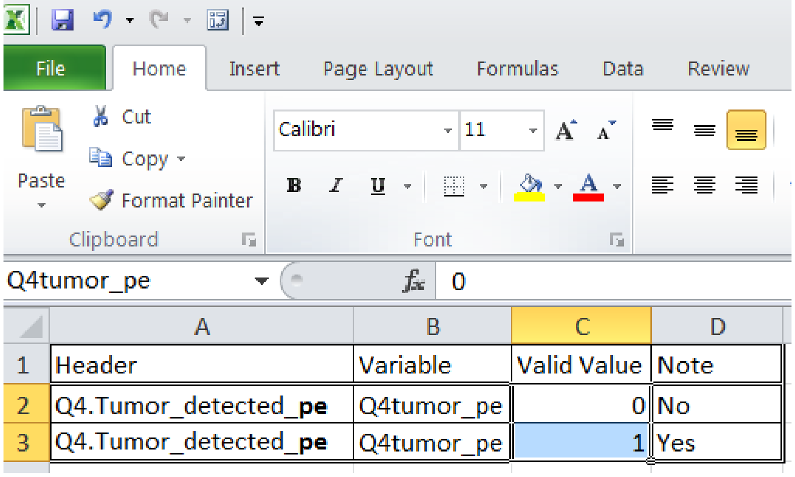
4.3.1 Drop-down List

Creating a drop-down list is a multi-step process. First a list of valid entries needs to be created. The drop-down list of valid values is compiled from cells elsewhere in the workbook. To create the valid values list select a tab in the workbook that is not being used to collect data. Add columns for the header, variable name, and the valid value. A note field can be added as well.

Header	Variable	Valid Value	Note
Q4. How was the Tumor Detected; Choice Physical Exam	Q4tumor_pe	0	No
Q4. How was the Tumor Detected; Choice Physical Exam	Q4tumor_pe	1	Yes

The next step is to define a “Name” for the valid values.

- Select the cell or range of cells in the Valid Value column that apply to the variable in question;
- Click the “Name” box at the left end of the formula bar;
- Type a name for the cells and press enter;

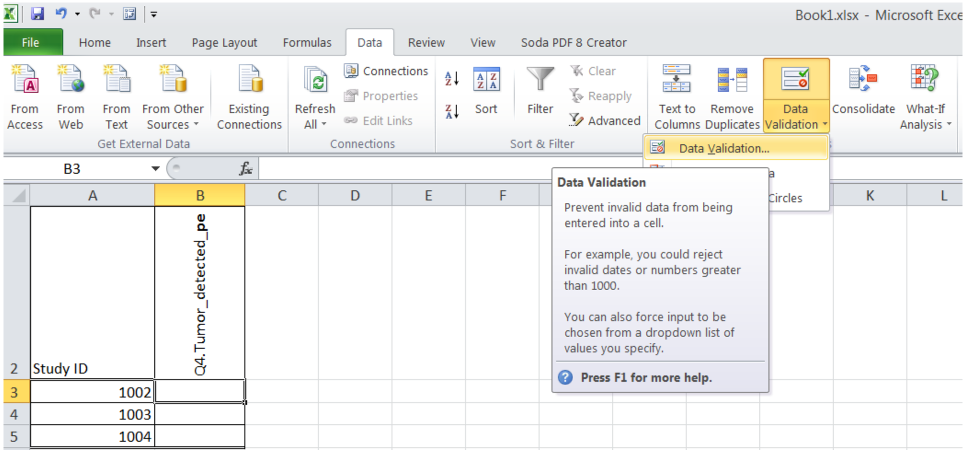


Q4tumor_pe | f_x 0

	A	B	C	D
1	Header	Variable	Valid Value	Note
2	Q4.Tumor_detected_pe	Q4tumor_pe	0	No
3	Q4.Tumor_detected_pe	Q4tumor_pe	1	Yes

“Name” box on formula bar, valid values highlighted

- Select the cell where the drop-down list needs to be created;
- On the Data tab, in the Data Tools group, click Data Validation;



B3

	A	B
2	Study ID	Q4.Tumor_detected_pe
3	1002	
4	1003	
5	1004	

Data Validation

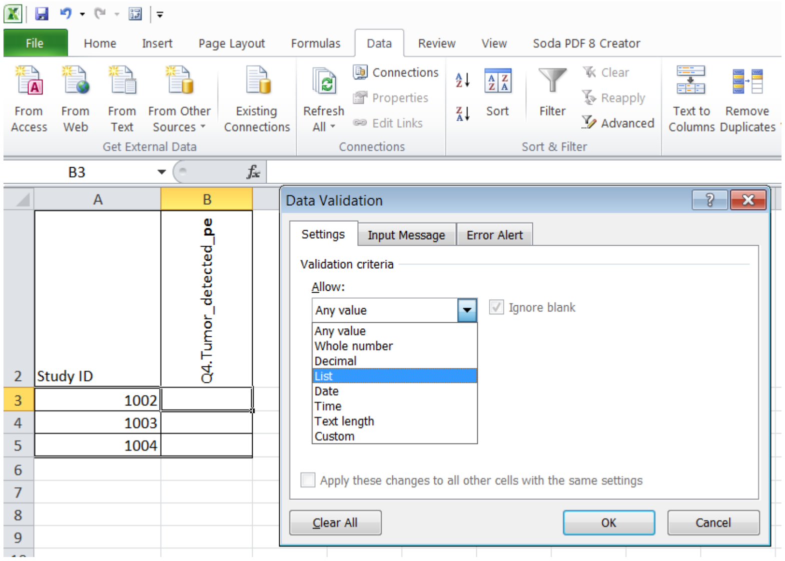
Prevent invalid data from being entered into a cell.

For example, you could reject invalid dates or numbers greater than 1000.

You can also force input to be chosen from a dropdown list of values you specify.

Press F1 for more help.

- In the Data Validation dialogue box click the Settings tab;
- In the Allow box, click list;



	A	B
2	Study ID	Q4.Tumor_detected_pe
3	1002	
4	1003	
5	1004	

Validation criteria

Allow:

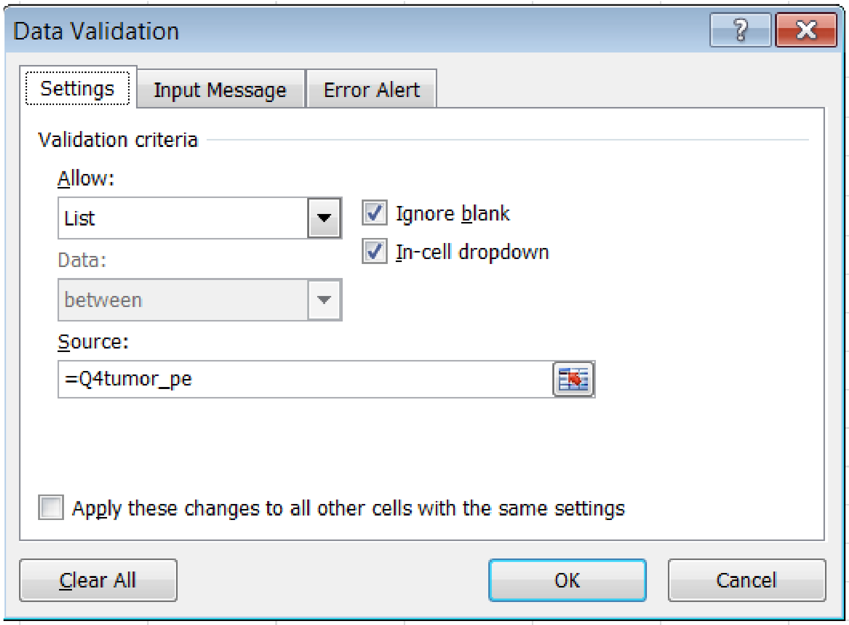
Any value

Any value
Whole number
Decimal
List
Date
Time
Text length
Custom

☑ Ignore blank

□ Apply these changes to all other cells with the same settings

- Specify the location of the list of valid entries by clicking the Source bar and then highlighting the cells of that contain the valid values, or type in the valid list name into the Source box;



Validation criteria

Allow:

List

☑ Ignore blank

☑ In-cell dropdown

Data:

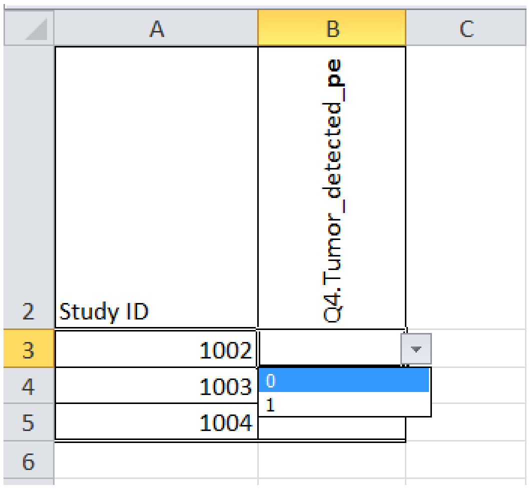
between

Source:

=Q4tumor_pe

□ Apply these changes to all other cells with the same settings

- Check the “Ignore Blanks” and “In-cell dropdown” boxes;
- Excel has a default error message (“The value you entered is not valid”) that can be used;
- Click OK.



	A	B	C
2	Study ID	Q4.Tumor_detected_pe	
3	1002		
4	1003		
5	1004		
6			

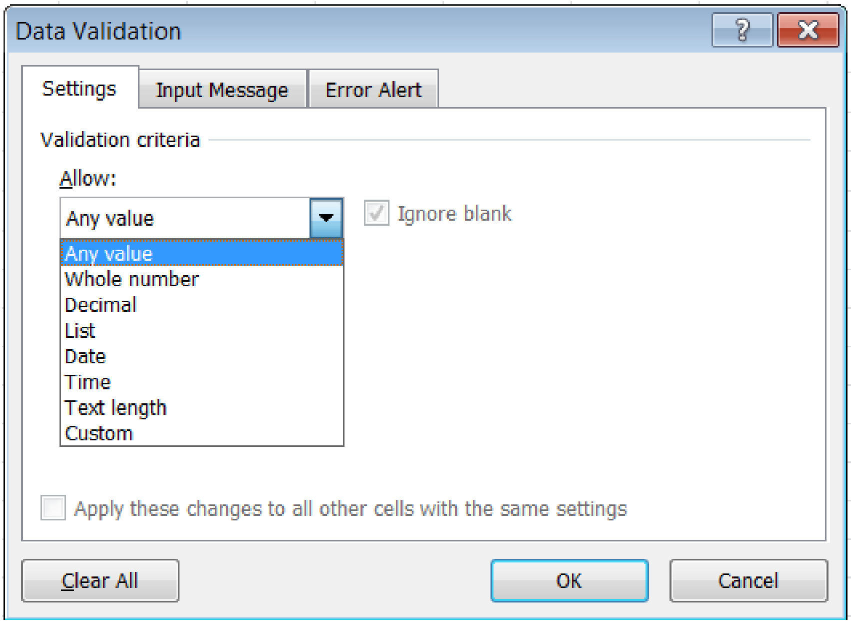
0
1

The Valid Values List is now active in the selected cell. Use Excel's copy and paste command to update the rest of the column

4.3.2 Restricting values

To restrict values based upon dates, numbers, or ranges (dates, times, numbers) follow the steps below:

- Select the cell where the drop-down list needs to be created;
- On the Data tab, in the Data Tools group, click Data Validation;
- In the Data Validation dialogue box click the Settings tab;
- In the Allow box, click the appropriate selection ;
- Fill in the requested information (start & end ranges)
- Click OK



Data Validation

Validation criteria

Allow: Any value

☑ Ignore blank

□ Apply these changes to all other cells with the same settings

The Restriction is now active in the selected cell. Use Excel's copy and paste command to update the rest of the column.

4.4 Variable naming conventions

Keep variable names short, with no spaces, or special characters. They should be described in the data dictionary. Some Examples are given below:

- Date_DX (Date of diagnosis)
- DOB (Date of birth)
- Q4tumor_pe (Q4. How was the tumor detected; choice Physical Exam)

4.5 Handling of missing or not applicable values

A good practice is to not leave a missing value blank, but instead use a value that is outside the range of possible values or a generic code to indicate a value that is missing or not applicable (e.g. -999 in a number field). That way there is a clear indication that a variable for a specific record is indeed missing or not applicable, instead of the possibility that it was missed during data entry. The same value or code can be used for multiple variables.

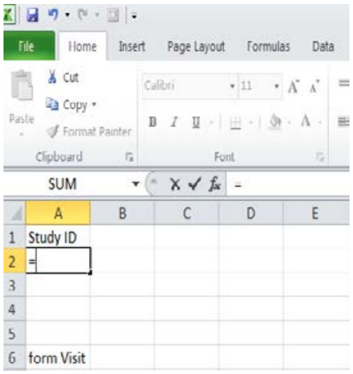
4.6 Linking spreadsheets in a workbook

Being able to type in the study id on the first Excel-based data collection form and having it auto-populate on any additional forms in the workbook is one common application for linking spreadsheets in a workbook. See the example below using form Visit and form Enrollment below:

On the cell that needs to be linked type "=" (form Visit).

- Click on the cell in the other spreadsheet in the workbook that the cell is going to be linked to (form Enrollment).
- Press enter on the keyboard (back to form Visit).
- Once the initial link is set, the Copy & Paste function can be used to link additional cells.

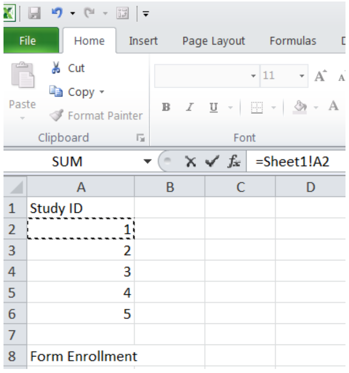
Excel will go back to the first cell, showing the value of cell from the other worksheet, as well as a formula in the formula bar.



SUM | =

	A	B	C	D	E
1	Study ID				
2	=				
3					
4					
5					
6	form Visit				

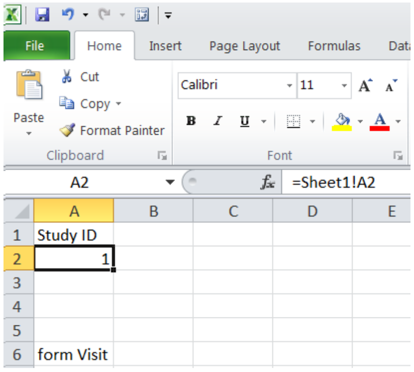
Select the cell on the visit form That needs to be linked and type "=".



SUM | =Sheet1!A2

	A	B	C	D
1	Study ID			
2	1			
3	2			
4	3			
5	4			
6	5			
7				
8	Form Enrollment			

Select the cell on the Enrollment form that should be linked to, press enter.



A2 | =Sheet1!A2

	A	B	C	D	E
1	Study ID				
2	1				
3					
4					
5					
6	form Visit				

Excel takes us back to the Visit form, notice the formula in the formula bar.

APPENDIX A: Excel Document Checklist

.

Document Section	Setting up a general template	Yes	No
3.1	Are no merged cells contained in the data area of the table (e.g. only headers and for titles)?		
3.2	Do all active worksheets in the workbook have clear and concise names?		
3.3	Are table header rows formatted to repeat on the top of the table as it goes from one page to another?		
3.4	If color is used to emphasize the importance of text, is there an alternate method?		
3.5	Has the Document been reviewed in Print Preview for a final visual check?		

Document Section	Data collection form considerations		
2.2	Has a Data Dictionary been created?		
4.1	Have the formats been checked?		
4.2	Do the variables only contain one piece of information (i.e. date/time should be two variables)?		
4.2.1	Is identifiable data (i.e. MRN #) not used for the study identifier?		
4.2.2	Has attention been paid to the type of variable (mutually or non-mutually exclusive) so that the data collection form is set up appropriately?		
4.3	Has data validation been used where appropriate?		
4.4	Are variable names short, with no spaces or special characters?		

APPENDIX B: Resources

Checklists

- Journal Review Checklist (PDF)
  Grant Writing Checklist (PDF)
  Chart Review Checklist (PDF)

Documents

- REDCap Process for Cardiology (DOC)
- RedCap Data Dictionary(DOC)

Brochures

- A Review of Statistical Analysis Software (PDF)
- Avoiding pitfalls that result in bad data (PDF)
- Database ownership (PDF)
- Don’t Monkey Around Use Survey Monkey (PDF)
- Guidelines for detecting bad data (PDF)
- How Quantitative Health Sciences can satisfy your research needs (PDF)
- Markedly Good Data Using Remark (PDF)
- Sound principles for simple statistics (PDF)
- Using REDCap to Capture Good Data (PDF)
- Working with spreadsheets (PDF)