## Missing values/not applicable values

Instead of leaving a missing value blank, use a value that is outside of the range of possible values to indicate missing data.

For numeric data, if all values are positive, " -9" could be used to indicate missing data. When possible, it is good to use the same value code for all variables. The same applies to values that are not applicable.

### Multiple pieces of information in a variable

A variable only should carry one piece of information. For example, blood pressure should be two variables, SBP and DBP.

### Spaces between rows

There should be no spaces between rows. The computer reads these as missing records.

## Brochures

- Database ownership (1 of 3).
- Avoiding pitfalls that result in bad data (2 of 3).
- Guidelines for detecting bad data (3 of 3).
- How Quantitative Health Sciences can satisfy your research needs.
- Sound principles for simple statistics.


***QHS Section***

**Pippa M. Simpson, PhD**
*Director*

**Raymond G. Hoffmann, PhD**
*Associate Director*

Shun-Hwa Li, PhD
*Senior Biostatistician*

Ke Yan, PhD
*Senior Biostatistician*

Mahua Dasgupta, MS
*Biostatistician*

Melodee Nugent, MA
*Biostatistician*

Chris Cronk, ScD
*Senior Epidemiologist*

JoAnn Gray-Murray, PHD
*Qualitative Researcher*

**Database Support**

Kathy Divine, MS
*Database Administrator*

Haydee Zimmerman, BA
*Database Analyst II*

Kim Gajewski, BA
*Database Analyst II*

Robert Thielke, PhD
*Manager IS II*

Children's Research Institute
*A member of Children's Hospital and Health System.*

MEDICAL COLLEGE OF WISCONSIN

Phone: (414) 955-7675 • Fax: (414) 955-6331
www.chw.org/qhs




Children's Research Institute
*A member of Children's Hospital and Health System.*

QHS
Quantitative
HEALTH SCIENCES

# Working with spreadsheets

Quantitative Health Sciences was established to provide help in the design and analysis of research studies.

# Common problems

***Contact the statistician BEFORE entering your data. Although the following suggestions can expedite data analysis, every project is different and may not be best served by these guidelines.***

## Data dictionary

A data dictionary describes the variables and values that variables can take. It is best to create your data dictionary first and then enter the data. 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, especially when the variable names are not adequately descriptive. For example:

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

### Dates

Dates should be formatted to be recognized by Excel as being a date. It also is recommended to enter the complete year portion of the date as a four-digit number, such as 2007.

### Extraneous text

Although helpful to humans, comments in the spreadsheet can cause problems for the computer application that is processing the data.

Annotation to define values causes problems. For example, entering 10 years, 2 months, 3 days for age cannot be analyzed. A solution is to create three columns with one for each figure: years, months and days.

## Formats

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.

**Free text fields/comment fields:**
These cannot be analyzed.

**Variable values:**

- Binary data.
- Consider using 1=Yes, 0=No. Y/N and T/F also are acceptable.

## Multiple categories

**Mutually exclusive example**: A patient only can take on one of the multiple values. 1=Low, 2=Medium, 3=High or L/M/H.

**Non-mutually exclusive 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. The binary variables could be PE, RI and LAB.

### Header rows

The first row always should contain the variable names and data values should start on the second row. Try to avoid multiple header rows.

## Inconsistencies

**Spelling differences.** Use short characters to minimize this problem. For example, spinal and spine as two different values. The symbols (+) and + are different.

**Upper/lower case.** Always use one case consistently. For example, "Cranial" and "cranial" are two different values.

**Differences in detail**. Detail may reflect reality, but is not necessary for analysis. For example, "Neck" and "Neck" may not be necessary for Radiation TX (Gy.).

## Long character values

Specifying long character values is not necessary and often can lead to inconsistencies and misspellings.

Numeric data may be preferential in many instances. For example, "1" for female and "2" for male.

Character data also may facilitate correct data entry. Use a single number or character when possible. For example, use "F" for female and "M" for male.

Short values also are acceptable, such as "ALL" for acute lymphoblastic leukemia.

## Long variable names

Ideally, variable names should be eight characters or less with no spaces, such as "DX" for diagnosis and "DATE_DX" for date of diagnosis. This does not allow for descriptive variable names, but details can be spelled out in the data dictionary.