

## UPMC HCC Biostatistics Data Submission Requirements

### Summary

Before work can begin on any statistical analyses, we will need:

- A complete validated data set in an electronic format.
- Written documentation of the variables involved, including the units of measurement and coding.
- Confirmation that IRB approval has been obtained for the study or project under discussion.

In general, we require a *spreadsheet*, which contains the data, and a corresponding *data dictionary*, which explains the meaning of the variables to be analyzed. The first row of the spreadsheet should contain the **variable names**, one per column, with the remaining rows containing the **values** of the variables (i.e., the data) for each subject in your study.

With respect to the data values:

- Each cell must contain a single numerical or character value, and nothing else.
- No cells should be left empty.
- Missing data should be entered as a period (.).

An example of an acceptable spreadsheet is shown below:

### Spreadsheet

|   | A       | B   | C       | D          | E    | F         | G    |
|---|---------|-----|---------|------------|------|-----------|------|
| 1 | patient | sex | priortx | date1      | sbp1 | date2     | sbp2 |
| 2 | 1       | 1   | 0       | 6/11/2000  | 120  | 8/3/2000  | 125  |
| 3 | 2       | 1   | 2       | 11/12/2001 | 130  | .         | .    |
| 4 | 3       | 2   | 3       | 2/2/2001   | 999  | 3/11/2001 | 130  |

We typically also need a *data dictionary*, which explains the meaning of the variables and the coding that is used for data recorded in the spreadsheet:

### Data dictionary

|    | A             | B             | C                                  | D                                       |
|----|---------------|---------------|------------------------------------|---|
| 1  | Variable Name | Variable Type | Variable Definition (and Units)    | Category Codes (values and definitions) |
| 2  | patient       | Categorical   | patient identifier                 | unique integer for each patient         |
| 3  | sex           | Categorical   | sex of patient                     | 1=Male                                  |
| 4  |               |               |                                    | 2=Female                                |
| 5  | priortx       | Categorical   | prior treatment                    | 0=none                                  |
| 6  |               |               |                                    | 1=radiation                             |
| 7  |               |               |                                    | 2=chemotherapy                          |
| 8  |               |               |                                    | 3=both                                  |
| 9  | sbp1, sbp2    | Numeric       | systolic blood pressure (mm Hg)    | 999= sbp >200                           |
| 10 | date1, date2  | Date          | Dates on which sbp1, sbp2 recorded | Not Applicable                          |

An example of an unacceptable spreadsheet is shown below:

|   | A           | B         | C        | D         | E              | F           | G |
|---|-------------|-----------|----------|-----------|----------------|-------------|---|
| 1 | patient     | Sex / age | Race     | Height    | Blood pressure | Weight      |   |
| 2 | Jones, Mary | F / 43    | Black    | 70 inches | N/D            | 160         |   |
| 3 | Smith, joe  | M / 24    | Oriental | 65 inches |                | broke scale |   |

This spreadsheet may be useful for record keeping in the lab or clinic, but it is unacceptable for statistical analysis because:

1. *patients* are identified by name
2. values for two variables (*sex* & *age*) are entered in the same cell: F / 43 and M / 24
3. units (inches, in this case) are included with values of *height*
4. both N/D and a blank cell are used to indicate missing data for *blood pressure*
5. a comment (“broke scale”) is entered in place of a missing value (.) for *weight*

The remainder of this document provides additional details on how to format your data; however, if your data look like the first example shown above, your format is probably ok.

## **Introduction**

The format typically required by the UPMC HCC Biostatistics Facility for analysis of data sets submitted to us is described below. Our intent is twofold: first, to minimize delays -- we do not have the personnel to provide data management, and will request that you re-format data submitted in an intractable format. Second, we want to encourage good data management practice. We suggest that you consult with us on data formats before you begin data collection; we will be happy to assist with the design of an appropriate data structure. Clinical trial data, with the exception of data from special correlative studies, should be entered into the appropriate database: the CRS database (CTMA) or the Theradex database (ACES).

We also urge that there be careful planning of any analyses that you want us to perform. We prefer to be involved in planning analyses and data collection before a major study is initiated (i.e., at the design stage), and believe that our participation will often lead not only to a better end result, but also results obtained more expeditiously.

With respect to statistical analyses, it is possible that results of the initial analyses may lead to the need for further analyses, and we are willing to accommodate this. However, we do not want a series of requests for a variety of post-hoc alternative analyses of the same data, especially those involving patient subsets that seek to salvage an ill-supported but fervently held hypothesis; in this instance, the additional requests will be placed in our “low priority” queue.

We urge that data be submitted on Excel worksheets because Excel facilitates data entry, editing and display; our statistical packages can read Excel worksheets directly, if our format specifications are followed; and you probably have access to Excel.

In addition to the data, we require a **data dictionary**, which explains the meaning of the variables and the coding that is used.

Although we perform checks for logic and consistency of data, please keep in mind that we expect that data submitted to us for analysis are accurate, appropriately formatted, and complete.

With respect to accuracy and completeness, if the data have errors which are later found and corrected, or the information on study subjects is incomplete, with the pertinent records being updated at a later time, then statistical analyses will typically have to be re-done. Re-doing work because of poor planning is not among our priorities.

The **Additional Information** section at the end of this document should be carefully reviewed:

- If you cannot supply data on Excel worksheets, or if your data format does not conform to the requirements summarized above and shown in detail below.
- If your data are from a retrospective chart review or similar study.

### **Data Format**

An Excel dataset is a rectangular array of **cells** arranged in **rows** and **columns** (see the spreadsheet on page 1).

**Row #1 (variable names):** The first row must contain the *variable names*.

**Rows #  $\geq 2$  (subjects):** Each row must contain data values for a single subject; an entry in the corresponding column must be made for each variable appearing in row #1.

In some instances, a variable may be repeatedly measured for each subject. If the number of measurements varies from subject to subject, it is best to use **example 2** (see below) as a template. If the number of measurements does not vary, you can use either **example 1** or **example 2** as your data format.

**Columns (variables):** Each column must contain only the values of the variable listed in row #1 (i.e., the column heading).

One column, typically the first, must contain the values of a “*subject identifier*” (e.g., an ID number that distinguishes the subjects in the analysis).

The dataset may be augmented with any number of columns containing descriptive text, but only if the text is not required for statistical analysis.

**Cells:** each cell for rows #  $\geq 2$  must contain the value of a single variable; cells should not be left empty.

**Formats for specific kinds of data follow:**

1. **variable names:** we prefer a single word or string of characters without spaces, but can accommodate almost anything. There can be no duplicate names
2. **subject identifiers:** enter a subject identifier as a number or as a character string.

If the data are for patients who were enrolled in a UPMC HCC clinical trial, the subject identifier should be the CTMA patient identifier in the clinical trial involved. This will permit us to link to the CRS database (CTMA). We do not want patient names or initials.

3. **data values:** enter numbers as integers (1,2,3,...) or decimals (3.1, 5.2, 8.6,...).
  - If a data value is known only to be greater than some bound (e.g., >5, >10, ...), or less than some bound (e.g., <5, <10, ...), enter a numeric code *that could not possibly be mistaken for data*. For example, if the values of a variable are all greater than zero, a value >5 could be indicated by -5.

There should be a different numeric code for each bound. Include a description of the codes in the data dictionary.

4. **categories:** for a categorical variable, such as *sex* or *race*, enter a value as an integer (0,1,...), or as a short, descriptive character string without spaces.

A single character (e.g., M, F for *sex*) is preferred when unambiguous, in part because typos and other data entry errors are minimized.

When using characters, always use either upper case or lower case; do not mix cases.

5. **dates:** enter a date as mm/dd/yyyy (e.g., 11/22/2001, 1/5/2002).
6. **missing data:** if data for any cells are missing, enter a period (.) in each such cell.

If data can be missing for several reasons, and it is important for the analysis to know why the data are missing, enter a unique numeric code for each reason.

A numeric code may also be used to indicate when data are missing because a variable is not applicable to some subjects: e.g., age at first pregnancy for men, and age at first pregnancy for women who were never pregnant.

*Use codes that could not possibly be mistaken for data. .*

7. **empty cells:** there should be none.

## **Data Dictionary**

We need a “data dictionary” to describe:

1. the meaning of the variables, and the units in which they are measured.
2. the codes used for:
  - numbers, when known only to be greater than or less than some value (e.g.  $>10$ ,  $<1$ ).
  - categorical variables (e.g., *young*, *old*)
  - missing data, if not represented by a period (.)

We suggest that, if you use Excel, the data dictionary be included with the data in a separate worksheet in the same workbook: e.g., see the data dictionary on page 1. Any additional information about the variables that you believe might be useful to us can be included in the data dictionary. Examples of data dictionaries are provided below; however, we do not require that you use the specific format shown.

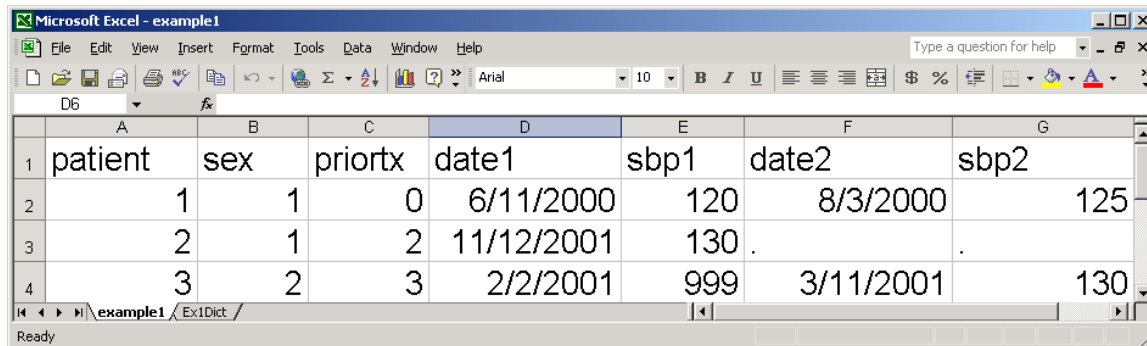
## Examples of Spreadsheets & Data Dictionaries

### Example 1: the “short” format for a spreadsheet

In the example below, systolic blood pressure (sbp) was measured 2 times (sbp1, sbp2) for patients 1 and 3; the second sbp measurement was not made for patient 2, and must be entered as missing. The first sbp measurement for patient 3 was greater than 200, and is therefore represented by a code, “999”.

Numeric codes are used for the categorical variables *sex* and prior treatment (*prior tx*).

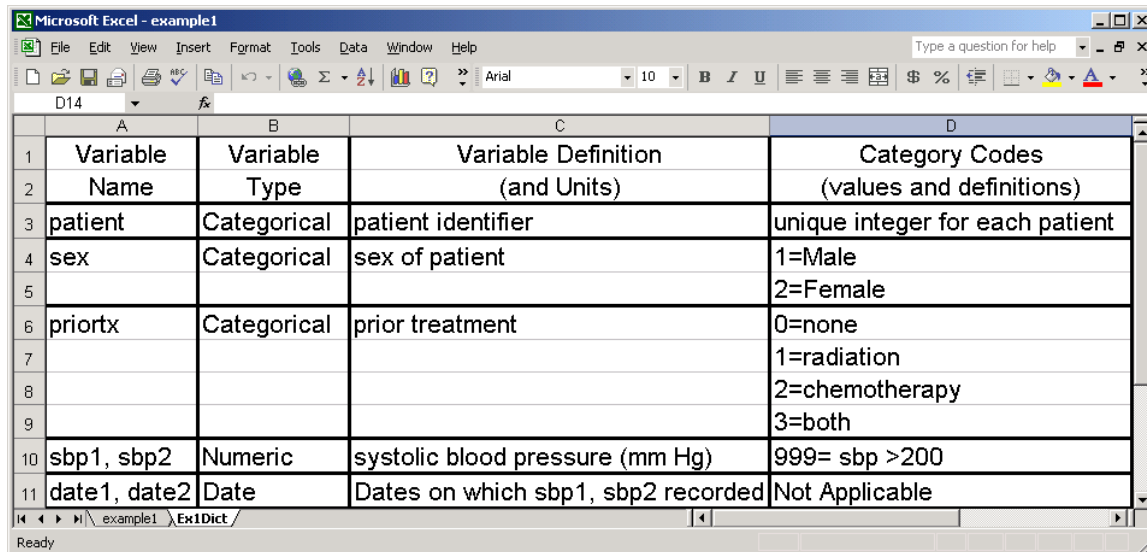
### Spreadsheet



The screenshot shows a Microsoft Excel spreadsheet titled "example1". The data is organized in columns: patient, sex, prior tx, date1, sbp1, date2, and sbp2. The rows represent individual patients.

|   | A       | B   | C        | D          | E    | F         | G    |
|---|---------|-----|----------|------------|------|-----------|------|
|   | patient | sex | prior tx | date1      | sbp1 | date2     | sbp2 |
| 1 | 1       | 1   | 0        | 6/11/2000  | 120  | 8/3/2000  | 125  |
| 2 | 2       | 1   | 2        | 11/12/2001 | 130  | .         | .    |
| 3 | 3       | 2   | 3        | 2/2/2001   | 999  | 3/11/2001 | 130  |

### Data Dictionary



The screenshot shows a Microsoft Excel spreadsheet titled "example1" containing a data dictionary. It defines the variables used in the spreadsheet, their types, and their category codes.

|   | A             | B             | C                                  | D   |
|---|---------------|---------------|------------------------------------|---|
|   | Variable Name | Variable Type | Variable Definition (and Units)    | Category Codes (values and definitions)           |
| 1 | patient       | Categorical   | patient identifier                 | unique integer for each patient                   |
| 2 | sex           | Categorical   | sex of patient                     | 1=Male<br>2=Female                                |
| 3 | prior tx      | Categorical   | prior treatment                    | 0=none<br>1=radiation<br>2=chemotherapy<br>3=both |
| 4 | sbp1, sbp2    | Numeric       | systolic blood pressure (mm Hg)    | 999= sbp >200                                     |
| 5 | date1, date2  | Date          | Dates on which sbp1, sbp2 recorded | Not Applicable                                    |

### Example 2: the “long” format for a spreadsheet

The same variables are recorded as in example 1, but in this instance there are no missing data. Character codes are used for categorical variables. (The prior treatment variable in example 1 has been split into two variables, one for prior radiation and one for prior chemotherapy; this is generally preferable to combining the two, as in example 1.)

#### Spreadsheet

|   | A       | B   | C        | D          | E          | F   |
|---|---------|-----|----------|------------|------------|-----|
| 1 | patient | sex | priorrad | priorchemo | date       | sbp |
| 2 | 1       | m   | n        | n          | 6/11/2000  | 120 |
| 3 | 1       | m   | n        | n          | 8/3/2000   | 125 |
| 4 | 2       | m   | n        | y          | 11/12/2001 | 130 |
| 5 | 3       | f   | y        | y          | 2/2/2001   | 999 |
| 6 | 3       | f   | y        | y          | 3/11/2001  | 130 |

#### Data Dictionary

|   | A             | B             | C                               | D                                       |
|---|---------------|---------------|---------------------------------|---|
|   | Variable Name | Variable Type | Variable Definition (and Units) | Category Codes (values and definitions) |
| 1 | patient       | Categorical   | patient identifier              | unique integer for each patient         |
| 2 | sex           | Categorical   | sex of patient                  | m = Male<br>f = Female                  |
| 3 | priorrad      | Categorical   | prior radiation                 | y = yes<br>n = no                       |
| 4 | priorchemo    | Categorical   | prior chemotherapy              | y = yes<br>n = no                       |
| 5 | date          | Date          | Date on which sbp recorded      | Not Applicable                          |
| 6 | sbp           | Numeric       | systolic blood pressure (mm Hg) | 999 = sbp > 200                         |

**Example 3:** This example shows some common data formatting errors:

### Spreadsheet

The screenshot shows a Microsoft Excel window titled "Microsoft Excel - Badexample3". The spreadsheet has columns A through G. The data is as follows:

|   | A           | B         | C        | D         | E              | F           | G |
|---|-------------|-----------|----------|-----------|----------------|-------------|---|
| 1 | patient     | Sex / age | Race     | Height    | Blood pressure | Weight      |   |
| 2 | Jones, Mary | F / 43    | Black    | 70 inches | N/D            | 160         |   |
| 3 | Smith, joe  | M / 24    | Oriental | 65 inches |                | broke scale |   |

1. *patients* are identified by name
2. values for two variables (*sex* & *age*) are entered in the same cell: F / 43 and M / 24
3. units (inches, in this case) are included with value of *height*
4. both N/D and a blank cell are used to indicate missing data for *blood pressure*
5. a comment ("broke scale") is entered in place of a missing value (.) for *weight*

**Example 3: corrected, and augmented with data for a third patient.**

### Spreadsheet

The screenshot shows a Microsoft Excel window titled "Microsoft Excel - Example3". The spreadsheet has columns A through H. The data is as follows:

|   | A       | B   | C   | D    | E      | F   | G      | H           |
|---|---------|-----|-----|------|--------|-----|--------|-------------|
| 1 | patient | Sex | Age | Race | Height | BP  | Weight | Wtcomment   |
| 2 | 1       | F   | 43  | B    | 70     | .   | 160    |             |
| 3 | 2       | M   | 24  | O    | 65     | .   | 999    | broke scale |
| 4 | 3       | F   | 57  | W    | 62     | 130 | 100    |             |

Because patient 2 weighed more than 500 pounds, his weight is coded as "999". The data dictionary is shown below:



### Data Dictionary for Example 3

|    | B            | C                               | D                               |
|----|--------------|---------------------------------|---------------------------------|
| 1  | Variable     | Variable Definition             | Category Codes                  |
| 2  | Type         | (and Units)                     | (values and definitions)        |
| 3  | Categorical  | patient identifier              | unique integer for each patient |
| 4  | Categorical  | sex of patient                  | M = Male                        |
| 5  |              |                                 | F = Female                      |
| 6  | Numeric      | Patient Age (years)             | Not Applicable                  |
| 7  | Categorical  | Patient Race                    | B = Black                       |
| 8  |              |                                 | W = White                       |
| 9  |              |                                 | O = Oriental                    |
| 10 |              |                                 | X = Other                       |
| 11 | Numeric      | Patient Height (Inches)         | Not Applicable                  |
| 12 | Numeric      | systolic blood pressure (mm Hg) | Not Applicable                  |
| 13 | Numeric      | Patient weight (pounds)         | 999 = weight > 500              |
| 14 | Alphanumeric | Weight Comment                  | Miscellaneous text              |

## **Additional Information**

**Alternative formats:** Contact us if your data are not in Excel worksheets. We can analyze data from most spreadsheet and database programs, and from most statistical packages; however, data should be formatted according to the rules described above. We also accept properly formatted text files.

- ***Data embedded in Word or WordPerfect documents will often need to be re-formatted and placed in an Excel worksheet.***

For large datasets with repeated measurements, it is often more efficient to record baseline data in one file, and the repeatedly measured data in another file. However, both files must have identical patient identifiers so that the two files can be properly linked. See example 4 below.

**Use of text for values of categorical variables:** We discourage use of text for the values of a categorical variable; our experience is that text is much more subject to various data entry errors than numeric or character codes. However, if your dataset already contains text entries, we can analyze it under the following conditions: there is no evidence of data entry errors; and each value of a categorical variable is identified by a single unique text string.

**Alternative formats for dates:** We can handle most date formats, but a single format should be used for an entire dataset. We prefer **mm/dd/yyyy** since it is less prone to data entry errors than most other formats. The year should be indicated with 4 digits.

**Numeric formats:** Scientific notation (1.3E+10, 5.6E-5, 2.3E+1,...) is an acceptable format, but may be more prone to data entry errors than decimal or integer formats. If you use Excel, you may mix the three numeric formats in a dataset, even for a single variable. Do not, however, mix text and numeric data for a variable: e.g., do not use both “1” and “M” and “2” and “F” for the variable *sex*.

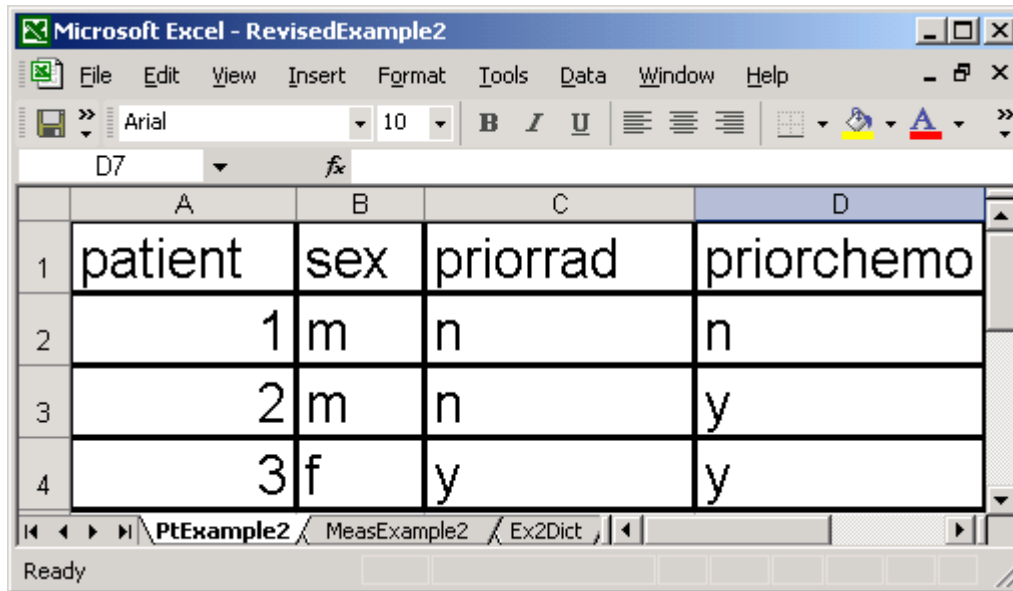
**Missing data codes:** Missing data codes should be used to identify the reasons why data are missing, unless that information is of no interest or would be redundant. For example, a woman’s age at first pregnancy could be missing either because the age is unknown, or because the woman was never pregnant. (In this case, one should use two numbers that could not possibly be a woman’s age at first pregnancy to code the missing values, e.g., 98 = age unknown and 99 = never pregnant.) If the number of children were also provided for all women, a missing data code to distinguish between these two cases would be redundant. However, if the age is unknown, it may still be useful to distinguish between the following two situations: 1) the age is expected to be determined in future follow-up, and 2) the age is not expected to be determined.

**Additional data required for retrospective chart reviews or similar studies:**

1. The data file should include every subject entered into the study, regardless of whether complete data were obtained for all subjects.
2. When measurements are recorded in a certain week or month of the study (say, at baseline or 1 month after surgery), include this information (e.g., coded as: 0 = baseline, 1 = 1 month) as well as the actual calendar date on which the visit occurred.
3. When information on patient response, duration of response, progression-free survival or overall survival is to be analyzed, we urge that the requisite information be recorded in our Excel data sheet, using the indicated coding: [Spreadsheet for Response, PFS & OS.xls](#)

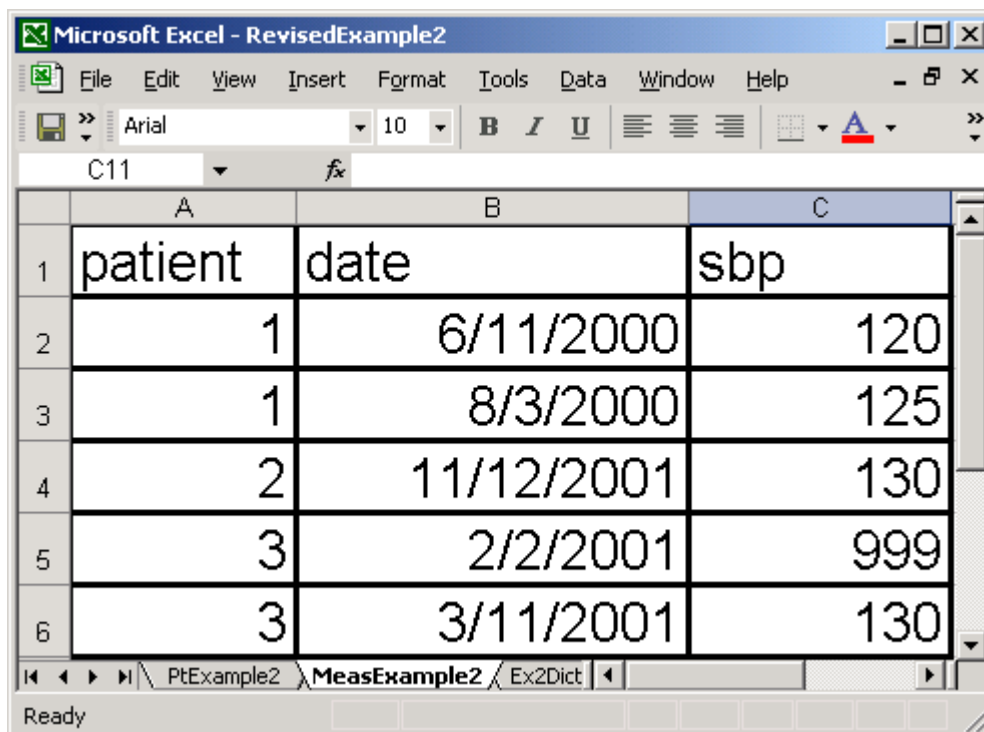
**Example 4: same data as in example 2, but separated into 2 files.**

**Baseline data file** (For UPMC HCC investigator-initiated clinical trials, we would obtain baseline data from the CRS database -- CTMA; you would not need to provide it.)



|   | A       | B   | C        | D          |
|---|---------|-----|----------|------------|
| 1 | patient | sex | priorrad | priorchemo |
| 2 | 1       | m   | n        | n          |
| 3 | 2       | m   | n        | y          |
| 4 | 3       | f   | y        | y          |

**Repeated measurements of systolic blood pressure (sbp) over time**



|   | A       | B          | C   |
|---|---------|------------|-----|
| 1 | patient | date       | sbp |
| 2 | 1       | 6/11/2000  | 120 |
| 3 | 1       | 8/3/2000   | 125 |
| 4 | 2       | 11/12/2001 | 130 |
| 5 | 3       | 2/2/2001   | 999 |
| 6 | 3       | 3/11/2001  | 130 |

*Data dictionary for example 4*

|    | A          | B           | C                               | D                               |
|----|------------|-------------|---------------------------------|---------------------------------|
| 1  |            |             | Baseline data                   |                                 |
| 2  |            |             |                                 |                                 |
| 3  | Variable   | Variable    | Variable Definition             | Category Codes                  |
| 4  | Name       | Type        | (and Units)                     | (values and definitions)        |
| 5  | patient    | Categorical | patient identifier              | unique integer for each patient |
| 6  | sex        | Categorical | sex of patient                  | m = Male                        |
| 7  |            |             |                                 | f = Female                      |
| 8  | priorrad   | Categorical | prior radiation                 | y = yes                         |
| 9  |            |             |                                 | n = no                          |
| 10 | priorchemo | Categorical | prior chemotherapy              | y = yes                         |
| 11 |            |             |                                 | n = no                          |
| 12 |            |             |                                 |                                 |
| 13 |            |             | Repeatedly measured data        |                                 |
| 14 |            |             |                                 |                                 |
| 15 | Variable   | Variable    | Variable Definition             | Category Codes                  |
| 16 | Name       | Type        | (and Units)                     | (values and definitions)        |
| 17 | patient    | Categorical | patient identifier              | unique integer for each patient |
| 18 | date       | Date        | Date on which sbp recorded      | Not Applicable                  |
| 19 | sbp        | Numeric     | systolic blood pressure (mm Hg) | 999 = sbp>200                   |