

Data Management Using EpiData

Wayne Enanoria, PhD, MPH
Center for Infectious Disease Preparedness
School of Public Health
University of California at Berkeley

Version 8 July 2007

Table of Contents

Introduction.....	1
Getting EpiData.....	1
Useful Resources for Learning EpiData.....	1
Data Entry using EpiData.....	2
Introduction.....	2
Steps of the Data Management Process.....	2
Different Types of Files.....	3
Creating a Data Entry Screen.....	3
Variable Types.....	4
Defining Variable Types.....	5
Example .QES File.....	5
Null Value Considerations.....	6
Preview Data Form.....	6
Creating a Data File (.REC File).....	7
Data Validation and Checks.....	7
Functions.....	7
Exporting Data from EpiData.....	9

Introduction

The purpose of this document is to provide an overview of the basic functions of EpiData and their use for managing epidemiologic data.

This document was created using EpiData Entry version 3.1 Build: 12Mar2006¹.

Getting EpiData

In order obtain a version of EpiData for data entry and for data analysis, download both programs by visiting <http://www.epidata.dk>. These two programs constitute EpiData software.

Useful Resources for Learning EpiData

There are many great resources for learning EpiData. One in particular is:

¹ Lauritsen JM. (Ed.) EpiData Data Entry, Data Management and basic Statistical Analysis System. Odense Denmark, EpiData Association, 2000-2006. [Http://www.epidata.dk](http://www.epidata.dk).

- Data Management for Surveys & Trials. A practical primer using EpiData
Steve Bennett, Mary Myatt, Damien Jolley, and Andrzej Radałowicz
URL: <http://www.epidata.dk/php/downloadc.php?file=dmePIData.pdf> .

Check the EpiData website for more examples that illustrate how to use the software.

Data Entry using EpiData

Introduction

In creating data management systems, it's important to think about several aspects of the project.

1. Data collection procedures
2. Data entry
3. Data validation and checking
4. Data manipulation
5. Data analysis
6. Reporting

EpiData has several features that assist epidemiologists in conducting epidemiologic investigations. EpiData contains the following features:

1. Text editor
2. Data Entry
3. Data Checking
4. Coding and Calculation
5. Import/Export capabilities
6. Data management (eg, append or merge datasets)
7. Documentation
8. Other Utilities

Steps of the Data Management Process

EpiData uses a general process for managing its data. The main features of the EpiData Entry program are:

1. Define Data
2. Make Data File
3. Checks
4. Enter Data
5. Document
6. Export Data

When you open EpiData, you can see these six functions as buttons at the top of the screen (below the toolbar). If you are creating a data management file from scratch, you will most likely start with step (1) Define Data and proceed in sequence.

Different Types of Files

EpiData uses different file types for different functions. The following are the different types of files:

1. .QES file (questionnaire file; contains the questions and file types for each field)
2. .REC file (data file)
3. .CHK file (check file; contains the data checking and validation rules for the data)
4. .NOT file (note file; contains any notes you create for a particular datafile)
5. .LOG file (log file; this file gets created by the data documentation function)

One usually creates a .QES file that mimics the data collection form or questionnaire from which you would like to create a data management system. Once the .QES file is created, one can then create a datafile (.REC file) in EpiData.

Creating a Data Entry Screen

The .QES file is where one identifies the data type, name, and length of each individual data element.

Defining Fields

For each field (also called “variable”) that you create, you must tell EpiData its name, type, and length (ie, the anticipated maximum number of letters and/or numbers). The variable type should be determined by the type of data that the variable is to contain. For example, an address field will contain text (and possibly numbers that can be saved as text). Below are some examples.

Name	Description	Type	Length
Case ID Number	CASEID	Numeric	4
First Name	FNAME	Text	20
Last Name	LNAME	Text	20
Soundex of Surname	LNCODE	Soundex	5
Interview Date	INTDATE	Date	10
Today’s Date	TODAYDT	Date	10
Age at Interview	AGEYRS	Numeric	3
Currently Ill?	CURRILL	Text (Yes/No)	1
ICD-10 Code	ICD10	Numeric	4

NOTE: a numeric field with length three will be able to hold numbers ranging from -99 to 999.

Remember that names for fields must adhere to the following rules:

1. Names must not exceed eight characters;
2. Names must begin with a letter, not a number;
3. Names must not contain any spaces or punctuation marks.

Variable Types

Text

Text data types are useful for storing text or numeric information such as names and addresses.

Text (Upper Case)

A special type of text data type is the upper case text. This data type will store whatever is entered as upper case (capital) letters, regardless of the way you enter the data. For example, if you type “wayne” (in all lower case letters) into a field that stores data as upper case text, EpiData will store the item as “WAYNE” in the dataset.

Numeric

Numeric variables store numbers (eg, age, birthweight, etc.). They can be used for storing categorical or continuous variables, and can store integers (whole numbers) or real numbers (numbers with fractions). If you know that you will be performing mathematical operations with a particular field, it must be of the numeric field type in order to do the operations (using EpiData Analysis).

Date

The date field type can store dates in two different formats:

- mm/dd/yyyy (American format)
- dd/mm/yyyy (European format).

Today’s Date

You can create a variable that automatically stores today’s date.

Yes/No

This data type is useful for storing binary categorical data such as “Yes” or “No” (which can also be stored as “1” or “0”).

Auto ID Number

Auto ID Number field type is a unique number that is assigned to each observation entered into a data file. This unique number usually starts at 1 and is assigned to the first record of the dataset and is incremented by one for every new record that is entered thereafter.

SOUNDEX

Soundex is a special type of text variable that applies SOUNDEX coding rules to text data as it is entered. The code is supposed to be a representation of the text as it sounds (rather than the way it is spelled). The SOUNDEX code consists of one uppercase letter, a hyphen, followed by three numbers.

In order to determine the soundex code for a particular text (eg, the last name “Enanoria”), do the following:

1. Eliminate any a, e, i, o, u, h, w, y (we are now left with “nnr”)
2. Assign numbers to the letters according to the following assignment.
 - a. 1=b, p, f, v
 - b. 2=c, s, k, g, j, q, x, z
 - c. 3=d, t

- d. 4=l
- e. 5=m, n
- f. 6=r

Double letters and adjacent letters with the same numeric code are coded only once. Zeroes are added if you run out of letters before you have three numbers. (We now have “556”)

- 3. Write the first letter as it is followed by a hyphen (eg, “E-556”).

For a detailed discussion of Soundex Coding System, see:

<http://freepages.genealogy.rootsweb.com/~bbunce77/CensusSoundex.html> .

Phone Number

NOTE: EpiData does not support PHONENUM variable types, as Epi Info does. However, telephone numbers can be entered into ordinary text fields.

Time

EpiData does not support time fields. However, the help file states that there are two functions, Time2Num and Num2Time, that allow one to convert numeric data to time data and vice-versa. Refer to the EpiData help files for more information.

Defining Variable Types

So for the above variables, I could represent them as the following in a QES file:

Description	Type	Length	Definition
CASEID	Numeric	4	<IDNUM>
FNAME	Text	20	<hr/> (20 underscore characters)
LNAME	Upper Text	20	<AAAAAAAAAAAAAAAAAAAA>
LNCODE	Soundex	5	<S >
INTDATE	Date	10	<mm/dd/yyyy>
TODAYDT	Date	10	<Today-mdy>
AGEYRS	Numeric	3	###
CURRILL	Text (Yes/No)	1	<Y>
ICD10	Numeric	4	###.#

In creating .QES files, one should not use any of the following characters except for defining field types: “_”, “<”, “>”, and “#”. These symbols are used for creating field types in EpiData.

Example .QES File

For example, your .QES file could contain the following in order to create the above mentioned fields and field types.

Personal Data		
Case ID Number	{CASEID}	<IDNUM>
1. What is the person's first name?	{FNAME}	<AAAAAAAAAAAAAAAAAAAA>
2. What is the person's last name?	{LNAME}	<AAAAAAAAAAAAAAAAAAAA>
	{LNCODE}	<S >
3. What is today's date?	{INTDATE}	<mm/dd/yyyy>
4. What is the person's age?	{AGEYRS}	###
5. Has the person been sick in the past week?	{CURRILL}	<Y>
5a. If "yes", code condition.	{ICD10}	###.#

Note: the names that you would like to assign to each data element should be given in curly brackets (eg, {CASEID}).

Alternatively, you can put curly brackets around whatever you want to name the field. For example, say we wanted to name question 5 with the name PERSICK instead of CURRILL. We could have written Question 5 as follows:

Question 5. Has the {PER}son been {SICK} in the past week?

The field name {PERSICK} would no longer be needed with this version of the question.

Null Value Considerations

In collecting information using a survey tool (questionnaire), there are oftentimes questions that do not apply to the respondent. For example, a question asking about a woman's pregnancy are not applicable to a male respondent. Therefore, one should think about how the answer to these questions will be coded in the dataset for male respondents so that it is clear that the questions are not applicable and should not be completed. This situation is different from a skipped question where a respondent just failed to answer a particular question but should have.

As a result, some questions are given a "not applicable" code such as "8", "88", or "888", or something like it. It is really important to think about null values (both true missing values and not applicable answers) and how they will be dealt with both in the data collection process as well as the data entry/management process so that there is consistency throughout the process. EpiData assigns all missing data with a special code: ".".

Preview Data Form

Once you've created your data form (QES file), it's a good idea to preview what the form will look like when you enter data into it via the Data Entry mode in EpiData.

To preview your data form:

1. On the toolbar, select Data file.
2. Select Preview data form.

Alternatively, you can press CTRL+T to preview your data form. What you see on screen with

the QES file is not always how it will appear when you enter data, so it's a good idea to check.

Creating a Data File (.REC File)

Once you have created your QES file, you can then create a data file (EpiData saves its data files with a .REC extension).

To create a .REC data file, you must:

1. Open a .QES file.
2. Click on the button under the toolbar that says "2. Make Data File".
3. Select "Make Data File".
4. A dialog box will open up that has two boxes. The first box says "Enter name of .QES file". This specifies the path and filename of the .QES file that you would like to create the .REC file from. The second box says "Enter name of data file". This enables you to specify the path and filename of the new data file you want to create.
5. Click "OK" when done.

Data Validation and Checks

There are several functions that are available in order to control how data is entered or certain "checks" of the data that should be performed. Whenever you establish checks for any of the variables in your dataset, these checks get "coded" and saved to a .CHK file.

Functions

Range, Legal

Specifies a range of values that can be entered (continuous scale), or specifies legal values to be entered (categorical).

The check code associated with a range looks something like this.

```
V1
  *Restrict data entry to values in range 1.0-2.10
  RANGE 1.0 2.10
END
```

One can also set up legal values for a particular field that controls what can be entered into the field. For example, you may want to restrict a field to only enter in certain values into it. Let's say we have a variable called "ATTEND" and we only want to allow 1 (Yes) or 2 (No). The check code would look like this.

```
ATTEND
  LEGAL
  1
  2
  END
END
```

Jumps

If certain values are entered in a particular variable, jumps tell EpiData where to go given the value specified. This feature is good for implementing skip patterns in a questionnaire.

Let's say I want to set up a conditional jump for the variable called "ANYELSE". If someone enters a "1", I want the cursor to jump to the field called "ELSENAME1"; if someone enters a "2", I want the cursor to jump to the field called "STYHOTEL".

The check code associated with this jump would look something like this.

```
ANYELSE
  JUMPS
    1 ELSENAME1
    2 STYHOTEL
  END
END
```

Must Enter

If set to "Yes", a value must be entered before the computer will allow the data entry to continue. No blanks will be accepted if a value must be entered.

A Must Enter field's check code might look like this.

```
ASSIGNID
  MUSTENTER
END
```

Repeat

This function tells EpiData to repeat the value for a particular variable that was entered in the previous record. This is good for entering in state for addresses if your investigation is limited to a geographical area in one state; it prevents you from having to enter in the same state for everyone.

```
STATE
  REPEAT
END
```

Value Label

For categorical variables, this defines the meaning of the values (eg, 1=female, 2= male). The "+" allows you to pick an existing label definition.

Edit

The dialog box allows you to add functional checks without having to know the check code associated with doing so. The "Edit" button allows you to see the check code and add additional code if desired.

Save

Saves any changes you made to a particular variable.

Exporting Data from EpiData

One can export data into the following formats using EpiData:

- Text
- Dbase III
- Excel
- Stata
- SPSS
- SAS