



APHEO
Association of Public Health Epidemiologists in Ontario



EpiData: The Relevance for Local/Regional Chronic Disease Risk Factor Surveillance

**Created as part of a collaborative project¹ :
Association of Public Health Epidemiologists in Ontario (APHEO), the EpiData Association and the Public Health Agency of Canada.**

**Authors:
Association of Public Health Epidemiologists in Ontario (APHEO) EpiData Expert Panel,
Jens Lauritsen, EpiData Association & University of Southern Denmark, and
Canadian Alliance for Regional Risk Factor Surveillance (CARRFS)**

V3.0 Feb 2011

¹ This project has been made possible through a financial contribution from the Public Health Agency of Canada

Table of Contents

Purpose and scope	3
Introduction – The STEPwise Approach to Surveillance (STEPS) Program, World Health Organization	3
Purpose	3
Design.....	3
Introduction – EpiData Entry and Analysis Software	4
Preparation	4
A. Read about data management and structures	4
B. Acquire and install software.....	4
C. Starting a New Project	7
PART 1: EpiData Entry	8
Planning the Questionnaire.....	8
Create the Questionnaire	8
Create the Data File.....	11
Enter Data in the Questionnaire.....	11
Helpful Hints.....	14
PART 2: Creating Check Files (Data Validation).....	15
Specifying Data Entry Rules/Checks	15
Value Labels	16
Range Legal	17
Jumps (Skips).....	19
Helpful Hints.....	22
PART 3: EpiData Analysis	23
Opening and preparing the dataset.....	23
Fine tuning the dataset for analysis (data cleaning).....	25
Descriptive Epidemiology	27
Analytical Epidemiology.....	33
Test of Association.....	33
Strength of Association.....	33
Stratified Analysis using CIplot.....	35
Helpful Hints.....	38
PART 4: Working with Programs.....	39
Questions/For More Information	41
Appendix 1 – STEPS Questionnaire	42

Purpose and scope

Participants should have prior understanding of chronic disease and risk factor surveillance methods and the associated statistical methods.

The structure of this guide includes the principles of defining and entering data followed by analysis principles for a dataset. Data used herein are partially modified for exercise purposes and cannot be published except by written permission.

This instruction is based on a modified dataset provided by the World Health Organization's STEPS program.

Introduction – The STEPwise Approach to Surveillance (STEPS) Program, World Health Organization

The WHO STEPwise approach to Surveillance (STEPS) is a simple, standardized method for collecting, analysing and disseminating data in WHO member countries.

By using the same standardized questions and protocols, all countries can use STEPS information not only for monitoring within-country trends, but also for making comparisons across countries. The approach encourages the collection of small amounts of useful information on a regular and continuing basis.

There are currently two primary STEPS surveillance systems, the STEPwise approach to risk factor surveillance and the STEPwise approach to Stroke surveillance. This field guide is based on the STEPwise approach to risk factor surveillance which focuses on obtaining core data on the established risk factors that determine the major disease burden. It is sufficiently flexible to allow each country to expand on the core variables and risk factors, and to incorporate optional modules related to local or regional interests.

Purpose

The WHO STEPwise approach to chronic disease risk factor surveillance provides an entry point for low and middle income countries to get started on chronic disease surveillance activities. It is also designed to help countries build and strengthen their capacity to conduct surveillance.

Design

The STEPS Instrument covers three different levels or "steps" of risk factor assessment:

1. Questionnaire - self-report measures that all countries should obtain. In addition to socio-economic data, data on tobacco and alcohol use, some measure of nutritional status and physical inactivity are included as markers of current and future health status.

2. Physical measurements - includes simple physical measurements, such as height, weight, waist circumference, and blood pressure.
3. Biochemical measurements - requires access to the appropriate standardized laboratories. Collecting and analysing blood samples is a relatively complex process and can be done only in the context of a comprehensive survey and in settings where appropriate resources are available.

In addition, EpiData is the software used by STEPS for data entry. For additional information on the STEPS initiative see: <http://www.who.int/chp/steps>

Introduction – EpiData Entry and Analysis Software

EpiData is a free software suite designed to assist epidemiologists, public health investigators and others to enter, manage and analyze data in the field. All software is available from <http://www.epidata.dk>. A number of field guides, software documentation notes, examples and other information are also available. Users are encouraged to:

- (1) Join the EpiData-list discussion group, and
- (2) Sign up for the information newsletters sent periodically each year.

To subscribe to the discussion group and sign up for the newsletters, visit the EpiData website at <http://www.epidata.dk>

If you work in an institution, the expectation is that the institution adds work or funding to the development of the software - the continued release of EpiData as freely downloadable from the internet depends on this.

Preparation

Follow steps A, B and C:

A. Read about data management and structures

Go through the document:

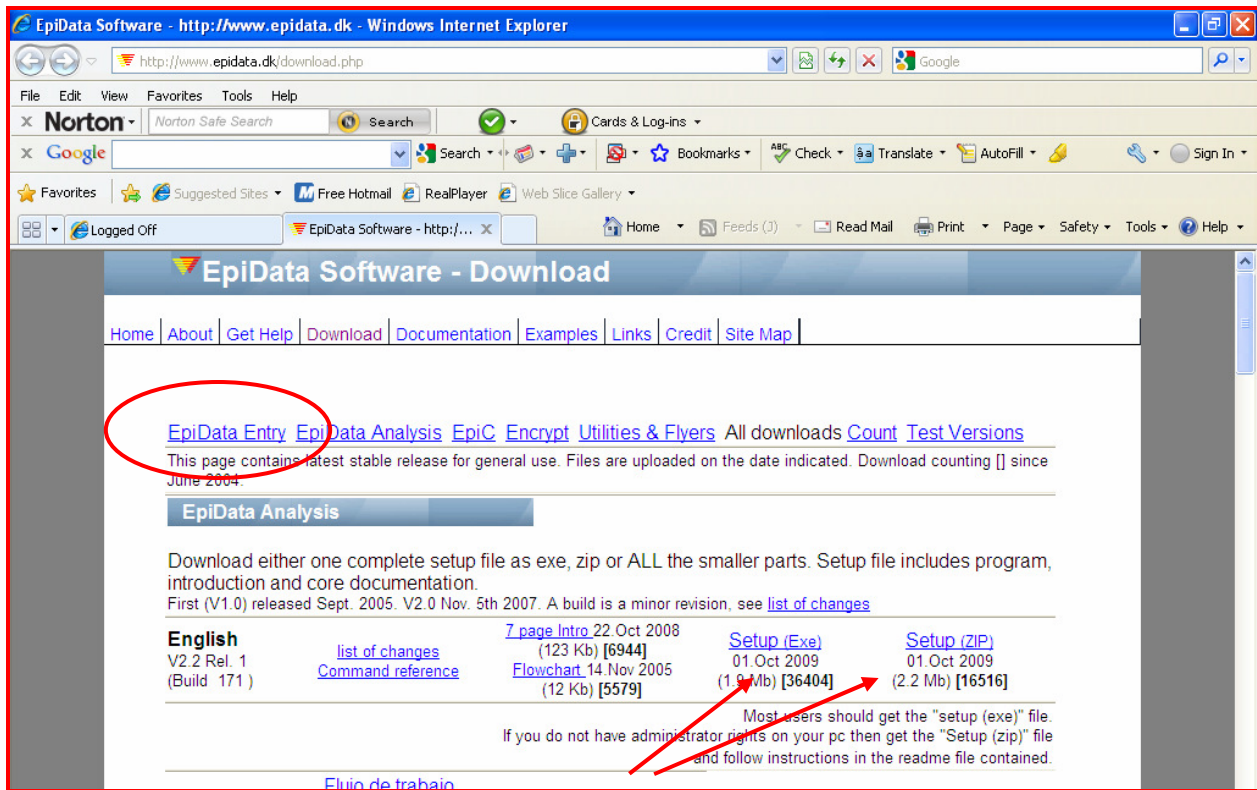
http://www.epidata.org/wiki/index.php/Field_guide:_from_question_to_data which discusses data structures, data management and documentation.

B. Acquire and install software

Download EpiData Analysis (v2.2 or later) and EpiData Entry (v3.1 or later) and install these on your computer. You should use the most recent version available from the download section of the website <http://www.epidata.dk/download.php> If installation on your

computer requires the permission of an administrator, you can install the software into your private folder (create a subfolder on the desktop) or on a USB key².

To download EpiData Analysis, you will need to download and run one of the setup files indicated below. Most users should get the "Setup (exe)" file. If you do not have administrator rights on your pc then get the "Setup (zip)" file and follow the instructions in the readme file.



To download EpiData Entry, download and run the "Complete Setup" file indicated below.

²It is also good practice to talk with the administrator. All EpiData Software is checked with up-to date anti-virus and other malicious software protection agents before uploading to epidata.dk.

link [French](#)

EpiData Entry

Entry is suited for entry and documentation of data. Download the "setup" file in the language of your choice. Setup file includes program and core documentation. A translation consists of documentation and texts for menu's etc. The actual program file is the same for all languages. Acknowledge translators.

Current version: **3.1 Build: (27jan2008)**. A build is a minor revision, see [list of changes](#)

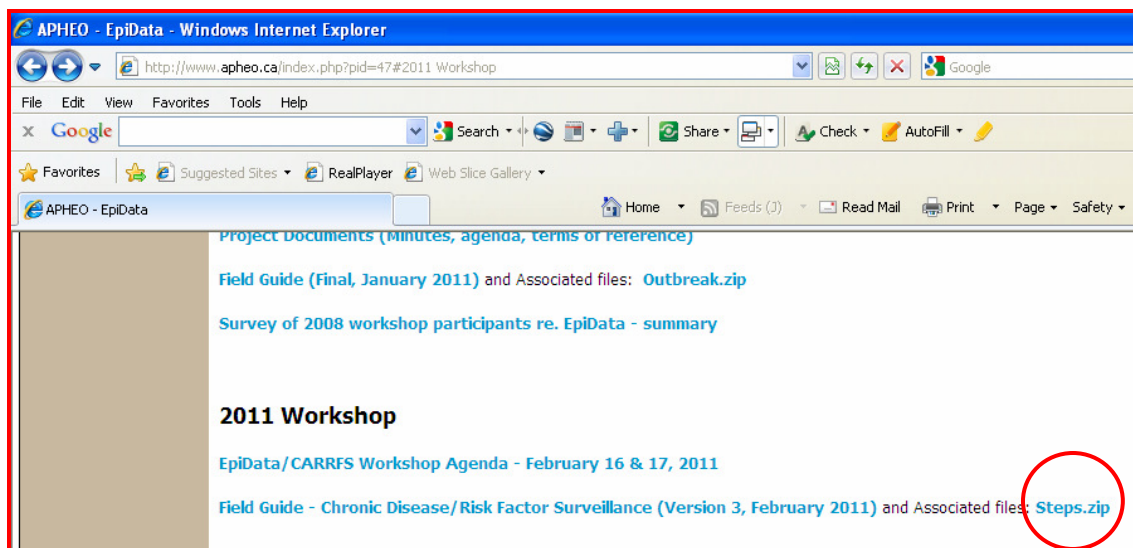
English	Flowchart 27 Aug 2005 (19 Kb) [9278]	4 page Intro 25 Aug 2005 (49 Kb) [8393]	Complete Setup 28 Jan 2008 (0.9 Mb) [44763]	Extended help (pdf) 19 Jan 2005 (0.5 Mb) [17300]
English	Get this if you only need latest build and have documentation and language files. Unzip and replace previous epidata.exe.			Zip - exe only 28 Jan 2008 (0.7 Mb) [7103]

C. Starting a New Project

Create and name a new folder

Create a project folder that will house all data files associated with your investigation – this folder can be located anywhere on your computer; you do not have to create one in the same drive as the EpiData program files. For the purposes of this exercise, create a new project folder on your desktop titled "STEPS".

Get the necessary exercise data file from the Association of Public Health Epidemiologists in Ontario's (APHEO) website ([http://www.apheo.ca/index.php?pid=47#2011 Workshop](http://www.apheo.ca/index.php?pid=47#2011%20Workshop)). Save the file into your project folder.



PART 1: EpiData Entry

Before you create a questionnaire and enter data, the following steps must be considered:

- What is the purpose of the investigation?
- Organizational aspects. e.g. Will the investigation take place in one or several sites?
- What sampling design of observations is appropriate and which sources of information can you use? (interview, registry data, laboratory data)
- Which data structure will enable us to do the investigation in view of sampling design, purpose, organization and types of data sources?
- What type of analysis is planned, including anticipated reporting needs?

Planning the Questionnaire

Determine its purpose.

Task 1: The first task is to define groups of information. Write down titles of types of information you would include, for example: Demographic factors, risk factors that are associated with common chronic diseases

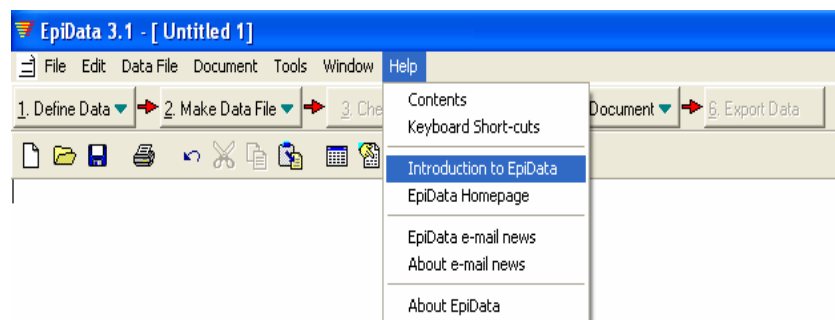
Task 2: Formulate the questions you want to include. A simple first version of a data structure could be:

age, sex, public health unit, risk factorsⁱ
where **i** indicates several of these.

Create the Questionnaire

Task 3: Now define the actual data structure.

First: Start EpiData Entry and open the introduction pdf file by clicking on the **Help** button and selecting "**Introduction to EpiData**" in the droplist. Read this document beforehand.



Task 4:

Field naming

Names of the fields in a data entry form are created automatically from the contents of the .QES file. There are two different ways of naming fields:

1) First word in the question is used as the field name - If the length of the first word is more than 10 characters then the first 10 characters of the first word will be used as the field name. This is the default.

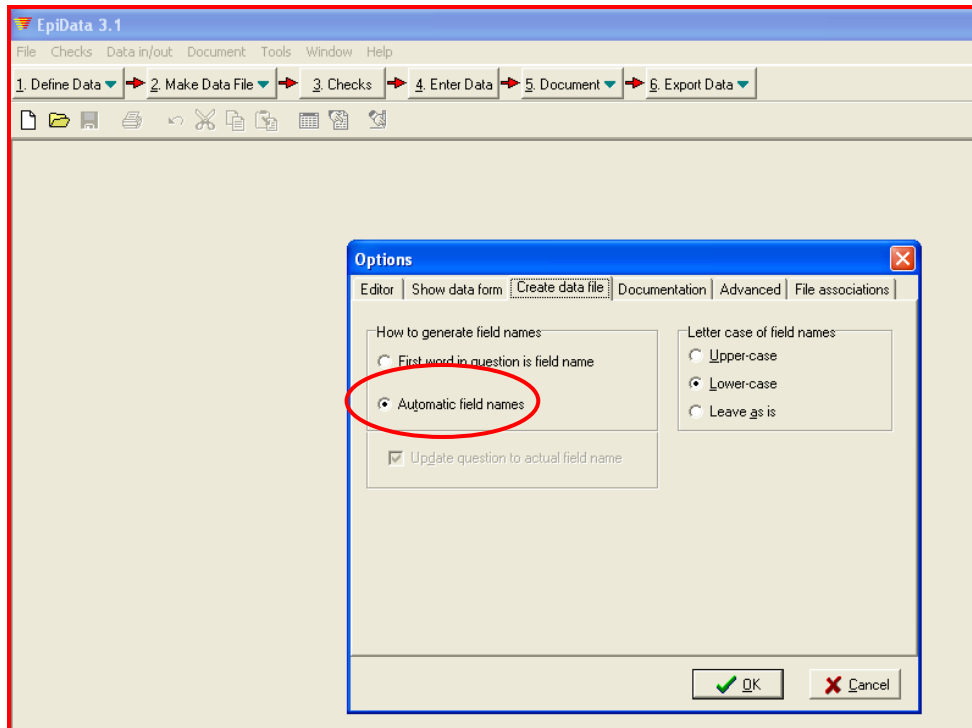
Examples:

- v1 Enter age of patient ### - the field name will be "v1".
- Enter age of patient ### - the field name will be "Enter". In this case it may be better to use the automatic field naming option.

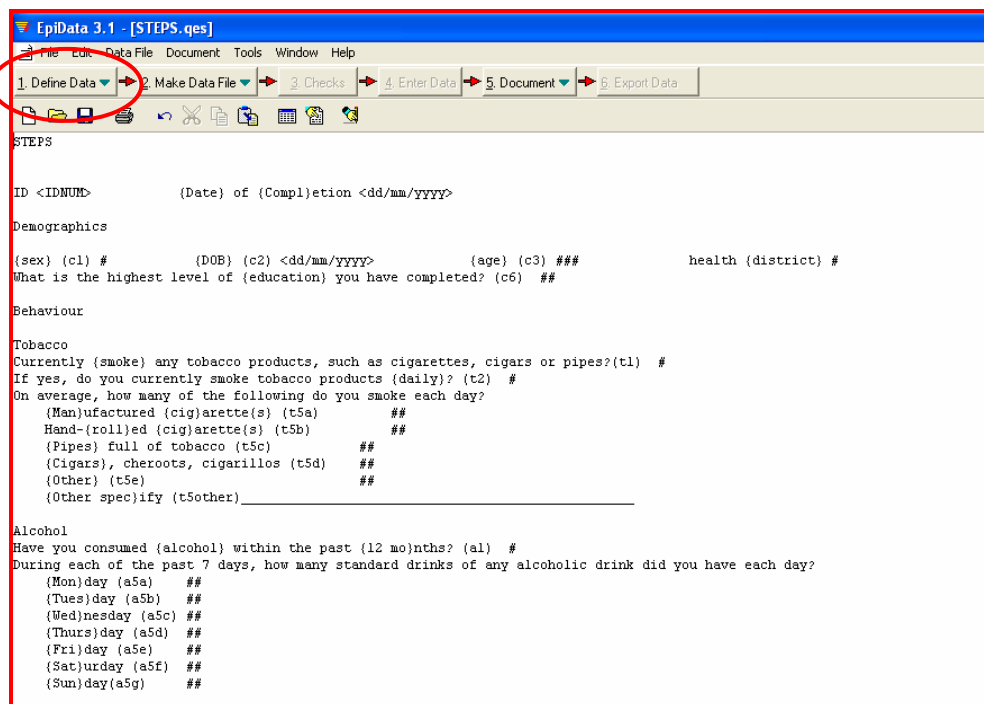
2) Automatic field naming – This is the option that is used in the STEPS.qes. The following rules are used when generating the field names:

- Text enclosed in curly brackets is used in preference to normal text. If the question is "{my} first {field}" then the field name will be MYFIELD.
- Common words are skipped (i.e. "what", "the", "of", "and" etc.). "What did you do?" generates the field name YOUUDO.
- Fields without a question get the same name as the previous field plus a number. If the previous field is named MYFIELD then the next field (if it has no question) is named MYFIELD1. If the previous field is named V31 then the next field is named V32. If no previous field exists then the default name FIELD1 is used.
- If the first character of the generated field name is a number then the letter N is inserted at the first character. "3 little mice" generates the field name N3LITTLEMI.


Before creating the questionnaire, ensure that Automatic field naming is selected. Go to File → Options → Create Data file.



Now you're ready to create your questionnaire. Write the text as shown in this picture (an example of a simple questionnaire/data structure) by selecting **"1. Define Data"** and then **"New .QES" File** in the droplist. The questionnaire is also available in [Appendix 1](#).



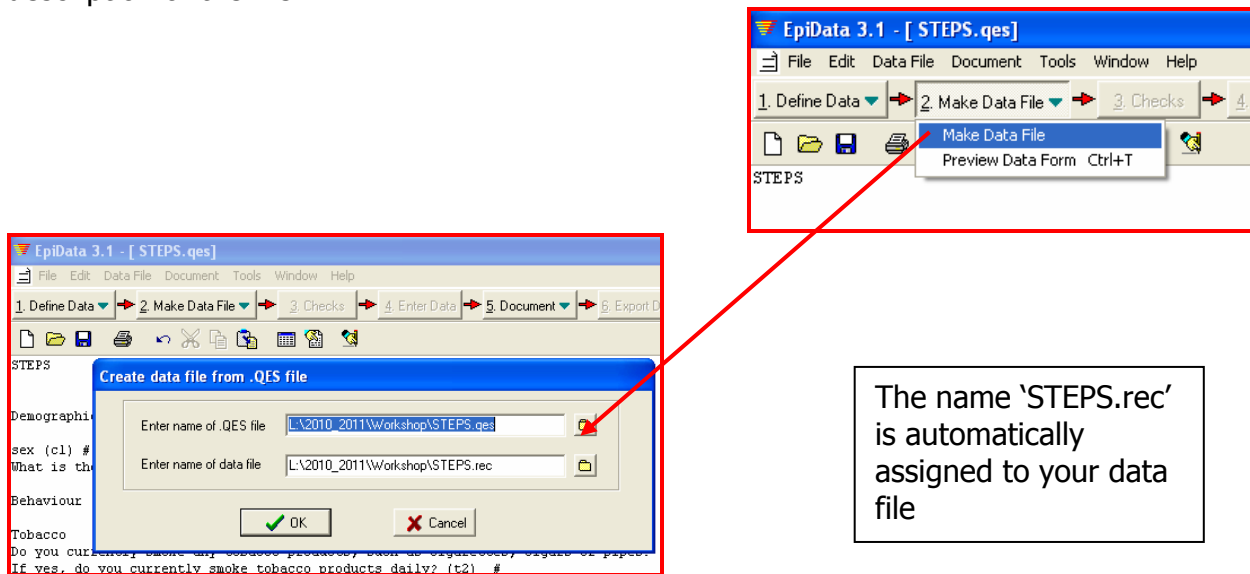
Each variable consists of a word which explains the content, e.g. "t1" and a specifier for the type of information, e.g., "#". The "#" indicates a numerical variable which will later be coded.

You can get further help with the 'field pick list'  (second icon from the right) to help specify variable types. Other field type options include text, Auto ID number, Boolean (Yes/No).

When finished, save the file in your project folder as "STEPS.qes" by clicking on the "save" icon.

Create the Data File

Task 5: Create a data file based on your questionnaire above. Choose "**2. Make Data File**" and select "**Make Data File**" in the droplist. You will be prompted to enter a description of the file.

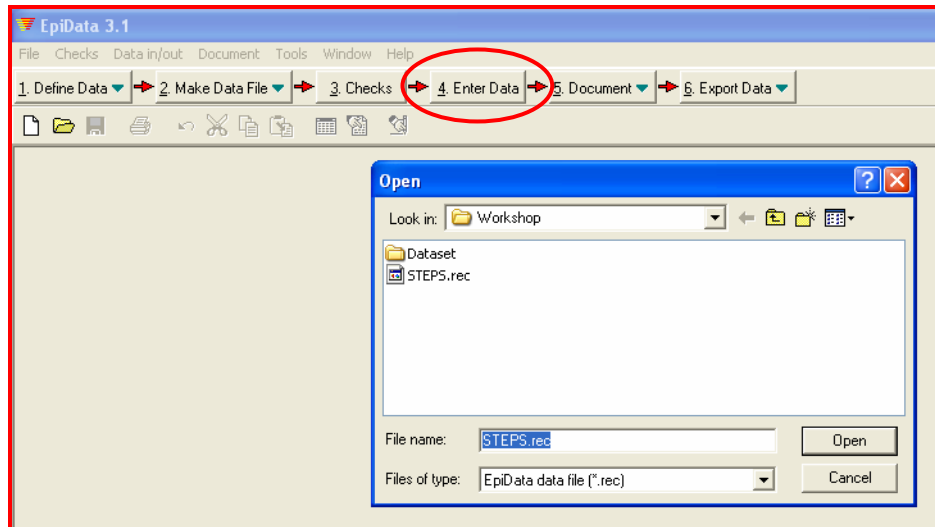


What can go wrong here is that you have not saved your .qes file first, or that you chose "preview data form" instead of "create data file".

Enter Data in the Questionnaire

During entry you **CANNOT USE THE MOUSE** – it is best to use the keyboard, tab and arrows. Some controls will not work if you use the mouse.

Task 6: Entering data in your newly created questionnaire
Close your STEPS.qes file (ensure it has been saved!). Choose "**4. Enter Data**" and select "STEPS.rec". Enter five "made up" records and see if you can enter an illegal date, e.g. February 29th in a "non-leap" year in the "Date of Birth" field.



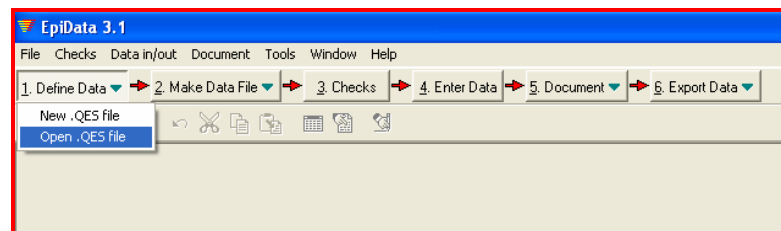
Task 7: Actual entry – how did it work?

You should now have noticed that by just defining the date of birth in the dd/mm/yyyy format, you already have sufficient control for not entering an illegal date. But in the other fields you could enter any number, whether it makes sense or not. Therefore, you must impose further controls.

Task 8: Revising your Questionnaire

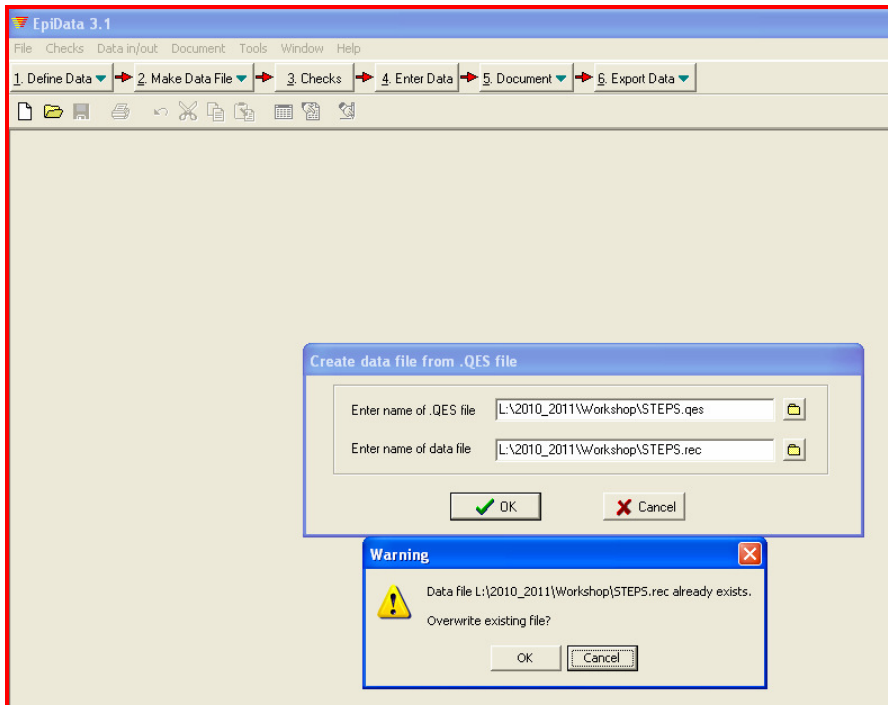
It turns out that you need two new variables: height and weight
Change the questionnaire to include these two new variables:

1. Reopen the STEPS.qes file and add the additional text:
 {Height} in centimeters (m3) ###.# {Weight} in kilograms (m4) ###.#
 Are you {pregnant}? (m5) #



Close the revised .qes file (you'll be prompted to save the file – give it the same name 'STEPS.qes')

2. Open for entry by clicking on "**4. Enter Data**" and select **OK** to update structure. Close after the update is finished



Helpful Hints

Field types in EpiData

Field type	Example	Explanation
ID number	<IDNUM>	automatic id number
Numeric	### ###.##	maximum field length: 14 characters
Text	_____	maximum field length: 80 characters
	<E >	encrypted field
Upper-case text	<A>, <A >	entries converted to upper case
Boolean	<Y>	accepts Y, N, 1, 0 and space as legal entries
Date	<dd/mm/yyyy> <mm/dd/yyyy> <yyyy/mm/dd>	10 characters in length
Today's date	<today-dmy> <today-mdy> <today-ymd>	filled automatically with the current date
Soundex	<S> <S >	a coding of words that can be used to anonymise
Tabulator code	@	changes the position of the fields in the form

Recreating a lost .QES file

If the .QES file that was used to create a data file is no longer available, it can be recreated:
Tools → QES File from REC File.

Deleting Records

During data entry, a record can be marked for deletion using [Shift] + [Delete]. The record is not removed from the data file but only marked for deletion.

To permanently delete all records marked for deletion: Tools → Pack File.

PART 2: Creating Check Files (Data Validation)

Specifying Data Entry Rules/Checks

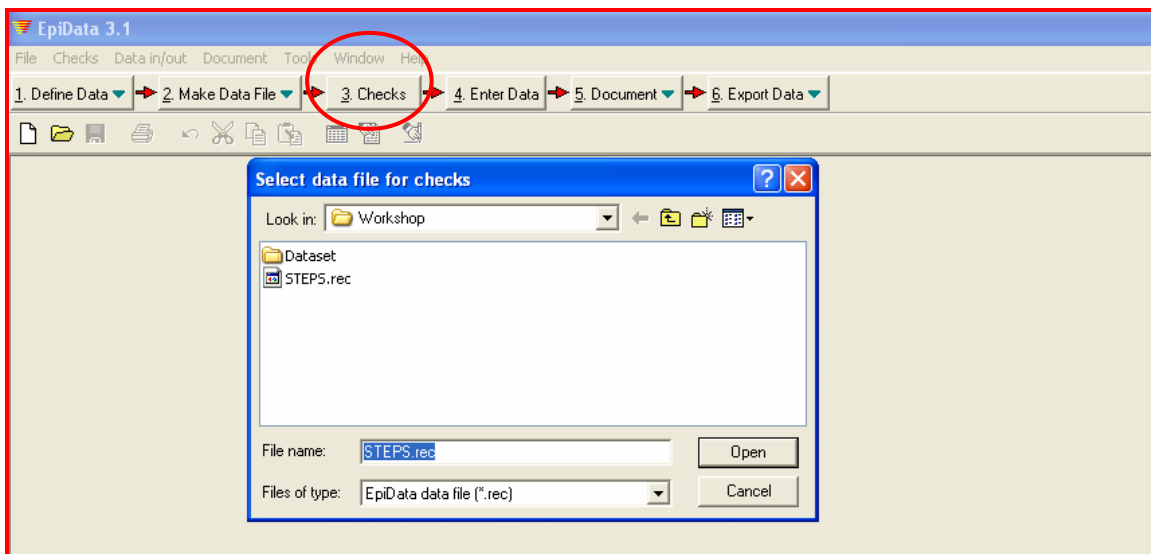
A strong component of EpiData is the ability to specify rules during data entry. For example, you can design your questionnaire to:

- Limiting entry of numbers or dates to a specific range or to a number of specified values and give text descriptions to the numerical codes entered.
- Specify sequence of data entry, e.g. fill out certain questions for males only (i.e., build in question jumps).
- Apply calculations during data entry, e.g. age at visit = date of visit - date of birth. Typically most calculations are done at the analysis stage.
- Forcing an entry to be made in a field

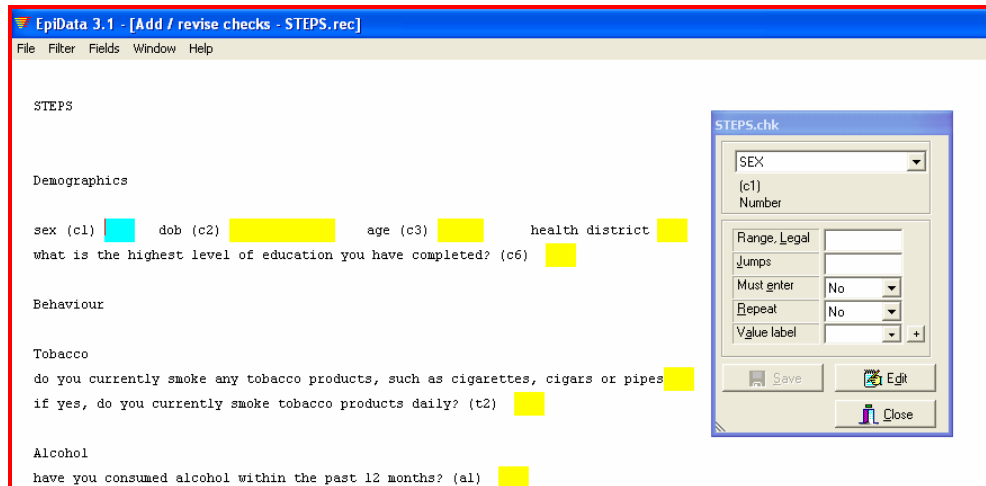
A check file must have the same name as the data file but with the extension .CHK instead of .REC.

Task 9: Add checks/controls by selecting the **"3. Checks"** section

Hint: Close data entry (F10) and read the section on adding checks in the Introduction to EpiData document.



Select the 'STEPS.rec' file. Your questionnaire will appear along with the 'check' window:

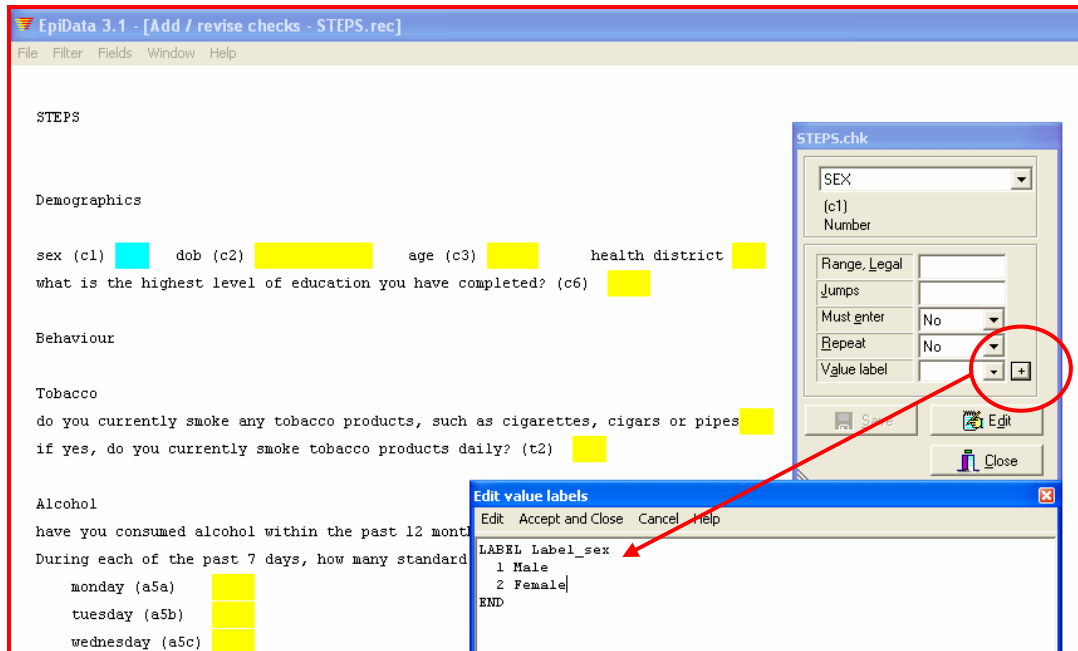


Value Labels

Earlier, you defined some of the data fields as numeric (#). Now, assign labels to the numeric values:

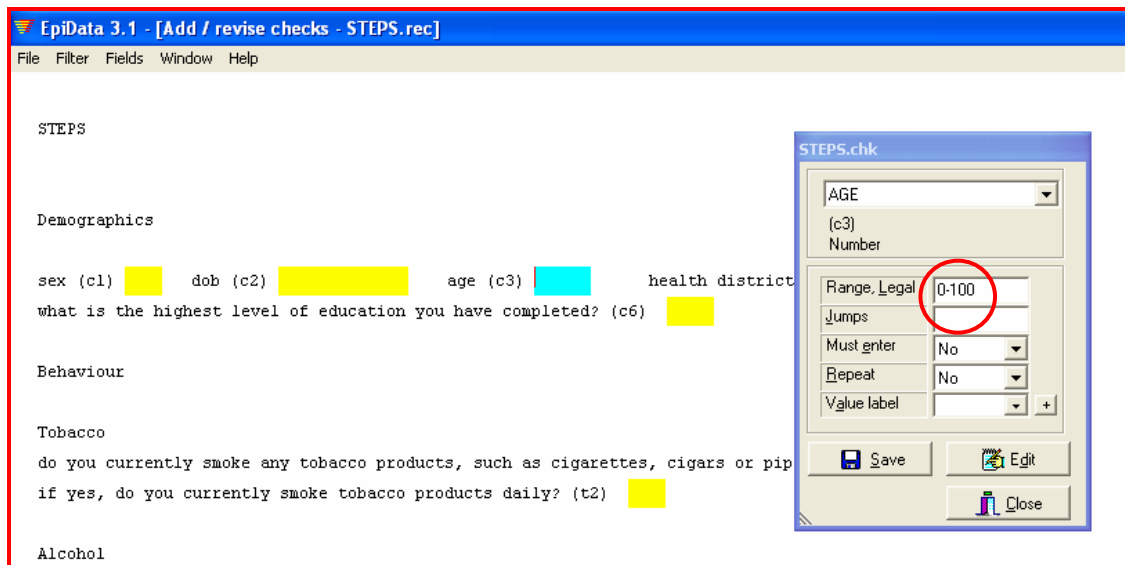
Sex	1 Male	2 Female
Education	1 "No formal schooling" 2 "Less than primary school" 3 "Primary school completed" 4 "Secondary school completed" 5 "High school completed" 6 "College/university completed" 7 "Post graduate degree" 88 Refused	
Smoke	1 Yes	2 No

To do so, click in the box next to the variable name – the field will now be highlighted (in blue). Press F9 or click on the small + icon next to the "Value Label" field. For each field, enter the corresponding codes for each value that will be entered then click on "Accept and Close".



Range Legal

To avoid improbable values being entered in certain fields, use the Range Legal function. Place your cursor in the 'age' field – it will now be highlighted in blue. In the Range, Legal box, enter the values that you will accept – in this case, enter 0-100. This ensures that no value above 100 will be accepted.

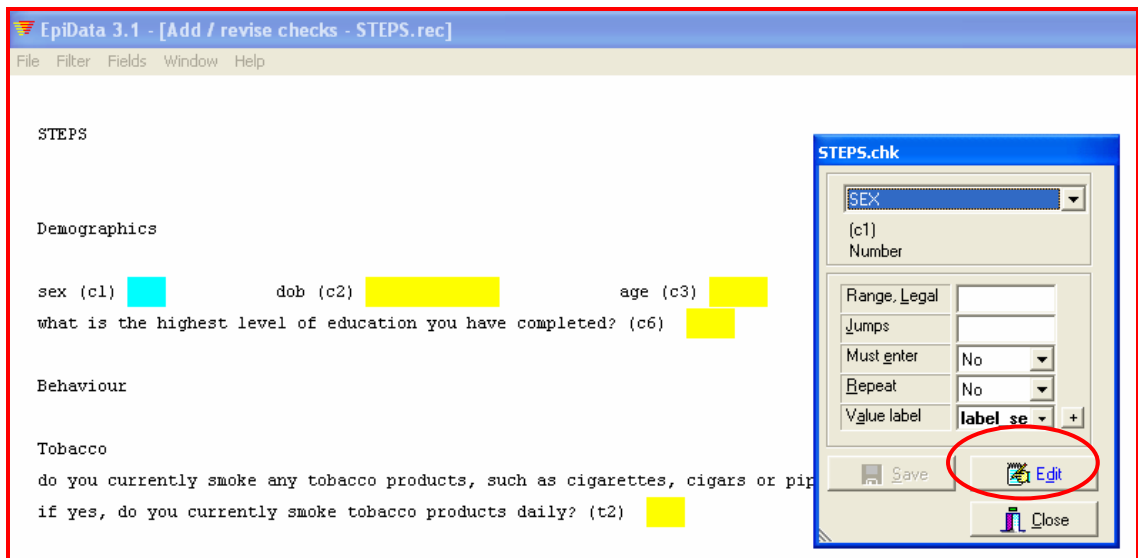


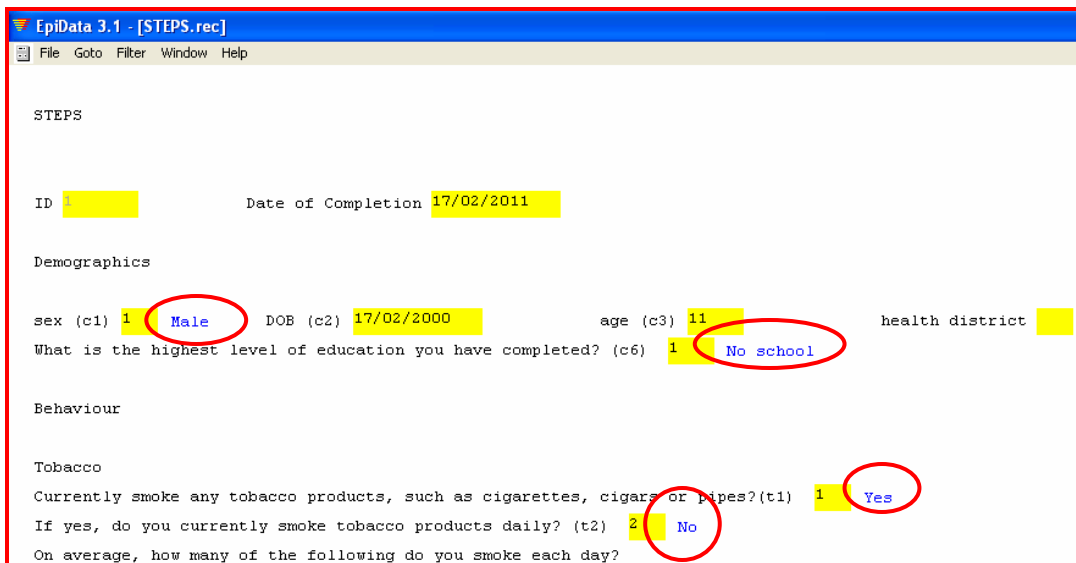
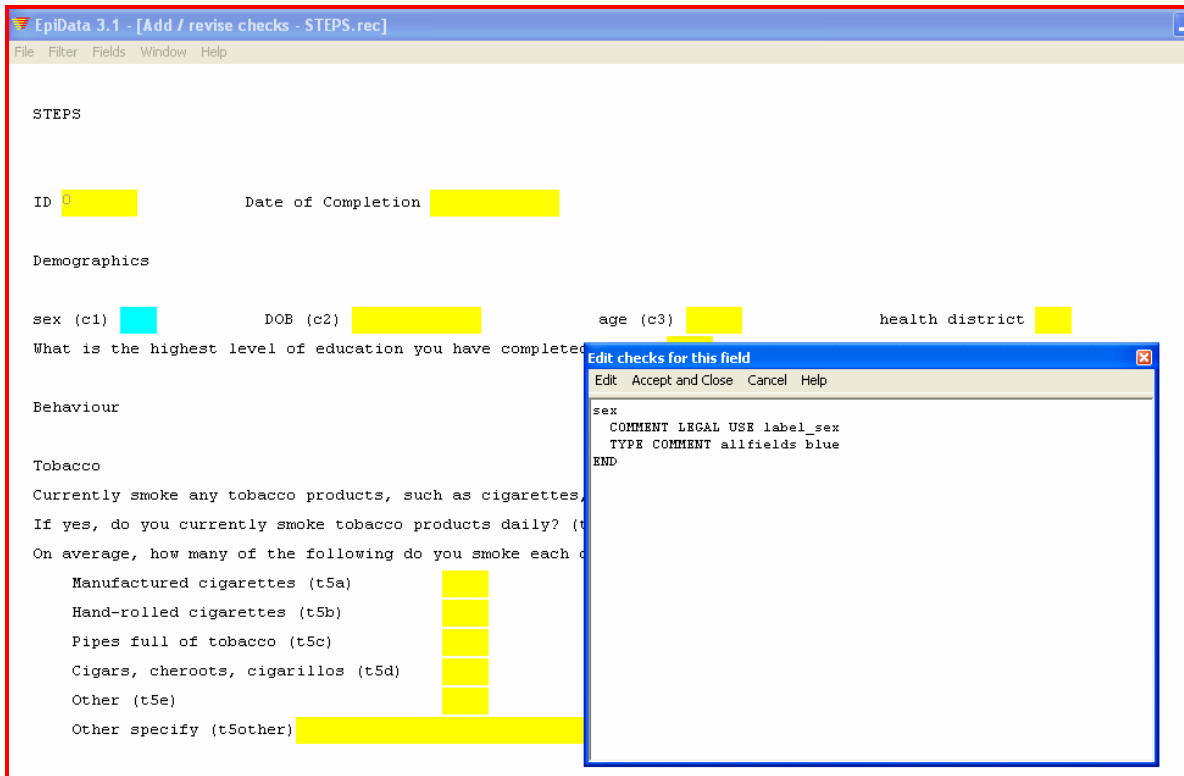
Task 10: Close the check portion described in task 8, and then select "4. Enter Data" to open the STEPS.rec (data file) again. Now when you try to enter improbable data (i.e., age = 500), you will notice that the 'checks' you created in task 9 restrict what can be entered.

HINT: pressing + in a given field opens a small box that specifies the allowable values (e.g., 1 – Yes, 2 - No).

Task 11: Adding extended check file definitions.

Select **"3. Checks"** and reopen the STEPS.rec file, highlight the field Sex, click "Edit" and add the command "type comment allfields blue" (without quotes) to the field as shown below. When a numerical value is entered in the field and the cursor is moved to another field, then the label associated with the value will appear beside the code.

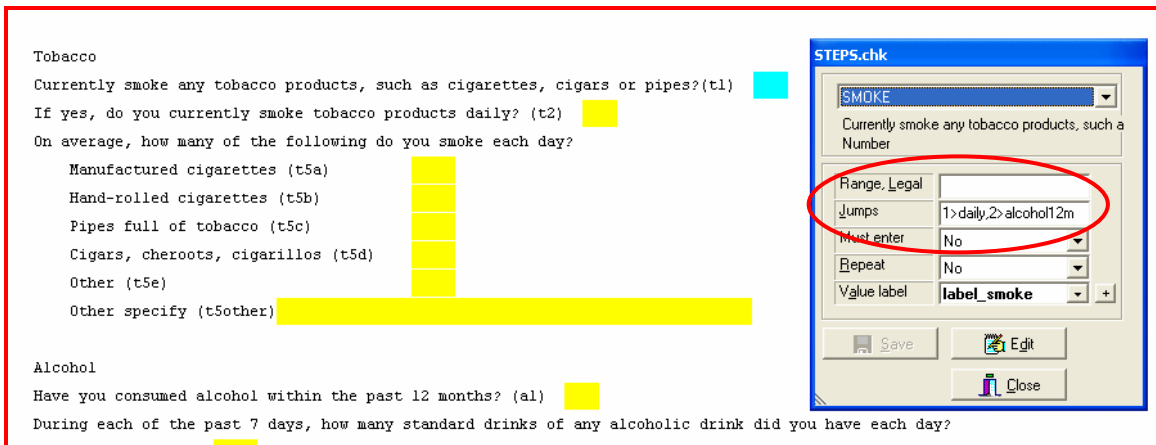




Jumps (Skips)

By default, data entry progresses from the first field to the last field in the form, one field at a time. Jumps are used to move to different fields. They are most useful for conditional jumps, for example "if response to question 8 is No, go to question 10".

If a respondent indicates that they do not currently smoke, you would want to skip the next question, which asks if they smoke daily. Make the currently smoke variable 'active' by clicking in it (the entry section will be highlighted blue). To tell that a jump is needed in case the value of the field SMOKE is 2 (No), type the following in the Jumps field: 1>daily,2>alcohol12m, meaning, if 1 is entered, go to the next question "If yes, do you currently smoke tobacco products daily?" and if 2 is entered, go to the question "Have you consumed alcohol within the past 12 months?"



Jump to a New Record

If, depending on the answer to a question, all other questions that follow are irrelevant and you want to move on to a new record, make the variable 'active' by clicking in it, click on the 'Edit' button and then add the command to the Check file. If, for example, a woman specifies that she is pregnant, the 'height' and 'weight' variables that follow (used to calculate BMI) should not be asked. The following can be entered into the Check file for the pregnant variable:

Jumps

- 1 Write If pregnant = Yes, write record to disk
- 2 Height If pregnant = No, go to the height question

End

Instead of specifying a field name, WRITE causes the 'Write record to disk?' dialog box to appear. After clicking Yes the next or a new record is loaded.

EpiData 3.1 - [Add / revise checks - STEPS.rec]

File Filter Fields Window Help

Tobacco

Currently smoke any tobacco products, such as cigarettes, cigars or pipes?(t1)

If yes, do you currently smoke tobacco products daily? (t2)

On average, how many of the following do you smoke each day?

Manufactured cigarettes (t5a)

Hand-rolled cigarettes (t5b)

Pipes full of tobacco (t5c)

Cigars, cheroots, cigarillos (t5d)

Other (t5e)

Other specify (t5other)

Alcohol

Have you consumed alcohol w...

During each of the past 7 da...

Monday (a5a)

Tuesday (a5b)

Wednesday (a5c)

Thursday (a5d)

Friday (a5e)

Saturday (a5f)

Sunday (a5g)

Physical Measurements

Are you pregnant? (m5)

Height in centimeters (m3)

Weight in kilograms (m4)

Edit checks for this field

Edit Accept and Close Cancel Help

```

pregnant
COMMENT LEGAL USE label_smoke
JUMPS
  1 WRITE
  2 height
END

```

PS.chk

PREGNANT

Are you pregnant? (m5)

Number

Range_Legal

Jumps 1>WRITE,2>

Must enter No

Repeat No

Value label label_sm +

Save Edit Close

Friday (a5e)

Saturday (a5f)

Sunday(a5g)

Physical Measurements

Are you pregnant? (m5)

Height in centimeters (m3)

Weight in kilograms (m4)

Confirmation

Save record to disk?

Yes No

Helpful Hints

Value Labels - If a label contains spaces, it must be surrounded by double quotes.

Examples:

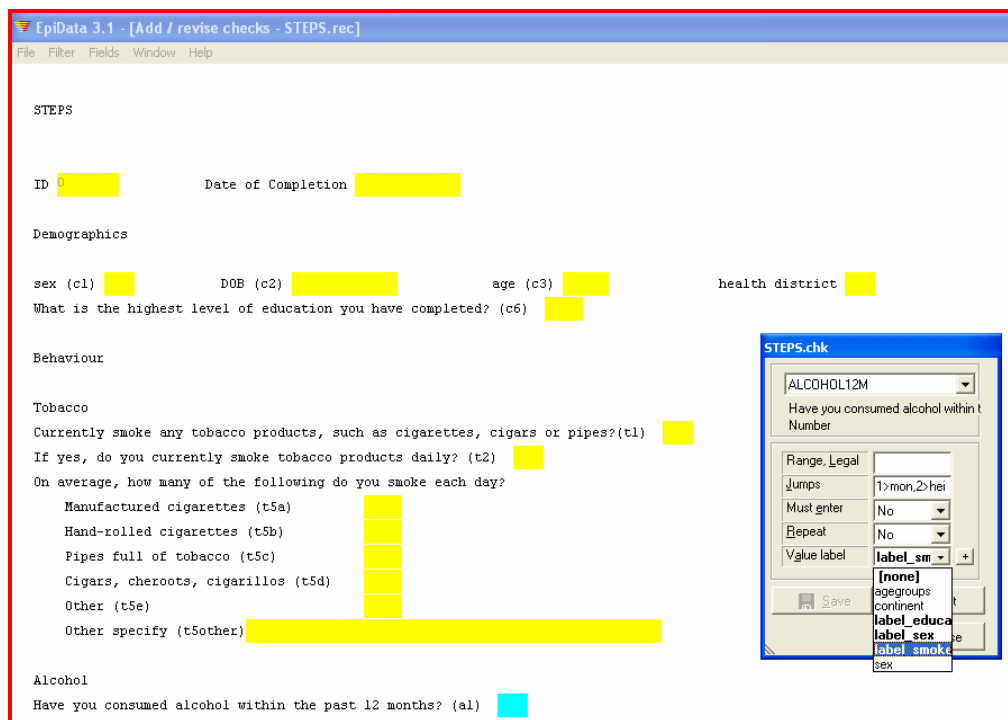
LABEL label_education

- 1 "No schooling"
- 2 "Less than primary"
- 3 "Primary completed"
- 4 "Secondary completed"
- 5 "High school completed"
- 6 "College/university completed"
- 7 "Post graduate degree"

END

Manually Editing a Check File - EpiData provides a text editor for editing check files, similar to the questionnaire file editor. To open a file in this editor, use File → Open, select "EpiData check file" from the "Files of type" drop-down list, and select the check file. You can now edit the whole check file all at once rather than having to select fields in the interactive dialogue.

Assign an existing label to a field - Click on the down arrow in the Value Label drop-down list and select the relevant label. Several fields can share the same value label, which only needs to be defined once. In this case, the label for consuming alcohol will use the label already defined for currently smoking.



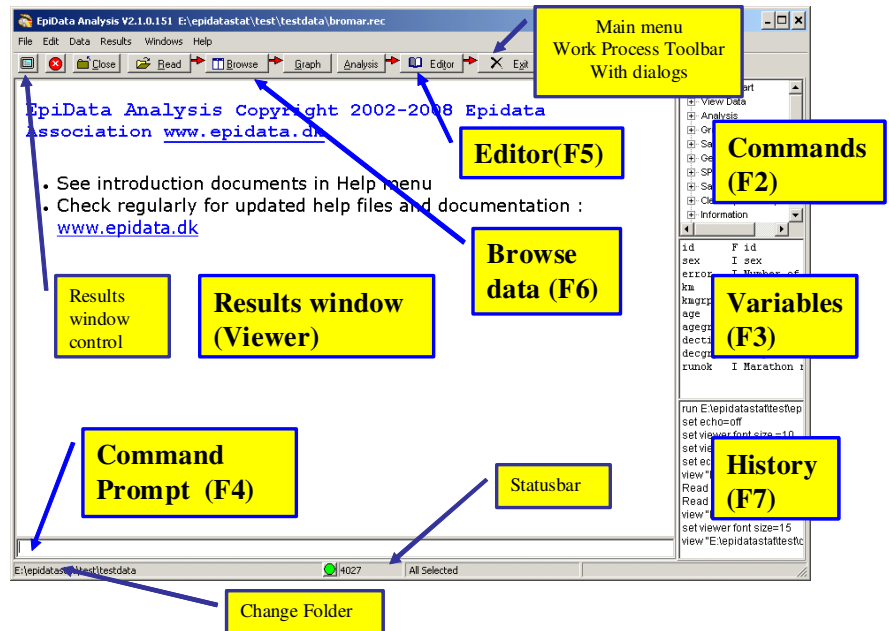
PART 3: EpiData Analysis

EpiData Analysis is used to analyze EpiData *.rec files, dBase *.dbf III files, and text *.csv files.

The EpiData Analysis screen is easy to navigate and functionality is accessible via the use of shortcut keys.

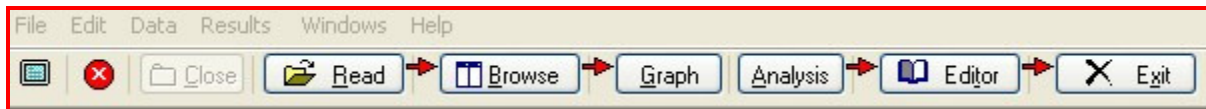
For the purposes of this field guide, areas of importance are:

- **The command prompt (F4) area located at the bottom of the screen**
- **The editor (F5)**
- **The results window.**
- **The dialogs in the work process toolbar.**



But also take some time to explore additional functionality available via other short cut keys.

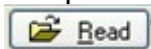
When you open EpiData Analysis, you see the following menu:



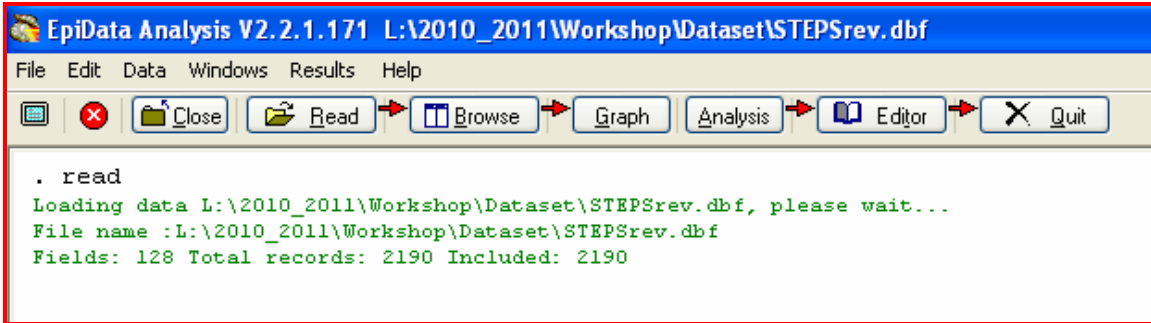
Opening and preparing the dataset

Note: when working with EpiData Analysis, the file that will be used is titled "STEPSrev.dbf". All files are available on the Association of Public Health Epidemiologists in Ontario's website at [http://www.apheo.ca/index.php?pid=47#2011 Workshop](http://www.apheo.ca/index.php?pid=47#2011%20Workshop).

To open a data file, click on **"Read"**:

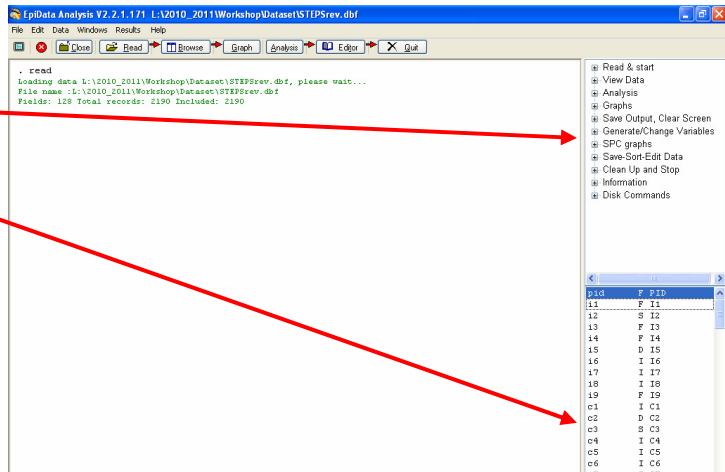


Navigate to the location where you saved your file (in your project folder) and select it.

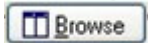


EpiData will provide some information about the data: name, number of records number of fields, etc. In this case you have 2,190 records in the dataset.

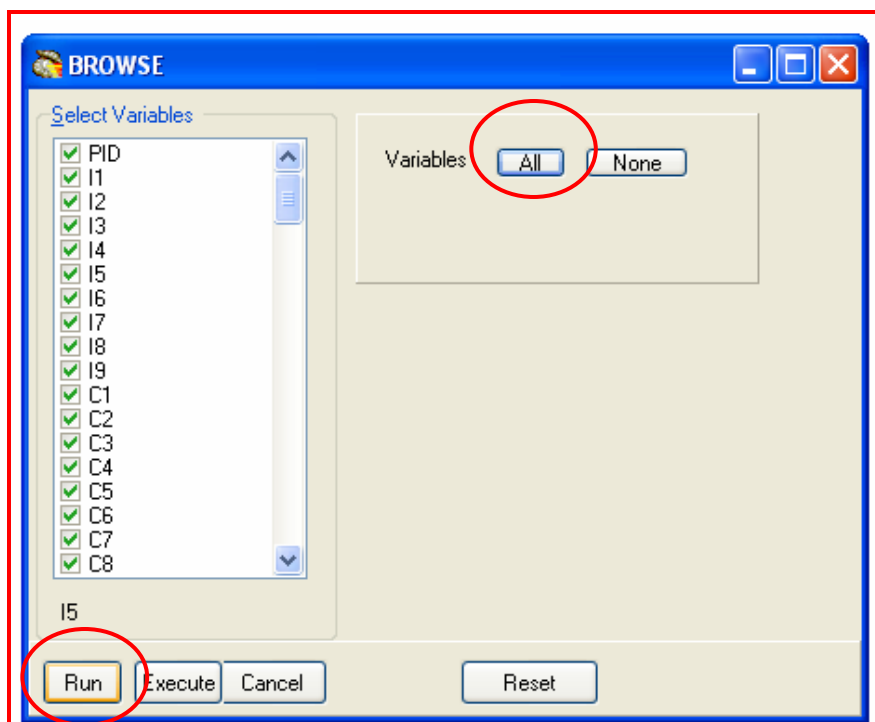
Press F2 to reveal all commands as well as F3 to display the file's variables:



Now take a look at the data by selecting **"Browse"**:



Click on the **"All"** button to select all available variables then click on the **"Run"** button.



This will produce a list:

	PID	I1	I2	I3	I4	I5	I6	I7	I8	I9	C1	C2	C3	C4	C5	C6
1	1	5	.	691	10	05/10/2006	1	1	2	9	2	24/01/1982	.	7	1	3
2	2	13	.	13	19	04/02/2006	1	1	2	12	2	11/03/1972	.	9	1	3
3	3	6	.	101	7	20/05/2006	1	1	2	11	2	28/11/1975	.	9	1	4
4	4	5	.	22	14	21/05/2006	1	1	2	10	2	29/01/1980	.	99	1	3
5	5	22	.	25	25	04/02/2006	1	1	2	17	2	15/06/1981	.	15	1	6
6	6	22	.	375	26	16/04/2006	1	1	2	11	2	27/08/1977	.	9	1	3
7	7	10	.	172	22	14/05/2006	1	1	2	11	2	05/04/1975	32	5	1	2
8	8	13	.	172	17	04/02/2006	1	1	2	8	2	19/11/1977	.	0	1	4
9	9	22	.	172	26	04/04/2006	1	1	2	19	2	06/06/1976	.	12	1	4
10	10	17	.	171	9	26/05/2006	1	1	2	14	2	09/03/1973	.	9	1	5
11	11	10	.	13	23	17/05/2006	1	1	2	16	2	14/04/1978	28	13	1	6
12	12	5	.	483	14	21/05/2006	1	9	9	11	2	11/02/1978	.	18	1	6
13	13	5	.	171	14	21/05/2006	1	1	2	100	2	15/02/1974	.	9	1	4
14	14	10	.	101	21	16/05/2006	1	1	2	17	2	03/09/1972	35	7	1	3
15	15	13	.	22	19	04/02/2006	1	1	2	9	2	07/10/1978	.	9	1	3
16	16	5	.	42	10	28/05/2006	1	1	2	9	2	20/05/1976	.	9	1	6
17	17	22	.	13	26	04/04/2006	1	1	2	17	2	07/10/1977	.	9	1	3
18	18	16	.	13	7	21/05/2006	1	1	2	12	2	24/06/1981	.	12	1	4
19	19	10	.	172	21	14/05/2006	1	1	2	16	2	06/08/1980	25	5	1	2
20	20	6	.	436	7	14/05/2006	1	1	2	9	2	18/02/1982	.	12	1	4

Fine tuning the dataset for analysis (data cleaning)

Before analysis can be done, the data must be cleaned and documented if it hasn't already been done during entry or through the creation of a check (.chk) file. This includes finding out whether all variables are valid, how many observations can be part of the analysis etc. Here, a few examples are shown; more aspects will be covered in part 3 of this document.

1. Changing Variable and Value Labels

Sometimes when you start analyzing your dataset, you realize that the names of variables or values are not all that meaningful. In these instances, it is important to spend some time preparing the dataset, but it is always good practice to define labels at three levels:

(1) At whole file level (labeldata)

- To add labels to the whole data file: LABELDATA

In the command prompt area (F4) write:

LABELDATA "STEPwise approach to chronic disease risk factor surveillance (STEPS)"
and press <enter>.

(2) At variable level (label)

- To add (or change) variable names, use the command: LABEL

In the command prompt area (F4), write:

LABEL t1 "Currently smoke" and press <enter>.

(3) At value or category level ((labelvalue)

- To add labels to the values of t1 use the command: LABELVALUE

In the command prompt area (F4) write:

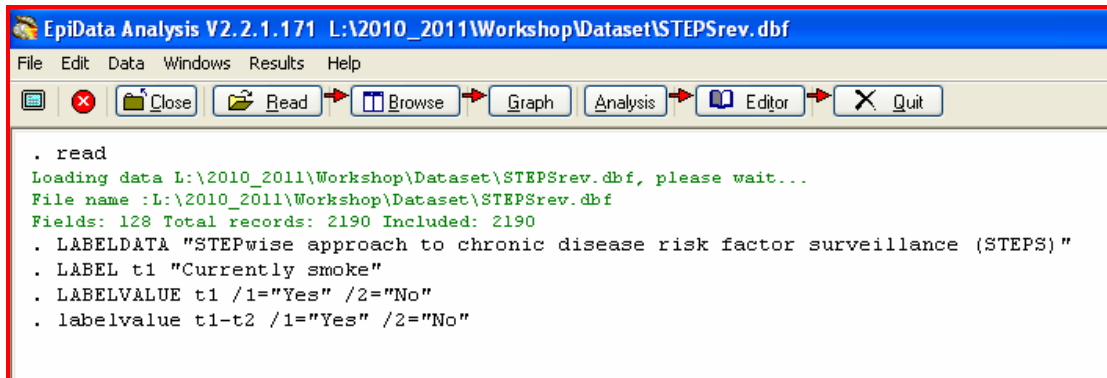
LABELVALUE t1 /1="Yes"/2="No" and press <enter>.

There is an easy way to add value labels for many variables that are in sequential order:

labelvalue t1-t2 /1="Yes"/2="No" and press <enter>.

Notice: no space between the variable names only one dash (-).

Once all commands have been run, the commands appear in the results window.



The screenshot shows the EpiData Analysis V2.2.1.171 interface. The title bar indicates the file path: L:\2010_2011\Workshop\Dataset\STEPSrev.dbf. The menu bar includes File, Edit, Data, Windows, Results, and Help. The toolbar contains icons for Close, Read, Browse, Graph, Analysis, Editor, and Quit. The command prompt area displays the following text:

```
. read
Loading data L:\2010_2011\Workshop\Dataset\STEPSrev.dbf, please wait...
File name :L:\2010_2011\Workshop\Dataset\STEPSrev.dbf
Fields: 128 Total records: 2190 Included: 2190
. LABELDATA "STEPwise approach to chronic disease risk factor surveillance (STEPS)"
. LABEL t1 "Currently smoke"
. LABELVALUE t1 /1="Yes" /2="No"
. labelvalue t1-t2 /1="Yes" /2="No"
```

2. Creating New Variables and Adding Conditional Values

Age

During the data entry process, the age of each respondent was collected in two ways. Each was asked their birth date as well as their age.

First, assess the age (c3) variable by running a frequency:

freq c3

Of the 2,190 records in the dataset, only 580 include an age. In order to deal with the missing data, create a new age variable based on 2 date variables: date of birth (variable c2) and date of completion of the survey (variable i5). Use the "DEFINE" and "LET" commands.

In the command prompt (F4) area, write:

define newage ###

let newage=trunc((i5-c2)/365.25)

Check the new variable by running a frequency:

freq newage

Now, 1,946 records contain an age.

Now that you want to analyze the data, group ages (variable newage) into categories by creating a new variable called "agegrp".

You are going to use the "DEFINE" and "RECODE" commands.

In the command prompt (F4) area, write:

* create your age group variable as an integer:

```
define agegrp #
```

```
recode newage to agegrp 0-9=1 10-19=2 20-44=3 45-64=4 65-hi=5
```

The command "recode" also adds the value label in the intervals indicated. You will see a new variable called agegrp in the variable list (press F3). Note: Descriptive comments can be added to your code using '*' at the start of the sentence.

Descriptive Epidemiology

Describing the cases in terms of person, place and time is good epidemiological practice and can help you develop hypotheses about exposures and outcomes. Let's start with **person**:

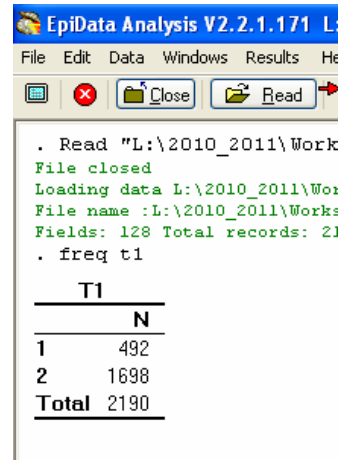
In order to know how many current smokers you have, you can do a frequency distribution of this variable (t1).

In the command prompt area (F4), write:

```
freq t1
```

You can see that there are 492 current smokers and 1698 non-smokers.

- (Where 1=smoker and 2=non-smoker)



```
. Read "L:\2010_2011\Work
File closed
Loading data L:\2010_2011\Wor
File name :L:\2010_2011\Works
Fields: 128 Total records: 21
. freq t1
```

T1	
	N
1	492
2	1698
Total	2190

Now, analyze the demographic data to determine the age and sex structure of the population at risk. Use the new variable you just created above – agegrp. Once again, at the command prompt (F4), type:

```
tables c1 agegrp /c
```

This will produce a cross tabulation of sex by age group along with column percents. Your table should look like this:

WHO age groups	Sex		% Total	%
	Male	Female		
20 to 44 years	724 {83.4}	1006 {93.3}	1730	{88.9}
45 to 64 years	144 {16.6}	72 {6.7}	216	{11.1}
Total	868 {100.0}	1078 {100.0}	1946	
Percents: {Col}				

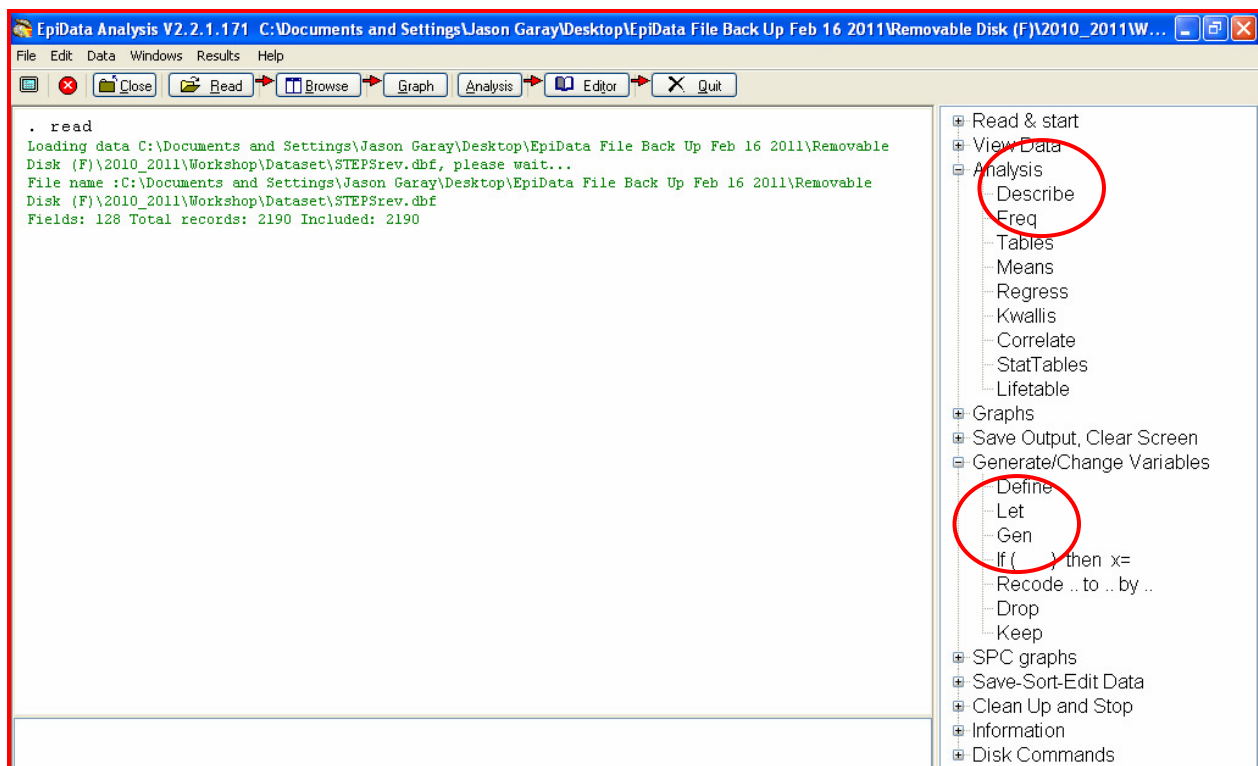
You can see that the majority of the population is between the ages of 20 and 44 years (89%). You can add row percents as well by adding /r to the end of the command line. If you did this, you would also see that there are more females (58%) than males (42%) in the population.

Body Mass Index (BMI)

Respondents were asked for their height (variable m3) and weight (variable m4), which allows BMI to be calculated. Press F2 to show all commands – under Generate/Change Variables, select 'Gen'. 'Gen' creates a new numeric variable based on an expression.

The metric imperial BMI formula is: $BMI = \text{weight in kg} / \text{height in m}^2$. Since the height was collected in centimeters, our formula will need to reflect the change from centimeters to meters. Also, we will need to exclude pregnant women (variable m5) from our calculation. Type the remainder of the formula in the command prompt area:

```
gen bmi = (m4/((m3/100)*m3/100))
if (m3=999) then bmi = .
if (m4 =999) then bmi = .
* bmi not correct for pregnant women:
if (m5 = 1) then bmi = .
```



Next, use the 'Describe' command (under the Analysis section) to view summary information about the 3 continuous variables – height, weight and bmi by typing the following formula:

Describe m3 m4 bmi

```

EpiData Analysis V2.2.1.171 L:\2010_2011\Workshop\Dataset\STEPSrev.dbf
File Edit Data Windows Results Help
Close Read Browse Graph Analysis Editor Quit

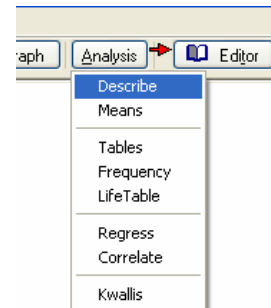
. read
Loading data L:\2010_2011\Workshop\Dataset\STEPSrev.dbf, please wait...
File name :L:\2010_2011\Workshop\Dataset\STEPSrev.dbf
Fields: 128 Total records: 2190 Included: 2190
. Run "C:\DOCUME~1\Brenda\LOCALS~1\Temp\TMP47F8.Tmp"
. gen bmi = (m4 / ((m3/100) *m3/100))
Var Name bmi of type Float and length: 10 decimals 4
. if (m3=999) then bmi = .
. bmi = .
. if (m4 =999) then bmi = .
. bmi = .
* bmi not correct for pregnant women:
. if (m5 = 1) then bmi = .
. bmi = .
. describe m3 m4 bmi

```

Variable	N=2190	Sum	Mean	(95% cfi)	Min	p5	p10	p25	Median	p75	p90	p95	Max	
M3	2190	405035.0	184.95	180.15	189.74	59.00	154.00	158.00	163.00	170.00	177.00	182.00	184.00	999.00
M4	2190	177661.0	81.12	76.25	85.99	34.00	47.00	51.00	57.00	65.00	75.00	87.00	97.00	999.00
bmi	2088	49493.0	23.70	23.20	24.21	12.21	16.84	17.78	19.61	22.21	25.83	29.99	32.46	376.48

For each variable, the total number of records used in the calculation are given (N) as well as the sum, mean age, 95% confidence interval, the minimum value, percentiles (5, 10, 25, 50=Median, 75, 90 and 95) and the maximum value.

Note: the 'Describe' command can also be accessed from the main menu toolbar



The extremely high BMI could be due to errors – this should be checked. List the observations with very high values by typing the following at the command prompt:

```
list m3 m4 c1 m5 bmi if bmi > 100
```

Graph

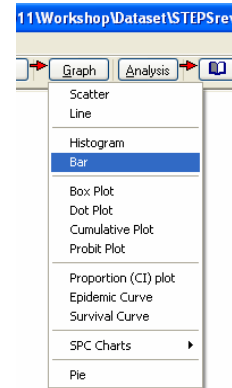
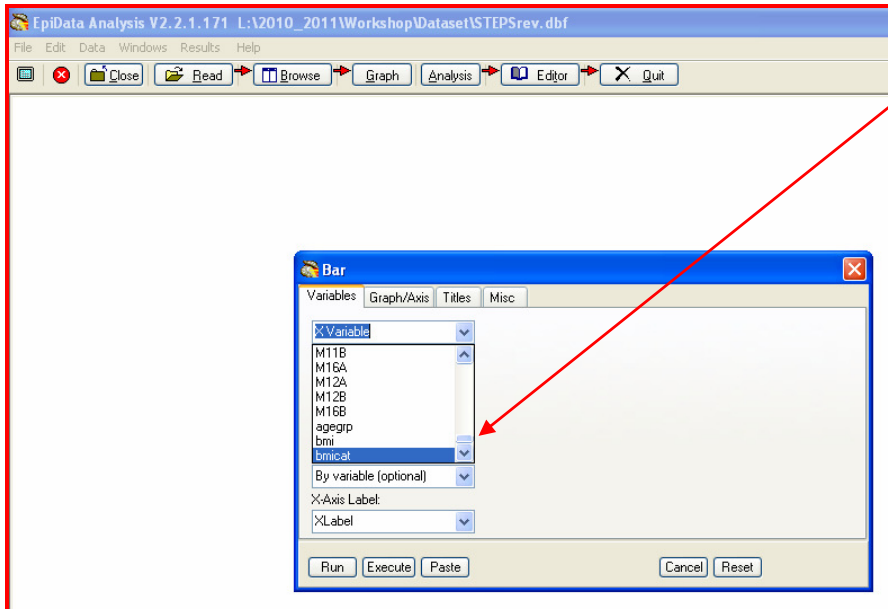
Creating a graph of BMI Distribution is relatively easy and straightforward. Here it is shown with a categorized version of BMI. After completion of the graphs based on the categorized values you should look at the distribution with a histogram and a normal probability plot. Both of these are shown in the toolbar under "Graph".

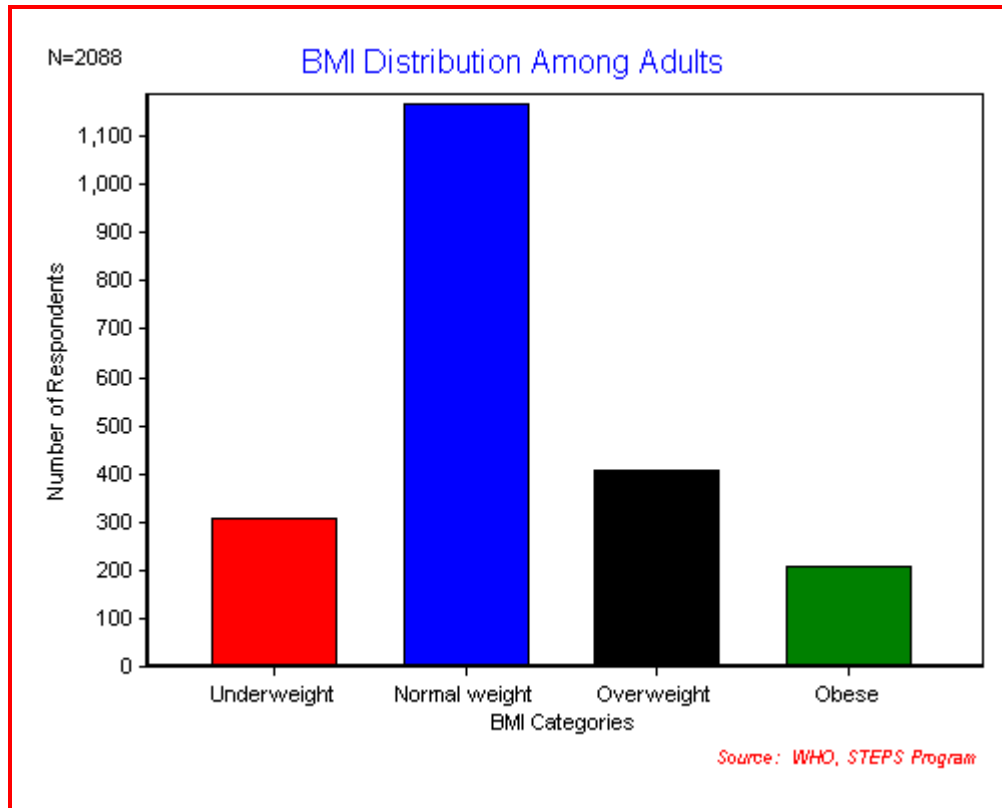
First, create a new variable 'bmicat' in order to categorize the continuous variable 'BMI'. This is accomplished by using the commands 'define' and 'recode'. In the command prompt area, type:

```
define bmicat #
recode bmi to bmicat lo-18.4999=1 18.50-24.9999=2 25.0-29.9999=3 30.0-hi=4
label bmicat "BMI Categories"
```

labelvalue bmicat /1="Underweight" /2="Normal weight" /3="Overweight" /4="Obese"

To create a simple bar chart, select 'Bar' from the Graph icon. Choose the variable that you want graphed (bmicat). Work your way through the tabs to produce your graph.





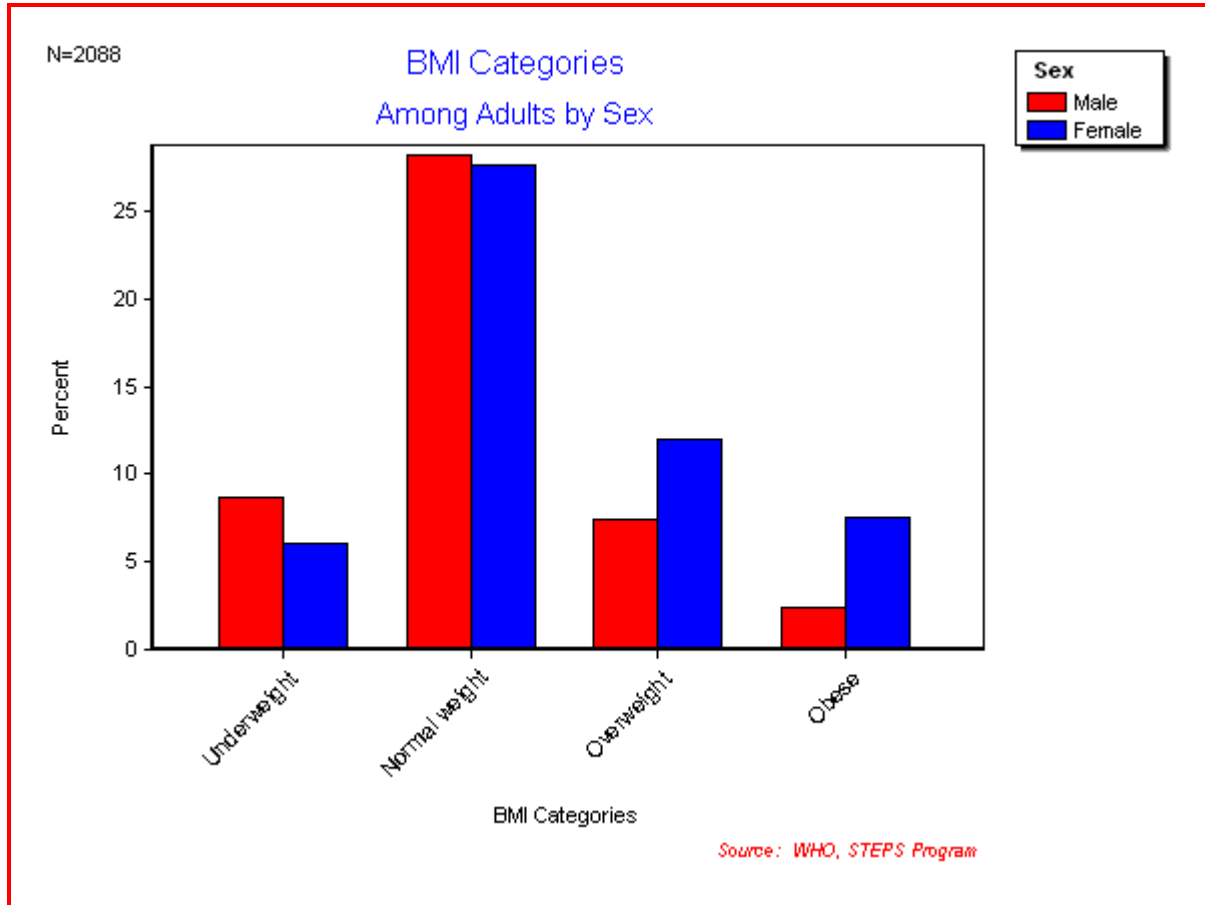
The command that produced this graph:

```
BAR bmicat /Frame /N /SizeX=500 /SizeY=400 /Ti="BMI Distribution Among Adults"
/Fn="Source: WHO, STEPS Program" /Xtext="BMI Categories" /Ytext="Number of
Respondents" /Xlabel=bmicat
```

Where:

/Frame =	Show a frame around the chart
/N =	Show size of the population of interest
/SizeX and /Size Y =	Height and width of the chart
/Ti =	Title
/Fn =	Footnote
/Xtext =	X axis title
/Ytext =	Y axis title
/Xlabel =	X axis category titles

Next is a clustered bar chart, comparing weight categories of males and females.



The command required to produce this graph:

```
BAR bmicat /by=c1 /PCT /Frame /Legend /X45 /SizeX=600 /SizeY=450 /sub="Among
Adults by Sex" /Fn="Source: WHO, STEPS Program" /Xtext="BMI Categories"
/Ytext="Percent" /Xlabel=bmicat
```

Where:

/by =	The 'by' variable, in this case, sex
/PCT =	Show percent
/Frame =	Show a frame around the chart
/N =	Show size of the population of interest
/X45 =	x-axis labels at 45 degree angle
/Legend =	Show the legend
/SizeX and /Size Y =	Height and width of the chart
/sub =	Subtitle
/Fn =	Footnote
/Xtext =	X axis title
/Ytext =	Y axis title
/Xlabel =	X axis category titles

Analytical Epidemiology

Test of Association

Now, let's run a chi square test to test for associations between two variables. Calculate the chi square statistic by running a cross tabulation between the BMI categories and sex (variable c1). Type:

`tab bmicat c1 /R /T` Where /R= row percentages and /T = Chi square test

Sex	BMI Categories				% Total	%
	Underweight	% Normal weight	% Overweight	% Obese		
Male	181 (18.6)	589 (60.4)	155 (15.9)	50 (5.1)	975 (100.0)	
Female	127 (11.4)	578 (51.9)	250 (22.5)	158 (14.2)	1113 (100.0)	
Total	308 (14.8)	1167 (55.9)	405 (19.4)	208 (10.0)	2088	

Percents: (Row)
 Chi²= 79.157 df(3) p= 0.0000 ←

The Chi square statistic is significant therefore; we can reject the hypothesis of independence between sex and BMI.

Strength of Association

A popular measure of the strength of association between two variables is odds ratio. The odds ratio is used to assess the risk of a particular outcome (or disease) if a certain factor (or exposure) is present. The odds ratio is a relative measure of risk, telling us how much more likely it is that someone who is exposed to the factor under study will develop the outcome as compared to someone who is not exposed.

We'll look at the risk of developing hypertension (outcome) among those that are obese (exposure). First, assess both variables by running frequencies:

`freq h2 bmicat`

H2	N
1	194
2	1926
9	70
Total	2190

BMI Categories	
	N
Underweight	308
Normal weight	1167
Overweight	405
Obese	208
Total	2088

The hypertension variable includes 70 missing responses (coded 9). Also the bmicat variable includes all other weight categories, when we're only interested in those who are obese. The next step would be to create two dummy variables so that we could calculate the odds ratio from a 2 x 2 table.

Type the following in the command prompt window:

```

define hyp #
hyp=1 if (h2 = 1)
hyp=0 if (h2 = 2)
label hyp "Hypertension"
labelvalue hyp /1="Yes"
labelvalue hyp /0="No"
freq hyp /C /CI

```

This will remove the missing responses. As well, 95% confidence intervals (/CI) will be calculated.

Repeat similar steps in order to create a new 'obese' variable:

```

define obese #
recode bmicat to obese 1-3=0 4=1
label obese "Obese (BMI >30)"
labelvalue obese /0="Not Obese"
labelvalue obese /1="Obese"
freq obese /C /CI

```

Now, the variables are amenable to the calculation of an odds ratio. Type the following command:

tab hyp obese /O /T Where /O = odds ratio, /T = chi square

Hypertension			
Obese (BMI >30)	Yes	No	Total
Obese	32	171	203
Not Obese	152	1670	1822
Total	184	1841	2025

Exposure: Obese (BMI >30) = Obese
Outcome: Hypertension = Yes

Chi² = 12.177 df(1) p = 0.0005
Odds Ratio = 2.06 (95% CI: 1.32-3.14)

This odds ratio is greater than one, and the 95% confidence interval for the odds ratio is (1.32, 3.14) so does not include an odds ratio of one. Thus we would conclude that those who are obese are at significantly increased risk of hypertension.

Stratified Analysis using Cplot

Stratified analysis involves examining the exposure-disease association within different categories of a third factor. It is an effective method for looking at the effects of two different exposures on disease.

A plot of proportions of cases for subgroups defined by other variables can be an effective way of finding high or low risk groups quite quickly. Define a search pattern for subgroups before you do any analysis and be careful with statistical testing, since a broad search and test strategy will modify any p-value you should regard as significant.

You want in this example to look for variation of hypertension in subgroups defined by other variables. Let's use obesity and sex. Type the following command:

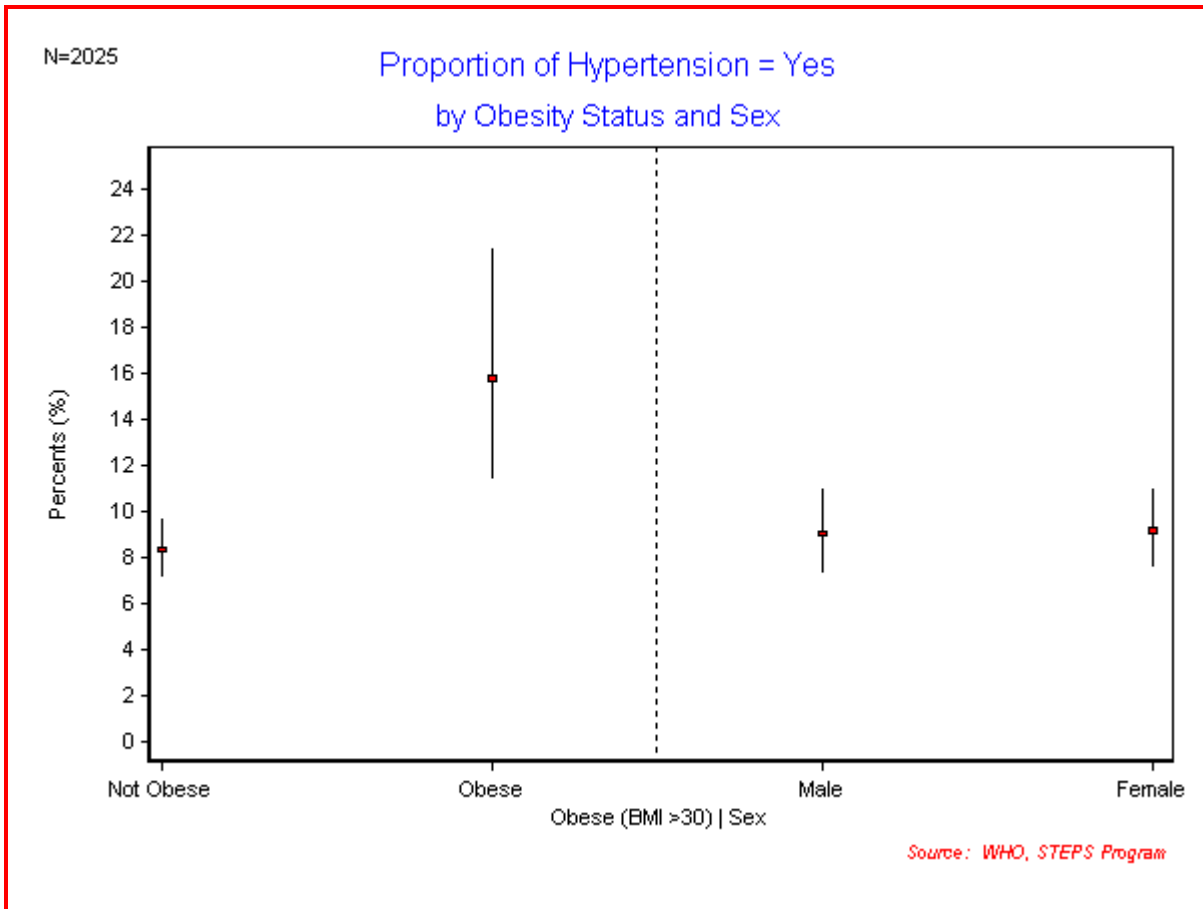
```
cplot hyp obese c1 /NM /NOCI /NOTOT /N /Frame /ymin=0 /ymax=25 /SizeX=600 /SizeY=450 /sub="by Obesity Status and Sex"/Fn="Source: WHO, STEPS Program"
```

Where:

/NM = Records with missing in any variable are excluded
 /NOCI = Hide confidence intervals in tables and graph (e.g. if graphs overlap for groups)
 /NOTOT = Remove crude (total) estimate from the graph
 /N = Show size of the population of interest
 /Frame = Show a frame around the chart
 /ymin and /ymax = Sets axis scale
 /SizeX and /SizeY= Height and width of the chart
 /sub = Subtitle
 /Fn = Footnote

Crude: Proportion of Hypertension = Yes among all.						
variable	stratum	Total N	n_{Hypertension=Yes}	%	(95% CI)	
Obese (BMI >30)	Not Obese	1822	152	8.3	(7.2-9.7)	
	Obese	203	32	15.8	(11.4-21.4)	
Sex	Male	943	85	9.0	(7.3-11.0)	
	Female	1082	99	9.1	(7.6-11.0)	

Crude: Proportion of Hypertension = Yes among all.



You'll notice that the proportion of hypertension is rather high among obese persons. There doesn't appear to be a difference in hypertension among males and females.

This could be investigated further by using crosstables and stratifying by sex:
tab hyp obese c1 /O /T

Hypertension			
Obese (BMI >30)	Yes	No	Total
Obese	32	171	203
Not Obese	152	1670	1822
Total	184	1841	2025

Unstratified table
 Exposure: Obese (BMI >30) = Obese
 Outcome: Hypertension = Yes

Chi²= 12.177 df(1) p= 0.0005
 Odds Ratio = 2.06 (95% CI: 1.32-3.14)

This is the unstratified table which shows a significant difference in hypertension among obese vs. not obese persons.

Sex: Male Hypertension			
Obese (BMI >30)	Yes	No	Total
Obese	15	35	50
Not Obese	70	823	893
Total	85	858	943

Exposure: Obese (BMI >30) = Obese
 Outcome: Hypertension = Yes

Chi²= 28.354 df(1) p= 0.0000
 Cells expected < 5: 1 (25 pct.)
 Odds Ratio = 5.02 (95% CI: 2.43-9.99)

The odds ratio for hypertension among obese males (OR=5.02, 95% CI 2.43-9.99) is significant.

Sex: Female Hypertension			
Obese (BMI >30)	Yes	No	Total
Obese	17	136	153
Not Obese	82	847	929
Total	99	983	1082

Exposure: Obese (BMI >30) = Obese
 Outcome: Hypertension = Yes

Chi²= 0.825 df(1) p= 0.3638
 Odds Ratio = 1.29 (95% CI: 0.70-2.28)

The odds ratio for hypertension among obese females (OR=1.29, 95% CI=0.70-2.28) is not statistically significant. There appears to be some effect modification.

The final table shows the odds ratio adjusted for sex:

Hypertension by Obese (BMI >30) adjusted for Sex							
	N = 2025	N	OR	[95% CI]	Chi²	Df	p
Crude		2025	2.06	(1.32-3.14)	12.177	1	0.000
Adjusted		2025	2.09	(1.33-3.21)	12.29	1	0.000
Sex: Male		943	5.02	(2.43-9.99)	28.354*	1	0.000
Sex: Female		1082	1.29	(0.70-2.28)	0.825	1	0.364

Summary Estimates
 Total 2 strata. 2 informative & 0 non-informative.
 Exposure: Obese (BMI >30) = Obese
 Outcome: Hypertension = Yes
 *: Small Expected Numbers, use P_{exact} /ex

Helpful Hints

Frequencies

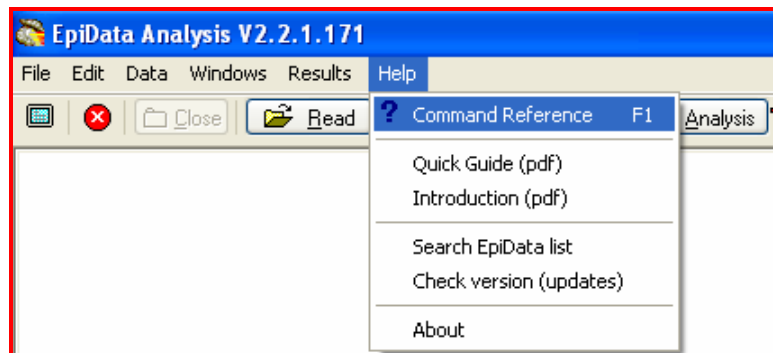
To run frequencies on all variables in the dataset, use the command *freq **

Tables

To place row or column percentages into separate columns, use the command *tables bmicat c1 /c /r /pct*

Command Assistance

For a listing of all EpiData Analysis commands and functions, access the reference guide via the help menu



PART 4: Working with Programs

When you are analyzing a dataset it's good practice to save your commands so you can use them later so there's no need to 'recreate the wheel'. This is especially useful if you are working with a database in a routine system like a surveillance system.

You should use programs if you want to recode variables, make calculations using other variables or any other kind of manipulation. This method does not make permanent changes to your original dataset. EpiData Analysis provides this functionality – it's similar to the creation of syntax files in SPSS.

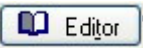
Saving, recalling and executing programs:

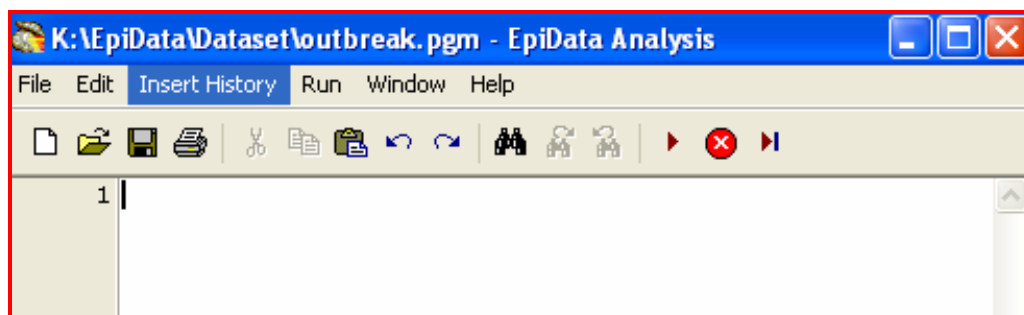
One way to make a program file is to enter commands interactively at the command prompt (F4). If you wish to save the commands you've already entered, use the SAVEPGM command and the name of a file, for example:

```
SAVEPGM STEPS.PGM
```

All the commands used during the working session will be placed in the file, which can then be edited later to remove unwanted commands or add new ones.

You can also use the EpiData Analysis Editor.

Click on "**Editor**"  and then select "**Insert History**" to populate the screen with all previously used commands. Select "**File**" then "**Save As**" to save the information as a program (.pgm).



Once you have saved your commands in a .pgm file, you can open it whenever you want using the EpiData Editor (FILE-->Open)

To execute a program, you have different alternatives:

- i) In the EpiData Editor you can run all (**F9**)

ii) Or in the EpiData Editor you can select the group of lines of the program that you want executed and then run the selected lines only (**F8**). With no selection only the current line is executed.

iii) Edit the program in the editor, save, and from the command prompt issue the "RUN" command and find the .pgm file you saved.

Questions/For More Information

Jens Lauritsen, EpiData Association
www.epidata.dk

Association of Public Health Epidemiologists in Ontario
EpiData Expert Panel
<http://www.apheo.ca/index.php?pid=47>

Appendix 1 – STEPS Questionnaire

STEPS

ID <IDNUM> {Date} of {Compl}etion <dd/mm/yyyy>

Demographics

{sex} (c1) # {DOB} (c2) <dd/mm/yyyy> {age} (c3) ###

health {district} #

What is the highest level of {education} you have completed? (c6) ##

Behaviour

Tobacco

Currently {smoke} any tobacco products, such as cigarettes, cigars or pipes?(t1) #

If yes, do you currently smoke tobacco products {daily}? (t2) #

On average, how many of the following do you smoke each day?

 {Man}ufactured {cig}arette{s} (t5a) ##

 Hand-rolled {cig}arette{s} (t5b) ##

 {Pipes} full of tobacco (t5c) ##

 {Cigars}, cheroots, cigarillos (t5d) ##

 {Other} (t5e) ##

 {Other spec}ify (t5other)_____

Alcohol

Have you consumed {alcohol} within the past {12 mo}nths? (a1) #

During each of the past 7 days, how many standard drinks of any alcoholic drink did you have each day?

 {Mon}day (a5a) ##

 {Tue}sday (a5b) ##

 {Wed}nesday (a5c) ##

 {Thurs}day (a5d) ##

 {Fri}day (a5e) ##

 {Sat}urday (a5f) ##

 {Sun}day(a5g) ##

Physical Measurements

Are you {pregnant}? (m5) #

{Height} in centimeters (m3) ###.# {Weight} in kilograms (m4) ###.#