

Research Minute

Avoiding Nightmare Spreadsheets

Robert Wood, DrPH
Sandra Burge, PhD

Issue 11



Describe your coded data in a Data Dictionary

| Data Dictionary |
|---|
| Item name (column headers— avoid special characters and spaces in the name) |
| Item Description (for labeling in the printout—don't make this more than 20 characters) |
| Category labels if applicable (i.e. 1=Male, 2=Female) |
| Missing Value codes (best to leave any missing values as blank) |

Organizing Data for QI and Research Projects: Avoiding Spreadsheet Nightmares

Have you ever wondered what goes on behind the scenes when you read journal articles about medical research? One 'invisible task' of research is recording data and getting it into a usable form for statistical analysis. One could make the case that the majority of effort in preparing for statistical analyses is constructing a functional computer data file. If a data file is properly prepared, then statistical analysis is fast and easy. However, a badly constructed data file will prevent the research aims from being completed.

Data Preparation:

Apply the following rules to data entry into spreadsheets, survey form tools like Survey Monkey or Red

Cap or statistical software packages such as SPSS, SAS, STATA or EPI Info.

- Organize information into rows and columns.
- Rows are individual cases (such as patient records or resident responses to a survey).
- Columns are individual measures (such as age, sex, immunization status). Use as many as you have items to measure.
- If more than one response per case is possible (like 2 insurance plans for one person) use multiple columns, one for each option.
- Use one column to create a unique ID for each record. It is best if this ID is encrypted in some

manner to maintain the confidentiality of the computer records.

- If the same research subject has multiple records or records in multiple files they should all have the same matching ID.
- Code the responses—use numbers and not text whenever possible.
- When using text to code a response, words must be consistently spelled and formatted—even the case must match.
- Number or category? When a measure is characterized by a continuous number (such as age or blood pressure), record the actual number, not a category. You can always make categories later.

Not This!

| Drug | Height | weight | sex | age | HTN |
|--------|------------|--------|---------|-----------|----------|
| STATIN | 5.10 | 200 | m | 3/31/52 | 1.00 |
| Zocor | 5/10 | 100k | M | 55.5 | Yes |
| SaTIN | 5 ft 10 in | 60lbs | Male | 44 yrs | YES |
| - | 120 | ??? | 1 | 10 years | 2.00 |
| NONE | 1meter | 40 pds | 0 | 44 months | No |
| None | 6.10 | 80 12 | f | - | NO |
| Sim | 530 | obese | F | 2.50 | Not sure |
| no | 5.80 | 160 | Fem | 1.90 | - |
| yes | 4.10 | 3.00 | unknown | 2.10 | maybe |

This!

| Drug | Height_in | Weight_Lbs | sex | Age_yrs | HTN |
|-------------|-----------|------------|-----|---------|-----|
| Simvastatin | 70 | 200 | 1 | 61 | 1 |
| Simvastatin | 70 | 220 | 1 | 55 | 1 |
| Simvastatin | 70 | 60 | 1 | 44 | 1 |
| | 65 | | 1 | 10 | 2 |
| None | 36 | 40 | 0 | 3 | 2 |
| None | 73 | 80 | 0 | | 2 |
| Simvastatin | 63 | 300 | 0 | 2 | 3 |
| None | 68 | 160 | 0 | 1 | 3 |
| Simvastatin | 58 | 300 | | 2 | 3 |