Preparing Data for Analysis Using Microsoft *Excel*

Alan C. Elliott, Linda S. Hynan, Joan S. Reisch, Janet P. Smith

ABSTRACT

A critical component essential to good research is the accurate and efficient collection and preparation of data for analysis. Most medical researchers have little or no training in data management, often causing not only excessive time spent cleaning data but also a risk that the data set contains collection or recording errors. The implementation of simple guidelines based on techniques used by professional data management teams will save researchers time and money and result in a data set better suited to answer research questions. Because Microsoft *Excel* is often used by researchers to collect data, specific techniques that can be implemented in *Excel* are presented.

Key Words: data collection, database management system, research design, pilot projects, informatics

During a presentation at an IBM technicians meeting in 1966, a programmer named Wilf Hey coined the phrase ''Garbage In, Garbage Out.'' Now abbreviated GIGO, this term has become a catchphrase for the too common situation in which inaccurate data entered into a computer are used to produce misleading or erroneous results.[1]

Investigative teams, ranging from a few individuals to complex organizations at universities, governments, and corporations, are all involved in the planning, execution, and analysis of research. A critical component essential to each of these research projects is collecting data and entering results into a computer in preparation for statistical analysis. Often because of a lack of funding, experience, or both, these data are entered into the computer using an ad hoc process that results in poorly coded data, incorrectly formatted values, incomplete information, and typographical errors. It is not uncommon for researchers to discover that their data sets must be extensively ''cleaned'' before they can be properly analyzed. Written for research teams who do not have the services of a professional data management team, this article provides guidance on how to develop a strategy to create a well-designed and verified data set. Following these guidelines will save researchers time and money during all phases of the research project and will result in data that can be used in a statistical software program with minimal modification.

All of the strategies in this article can (and should) be employed whether the data are entered into the computer using *SAS* (SAS Institute, Cary, NC), *SPSS* (SPSS Inc, Chicago, IL), *Access* (Microsoft Corporation, Redmond, WA), or any number of programs. To illustrate the guidelines in this article and appendix, we use the Microsoft *Excel* (Microsoft Corporation) spreadsheet program. Although *Excel* was not designed to be a research data entry tool, it is commonly used because almost every researcher already knows the basics of how to use it. This article does not address the use of *Excel* for data analysis because its limited data analysis capabilities and sometimes confusing output make it suitable only for preliminary analyses. As Jonathan Cryer put it, ''Friends don't let friends use Excel for statistics.''[2] Other articles have also discussed the problems associated with performing statistical analysis using *Excel*.[3,4] Furthermore, *Excel* is limited to spreadsheets containing less than 256 variables (columns) and 65,536 records (rows).

In well-funded studies, a professional data management team works with investigators from the planning stage through forms development, database design, and data collection and entry and in the preparation of data for analysis. The characteristics of a professionally designed data management process include a thorough description of the data variables, validation of data values as they are entered into the computer, and the use of a double-entry data process into a relational database. Such processes use specialized programs for data entry rather than *Excel*.[5,6] Programs designed for professional data entry include the *SPSS Data Entry Builder*, *Key Entry III* from Southern Computer Systems (Birmingham, AL) *SAS/AF*, and *Access*. Some of these programs require the expertise of a programmer to create data entry screens, validation code, and data verification procedures. For smaller projects in which the use of a professional data management team is not

From the Department of Clinical Sciences (A.C.E., L.S.H., J.S.R., J.P.S.), Division of Biostatistics, UT Southwestern Medical Center, Dallas, Dallas, TX.

Address correspondence to: Alan C. Elliott, Department of Clinical Sciences, Division of Biostatistics, UT Southwestern Medical Center, Dallas, Dallas, TX 75390; tel: 214-648-2712; fax: 214-648-7673; e-mail: alan.elliott@utsouthwestern.edu.

DOI 10.2310/6650.2006.05038

TABLE 1 Variable Types Collected for Research

Variable Type	*Description*
ID number or code	An identification variable that uniquely identifies a subject or entity. This could be a patient identifier, a number assigned by the computer, or some other assigned code.
Demographic variables	These variables include measures such as age, gender, and ethnicity that describe the subject population.
Outcome variable(s)	(Dependent variable) This variable is the primary outcome measure. It can be discrete (eg, dead/alive, cured/not cured) or continuous (eg, time to recurrence, blood pressure, cost incurred).
Predictor variables	(Independent variables) These variables may include treatment (grouping) variables and other measured or observed information, such as weight, height, or smoking status.
Covariate measures	Covariate measures are variables that are related to the outcome variable and may be used to adjust the mean of the response variable and account for variability in a statistical model.
Verification measures	These variables may be used to verify the reliability of the data, such as a measure of how compliant a subject is in taking medicine or multiple laboratory values to assess the reliability of laboratory data.

feasible and data entry is performed using *Excel*, the savvy investigator can still implement the ''good practice'' techniques described here.

SELECTION AND DESCRIPTION OF DATA ELEMENTS

Accurate data collection begins with planning. Before collecting any data, an investigator should define research questions and determine what measurements are needed to answer them. Typically, a research data set includes at least one outcome variable (dependent variable) and one or more predictor (independent) variables. Other demographic, covariate, or verification measures may also be recorded. It is essential that a unique key identifier be included for every observed subject (or entity) in a data set. Table 1 describes types of variables collected for data analysis. For each project, the researcher should use this information to verify that all of the variables needed to perform an analysis are included. It is disheartening to realize too late that a variable needed to complete an analysis was not collected.

An important part of a well-designed study is the documentation of each variable in a table called a data dictionary. An example data dictionary is shown in Table 2. This table can be created in a word processor or spreadsheet program and, once created, defines the characteristics of variables in both the data entry spreadsheet and statistical program.

A brief explanation of each item of the dictionary follows:

- *Variable name*. Select simple variable names using naming conventions compatible with the program that will be used for analysis. General naming conventions that work for most programs (such as *SAS* and *SPSS*) include the following:
  - Make variable names short and explanatory. Typical variable names are ID, GENDER, B_DATE, COST, GROUP, IQ_SCORE, and DEATH.

TABLE 2 Example Data Dictionary

Column	*Variable Name*	*Label (Units)*	*Format*	*Codes and Ranges*	*Missing Values*
A	SUBJECT	Subject ID number	Text (4)	1000–9999	Not allowed
B	VDATE	Date client visited clinic	Date (MM/DD/YYYY)	None	. (dot) or 11/11/1111
C	AGE	Age at visit date	Numeric (3.0)	Range 0–100	−9
D	TEMP_F	Temperature (°F)	Numeric (4.1)	None	−9
E	GENDER	Gender	Text (1)	F = female M = male	X
F	ARRIVE	Mode of arrival	String (4)	Car Bus Walk	MISS
G	ANTIBIO	Was antibiotic prescribed?	Numeric (1.0)	1 = yes 0 = no	9

MISS = missing.

- Because some programs (such as older versions of *SPSS*) limit names to eight characters, it is best to comply with this restriction.
- Begin variable names with a letter (A–Z). Some programs also allow variable names to begin with an underscore (_). Variable names may include numbers (not as a first character) but not blanks or other special characters (an underscore is valid). For example, AGE_2004 or AGE2004 is valid but not 2004AGE or AGE 2004 (with a blank between 2004 and AGE). Each name in the data set must be unique.
- Name variables in a sequence when appropriate. For example, responses on a questionnaire might be named Q1, Q2, Q3, etc... The sequence of variables may then be referred to using a shortcut such as Q1-Q47 in statistical programs such as *SAS*.
- Most programs also allow the inclusion of a descriptive label for each variable.

- *Label*. Include a brief description of the variable that can be used as the variable label in the analysis program. For example, the label for AIS_SCORE might be ''AIS based on the ICD-9-CM scoring.''
- *Format*. Specify the format to be used to enter data. For example, use a single-digit integer to indicate the presence or absence of a condition (0 or 1), a five-digit number including one decimal point to indicate weight, or a date in the format MM/DD/YYYY. Measurement data values should include a sufficient number of digits but not too many. For example, recording a person's weight to the second decimal place would be unnecessary in most cases, even if the weighing device reported the data to that number of decimal places. For measurement variables, specify units such as pounds, inches, or liters. The most commonly used data types are as follows:
  - Numeric variables are those for which mathematical calculations make sense, such as age, salary, or weight. These variables are measured numerically. A binary code (0, 1) representing the presence or absence of a condition is usually coded as numeric (even though the number is an identifier and has no real numeric meaning.)
  - Text variables (also called string or character variables) are codes, descriptions, or nonmathematical numbers. For example, gender recorded as male and female or M and F is a text variable. Specifying an ID number such as ''23432'' as a text variable prevents it from being used to calculate a meaningless statistic, such as an average. If text variables are used in a data set, it is best to avoid using entries that contain a large number of characters. Data values such as ''pulmonary embolus'' and ''loss of operative reduction/fixation'' are lengthy to type into the spreadsheet and invite error. Use a code such as PE or LORF instead. If long descriptions must be used, it is best to use the list box selection criteria described later in this article to ensure consistency. Depending on the statistical program, there may be a downside to using text categorical variables. Some programs, *SPSS* in particular, will not allow the use of text variables as grouping variables in some analyses. Instead, *SPSS* requires that variables be classified as numbers (eg, 1, 2, and 3 for red, white, and blue).
  - Date variables can be used to represent dates and times of the day. Variables classified as dates warrant special attention. Make sure dates are defined with a four-digit year format to prevent any of the old ''Y2K'' problems from occurring. For example, if dates are entered using two-digit years in *Excel* and a date calculation is performed using its default date settings, dates ranging from 01/01/00 to 12/31/29 are considered in the years 2000 to 2029, whereas the date 01/01/30 (and after) is considered as being in the year 1930 (and after). Thus, if age is calculated on January 1, 2006, for a subject born on July 10, 1925, and entered as 07/10/25, *Excel* will calculate the age as −19. When dates using two-digit years are imported into a statistics program, a similar error may result. It is best to store dates as a single properly formatted date variable rather than storing month, day, and year as three separate variables. However, be aware that there are different ways of formatting a date. A typical US date format is MM/DD/YYYY, but some countries (and the military) use DD/MM/YYYY. If data are collected or recorded in another country, make sure the data entry procedure takes this into account.

- *Codes and ranges*. Define the range of values for each variable. For example, AGE might be limited to values from 0 to 100. Categorical variables should be limited to a specific list of possible values. For example, 0 = no and 1 = yes or AA = African American, H = Hispanic, C = Caucasian, and O = other. In most statistical programs, formats can be defined for coded variables, so output will display the descriptions (male and female) rather than a cryptic code (0 and 1). If a data set contains numeric variables that are recorded with entries such as ''>50'' or ''40-50,'' consider recoding this variable into a categorical variable. For example, this variable could be recoded using three coded values: 1 = ''0-50,'' 2 = ''51-100,'' and 3 = ''greater than 100.'' It is not possible to calculate averages and related statistics on data that consist of a mixture of numbers and ranges.
- *Missing values*. A missing value is a data element for which there is no available value. Missing data points can result from lost, never collected, or unknown information. There are several methods of handling missing values. If an *Excel* cell is left blank for a missing value and subsequently imported into *SAS* or *SPSS*, the blank value will be imported properly as a missing value. However, it is best to define an explicit missing value code as a confirmation that the data value has been

accounted for and has not been overlooked. For numeric variables, this code is typically an impossible value, such as −9 for age. Sometimes it is necessary to define more than one missing value code for a variable such as −9 for ''not available'' and −8 for ''not done.'' Missing values for a text variable might be defined, such as MISS or NA. For date values, use a blank to indicate a missing value or a dot (.) as a missing value code. Another option would be to use a date impossible for your study, such as 11/11/1111. If missing value codes are used in a data set, they must subsequently be defined within the statistics program in which the data will be analyzed.

DATA COLLECTION STRATEGIES TO ENSURE BETTER DATA ACCURACY

Data set design and collection strategies that lead to increased accuracy, reliability, and analyzability of a data set include the following:

- *Use open-ended questions with caution*. Questions such as ''List the medicines you are taking'' or ''What magazines do you read?'' are open-ended questions. Free-form answers to these types of questions are difficult to analyze using statistical procedures (although such information may be useful to analyze in a more subjective way). If your desire is to collect data useful for statistical analysis, construct questions that require subjects to select answers from a checklist (which should include ''unknown'' or ''other'' as a category) or be prepared to classify answers to open-ended questions into categories for analysis.
- *Avoid unnecessary data collection*. Collecting an excessive number of data elements (not pertinent to the research question) can lead to coding or ''fatigue'' errors. Collect only the data needed for the study.
- *Perform a pilot study*. Gather a small amount of data and perform a preliminary analysis before collecting the full research data. Many design and data collection flaws can be found in this way. Information gleaned from a pilot study will often help in planning a more effective larger study. It is also helpful to have a knowledgeable and critical colleague look over the data collection forms or questionnaires before they are used in an actual study.
- *Develop a data audit procedure*. Even though *Excel* is a convenient tool for data entry, it was not designed for data auditing. Therefore, the burden of adapting *Excel* (or any similar strategy) to the data entry process lies with the researcher. Good practice (and some governmental regulations) require that an audit trail be maintained for data changes in certain types of research (such as clinical trials). Professional data entry programs will automatically keep track of any records whose values are changed (including date, time, data entry person, etc), but there is no provision to do so in *Excel*. This means that when data are altered in any way, there must be a procedure in place to keep track of that change, such as with a written change form.

Addressing the issues related to the design of the study and the way in which data will be collected and recorded is an important step in increasing the accuracy of the data set. For practical guidelines for entering and verifying your data in *Excel*, see the Appendix to this article.

IMPORTING DATA

Once you have entered your data in *Excel* (or some other program), you must import that data into your statistical analysis program of choice before you can analyze it. Most statistics programs, such as *SAS* and *SPSS*, can import data files directly from *Excel*. If you have followed the guidelines in this article, the import will be straightforward, with few or no problems. However, you should always perform data checks once your data are imported to verify that a complete and accurate import occurred. In addition, you may want to add variable labels and define categorical codes once data have been imported.

CONCLUSION

If the data entered into your statistical program have errors, many analyses you perform will be wrong. To increase your chance of entering your data correctly into the computer, you must develop a data management strategy. This article described guidelines for creating such a strategy and provided information on how to use Microsoft *Excel* as your data entry tool. The guidelines described here, if followed, will help you create a cleaner, more accurate, and more appropriate data set that is well designed to answer research questions. The Appendix illustrates how these techniques can be implemented in *Excel*.

REFERENCES

1. English dictionary. Available at: http://www.english-dictionary.us/meaning/wilf_hey.asp (accessed July 18, 2005).
2. Cryer J. Problems with using Microsoft Excel for statistics. In: Proceedings of the 2001 joint statistical meetings [CD-ROM]. Alexandria, VA: American Statistical Association; 2002.
3. Knüsel L. On the reliability of Microsoft Excel XP for statistical purposes. Comput Stat Data Anal 2002;39(1):109–10.
4. McCullough BD, Wilson B. On the accuracy of statistical procedures in Microsoft Excel 97. Comput Stat Data Anal 1999; 31:27–37.
5. McFadden E. Management of data in clinical trials. New York: Wiley-Interscience; 2002.
6. Prokscha S. Practical guide to clinical data management. Denver (CO): Interpharm Press.

APPENDIX

Implementing Data Management Techniques in *Excel*

This appendix illustrates the data management techniques described in the article. Follow these examples in Microsoft *Excel* to create a data management strategy that will improve the accuracy and usefulness of your data.

Design Your Spreadsheet Using the Data Dictionary

Establishing the data dictionary is an essential first step in preparing data for entry into an *Excel* spreadsheet. With the dictionary in hand, following this list of guidelines will help design the spreadsheet for data entry:

- *Place variables names in row 1*. The first row of the data spreadsheet should contain only variable names. For example, Figure 1 shows a spreadsheet containing the seven variables specified in the previous data dictionary, one per column. These variable names are found in the data dictionary shown in Table 2. Case does not matter for variable names. ''Subject'' works equally as well as ''SUBJECT.''
- *Format columns to match the variable type*. To help prevent inaccurate values from being entered into the data spreadsheet, format the column cells to match the prescribed data values for that column. For example, column ''B'' contains a date variable that should be in the form MM/DD/YYYY. To format the VDATE column in *Excel*,

  1. Highlight the cells that will contain date values by clicking the column header, B in this example.
  2. From the *Excel* menu, select Format/Cells/Date.
  3. Select the MM/DD/YYYY format (which appears on the list of formats in the form of the current date, such as 07/10/2005). This will cause data in these cells to appear in the specified date format.

In a similar way, format the SUBJECT, GENDER, and ARRIVE columns as text. The TEMP_F column should be defined as a number with a single decimal place, and the ANTIBIO column should be a number defined with no decimal place because the data entered will be a 0, 1, or 9. The AGE column should be three digits with no decimal.

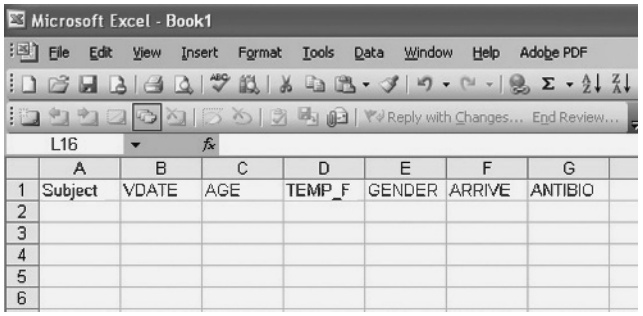

FIGURE 1 Variable names for a data set.

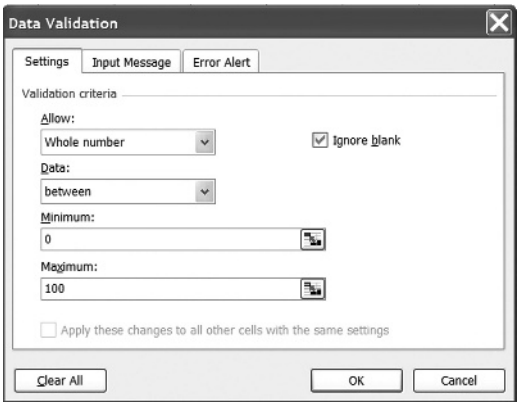

FIGURE 2 Creating data validation criteria.

Never format columns using the currency ($) or comma formats because that may cause problem when the data are imported into a statistics program.

- *Specify a range of allowed values*. Using the criteria in the data dictionary, specify a range of possible values allowed in a particular range of cells. The following steps can be used to specify that the format for the AGE variable in *Excel* column ''C'' can only contain values between 0 and 100 but will also allow you to override the check to enter a missing value code of −99:

  1. Select the range of cells to be marked for validation by highlighting a range of cells.
  2. Select the menu option Data/Validation. . .
  3. On the Settings tab option, select ''Whole Number'' from the ''Allow:'' criteria option.
  4. On the ''Data:'' option, leave it as ''between'' and enter 0 as the minimum and 100 as the maximum (Figure 2).
  5. Click on the Input Message tab. In the Title textbox, enter ''Age verification,'' and in the Input Message text box enter ''Age must be between 0 and 100 or −9 (override) for missing.''
  6. Click on the Error Alert tab. Select the yellow Warning icon in the Style pull-down box. In the Title text box, enter ''Age out of range.'' In the Error message box, enter ''Enter an age between 0 and 100 or −9 (override) for missing.'' Choosing the Warning icon rather than the Stop icon allows the data entry person to override the preset limits. This is recommended in this case because it is possible for a subject to be over the age of 100 and because you designated −9 as a missing value code.
  7. Click OK.

Once defined in *Excel*, the verification criteria will check values as they are entered to make sure the entry does not violate the specification. When a cell in the age column is selected, a yellow box appears with the input message ''Age

must be between 0 and 100 or −9 for missing.'' If a value outside the limits is entered into the cell, the dialog box shown in Figure 3 appears.

As mentioned earlier, the data entry person can override the warning by clicking ''Yes'' and then enter a value of −9. If the ''Stop'' option had been selected rather than ''Warning,'' *Excel* would not allow the entry person to override the data range. A third option, ''Information,'' displays a message when a value is entered out of range, but it does not prevent the entry of a number outside the specified range.

This example illustrates how to limit the entry of whole number values, but *Excel* also allows the specification of limits for decimals, dates, lists, and clock times. When one of these ''Allow'' criteria is selected in the ''Data Validation'' dialog box shown in Figure 2, the other entry options change to match the options allowed for that value type.

- *Limit data values to a list.* This list could be a list of US state abbreviations, names of months, days of the week, gender, hospital names, diagnoses, and so on. To limit an entry to M and F for a gender variable (and X for missing), for example, use these steps:

  1. Create the list of items in the same spreadsheet as the data. For example, in cell L2, place the value ''M''; in cell L3, place the value ''F''; and in cell L4, enter ''X'' (no quotation marks).
  2. Select the range of cells to be marked for validation by highlighting a range of cells.
  3. Select the menu option Data/Validation. . .
  4. On the Settings tab option, select List from the Allow: criteria option.
  5. For the Source, select the range of cells containing the admissible values. In this case, enter ''=$L$2:$L$4'' (no quotation marks). The dollar signs in the specification force the reference to be absolute. This range must be in the active spreadsheet. It is best to allow several blank columns between the actual data values and this list because the list can interfere with importing the data later.
  6. Create an input message by clicking on the Input Message tab in the Data Validation dialog box. Check the box titled ''Show input message when cell is selected.'' In the Title textbox, enter ''Select Gender,'' and in the Input Message textbox, enter ''Select M=Male, F=Female or X=Missing.''
  7. Click on the Error Alert tab. Select Warning from the Style pull-down menu. For the Title, enter ''Gender,'' and for the Error message, put ''Only uppercase M and F allowed in this field, or X for missing.''
  8. Click OK.

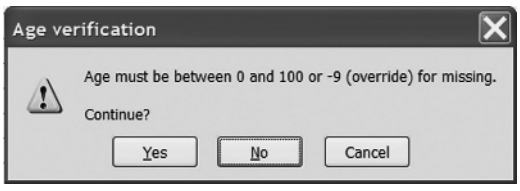

FIGURE 3 Warning that an invalid number has been entered.

Once these verification criteria are set up, when a cell within the specified range is selected, the message ''Select Gender'' and ''Select M=Male, F=Female or X=Missing'' appears in a yellow information box. If a value besides an M, F, or X is entered, the message ''Only uppercase M and F allowed in this field or X for missing'' appears in a warning dialog box. Clicking on the pull-down menu indicator (down arrow) in the box displays the list of defined values (M, F, or X).

Unfortunately, *Excel* will not match case, so it is possible for the entry person to enter a lowercase ''m'' instead of an ''M.'' Keep this in mind when you import these data into your statistics program.

These simple verification checks take only a few minutes to set up in *Excel*. If there is more than a small data set to enter, or if multiple people will be entering data, these validations will prevent the entry of obviously incorrect data.

- *Make each row of data represent a single subject (usually).* In most cases, data for a single subject or observation should be on a single row in the spreadsheet. A few analyses in *SPSS* and *SAS*, typically repeated measures models, expect data for a single subject in multiple rows. If you use multiple rows per subject, additional variable(s), such as visit number, date, or time, must be included so that each row is uniquely identified. Always use the single subject–single row option, unless the multirow format is required. If data are entered on a single row, the data can later be transformed into the multicolumn format within the statistics program to a multirow format if needed (or vice versa.)

Data Entry Guidelines

Along with the techniques described above, here are other suggestions that can ensure a cleaner data set:

- *Freeze column headings so they will not scroll off the screen.* When data are entered in *Excel*, it is easy for the column names to scroll off the screen. This makes it more likely to enter the data in the wrong column. To prevent this, freeze the variable names to always remain at the top of the screen. To freeze the variable names, click on A2 in the *Excel* spreadsheet (variables names are in column 1) and select Windows/Freeze Panes. In a similar way, if ID is in the first column of a data entry spreadsheet, freeze both variable names and the ID column of the spreadsheet by following these steps:

  1. Click on B2 in the data entry spreadsheet.
  2. Select Windows/Freeze Panes.

3. The variable names and ID column remain on the screen even when scrolled.
4. To Unfreeze the panes, click Windows/Unfreeze Panes.

- *Enter string variables in a consistent case*. String (text/categorical) variables should always be entered in the same case. When entering a letter-coded gender variable, consistently use either M and F or m and f. If cases are mixed ''M and m,'' the statistics program may see these letters as two different data values. When a comparison by gender is performed, the program finds four gender categories (F, f, M, and m). Forcing the data to match a list, as described above, is one way to prevent the mismatched case problem.
- *Do not leave any blank rows in the spreadsheet*. Blank rows are sometimes imported incorrectly into the statistics program (depending on the program) and may complicate an analysis.
- *Do not include unessential text or fancy formatting in the spreadsheet*. Extra text, colors, unusual fonts, separator lines, and other formatting options that are not meant to be imported into the statistics program may cause problems. Keep the data entry spreadsheet straightforward and simple.
- *Get rid of formulas*. If the data spreadsheet in *Excel* contains formulas, there may be unexpected errors when the data are imported into a statistics program. A technique to get rid of all formulas is to copy the entire data spreadsheet (leaving out any cells containing lists used in list entries), go to a new blank sheet, and select Edit, Paste Special. From the Paste Special dialog box, select the ''Values'' option and OK. This will paste the data into the new spreadsheet with all formulas removed. Save this new spreadsheet (under a new name) and use it to import the data into the statistics program.
- *Sort data with caution*. It is often helpful to sort a data set in *Excel* to put it into some order, such as patient number. However, be cautious when sorting data in *Excel* because it is easy to sort a single column while leaving the other columns intact, thus ruining the integrity of the data. Data sets should be saved before performing any sorting. To sort correctly, highlight all of the columns containing data and click Data/Sort. Follow the prompts to select which columns to use as sorting variables.

After designing a data entry spreadsheet using the guidelines above, the data entry should not only be more accurate but will also result in a data set that can be imported seamlessly into a statistical program, avoiding much of the time-consuming data manipulation and cleaning problems that must take place before data can be analyzed.

	A	B	C	D	E	F	G	H
1	Subject	VISITDATE	AGE	TEMP_F	GENDER	ARRIVE	ANTIBIO	
2	001	7/10/2004	23	101.5	M	BUS	1	
3	002	8/22/2005	43	98.6	M	CAR	0	
4	003	12/14/2004	23	98.2	F	CAR	0	
5	004	2/13/2005	33	100.0	F	CAR	1	
6								

FIGURE 4 First spreadsheet to compare (Sheet1).

	A	B	C	D	E	F	G	H
1	Subject	VISITDATE	AGE	TEMP_F	GENDER	ARRIVE	ANTIBIO	
2	001	7/10/2004	23	101.5	M	BUS	1	
3	002	8/22/2005	34	98.6	M	CAR	1	
4	003	12/14/2004	23	89.2	F	CAR	0	
5	004	2/12/2005	33	100.0	F	CAR	1	
6								

FIGURE 5 Second spreadsheet to compare (Sheet2).

Verify Data Using Double Data Entry in *Excel*

The gold standard for professional data entry is to enter data not once but twice. The two data sets are then compared, differences are examined, and corrections are made. To use this double data entry method, create two identical blank data entry spreadsheets. The data should then be entered into the spreadsheets by two different people. If it is impossible to use two different people, at least enter the data at two different sessions. Once the data are entered, compare the two spreadsheets for differences in *Excel* using the following technique. Figure 4 shows the first spreadsheet to compare (Sheet1), and Figure 5 shows the second spreadsheet (Sheet2).

If the two spreadsheets containing the entered data are not in the same worksheet file, copy the second spreadsheet and paste it into Sheet2 of the original worksheet. Note that these spreadsheets must have the data in the same order and data in identical cells. To compare these two spreadsheets, follow these steps:

1. In the Sheet1 spreadsheet, select Insert/Worksheet to insert a third worksheet (Sheet3). Copy the labels (row 1) from the Sheet1 worksheet to the Sheet3 (Difference) worksheet.
2. In Sheet3, place the cursor in cell A2 and enter the following *Excel* formula:

$$= \mathrm{IF}(\mathrm{EXACT}(\mathrm{Sheet1!A2}, \mathrm{Sheet2!A2}), 0, 1)$$

3. Copy this formula to all cells from A2 to G5 (the range of cells to compare). One method of copying this formula in *Excel* is to place the cursor in cell A2 and press CTRL-C (Copy). Then highlight the cells from A2 to G5 and press CTRL-V (Paste). This copies the formula to all of the specified cells. The Difference spreadsheet (Sheet3) looks like the one illustrated in Figure 6.
4. Notice the cells in the Difference spreadsheet. Cells containing a 1 indicate that the values of the two spreadsheets in that cell do not match.
5. When doing this comparison on the data set, examine the cells that are different (marked as 1) and make

	A	B	C	D	F	F	G	H
1	Subject	VDATE	AGE	TEMP_F	GENDER	ARRIVE	ANTIBIO	
2	0	0	0	0	0	0	0	
3	0	0	1	0	0	0	1	
4	0	0	0	1	0	0	0	
5	0	1	0	0	0	0	0	
6								

FIGURE 6 Sheet3 (Difference) spreadsheet.

	A	B	C	D	E	F	G	H
1	Subject	VDATE	AGE	TEMP_F	GENDER	ARRIVE	ANTIBIO	
2	0	0	0	0	0	0	0	
3	0	0	43/34	0	0	0	0/1	
4	0	0	0	98.2/89.2	0	0	0	
5	0	38396/38395	0	0	0	0	0	
6								

FIGURE 7 Sheet3 (Difference) displaying actual differences.

corrections. Once all of the corrections have been made, the cells in the Difference spreadsheet should all be 0 (zero).

To make the difference more informative, use the more complicated *Excel* formula below (in a single line):

$$= \text{IF}(\text{EXACT}(\text{SHEET1!A2}, \text{SHEET2!A2}), 0, \text{SHEET1!A2\&''/''\&SHEET2!A2})$$

This formula produces the spreadsheet shown in Figure 7.

The Figure 7 version of the differences shows the actual data values from the two sheets displayed so that the differences are more readily visible. For example, the digits for AGE in cell C3 are reversed on the two sheets (43 versus 34). Notice in the date comparison in cell B5 that date codes (38396/38395) are displayed rather than actual dates. Because these numbers are one digit apart, it means that the dates on Spread1 and Spread2 are 1 day apart. The original spreadsheet contains the date as February 12, 2005, and the other spreadsheet contains it as February 13, 2005.

Once you have verified that the two spreadsheets are identical, you are ready to import your data in a statistics program. If you have followed the guidelines in this article, your data set should accurately reflect the data that were collected.