# APPENDIX 3

# Organization and Administration of Data Processing

**C O N T E N T S**

## MANAGING THE DATA BASE

We suggest that you carry out data entry and data editing in small batches, usually completing data entry for an entire cluster at once. Data should be stored on the hard disk. After a batch has been entered, the data are passed through a consistency checking program and the results are printed out.

Any errors which are due to incorrect keying should immediately be corrected. The list of errors should be given to the supervisor to decide what action should be taken (for example, changing the data, consulting with the interviewer or leaving the error unresolved). Once corrections for errors have been made to the data, the consistency program should be rerun to verify the corrections. When there are no more errors that can be corrected in the office, the data should be backed up to diskette.

☞ Label your diskettes with the name of the files, the region and the date of the latest changes.

When the complete data base is ready for analysis, the checking program should be used again as a final check before frequencies are produced for each question.

## RECORD-KEEPING

Data processing operations frequently become messy because of the volume of data and the multiple procedures which have to be carried out. Good record-keeping can alleviate these problems.

Figure A3.1 shows a sample record-keeping form. This form has a line for each batch or cluster of data.

1. Record the numbers of questionnaires that return to the office from the field for each cluster. Check these numbers against the field work totals.
2. When a batch of questionnaires is submitted for data entry, record the date.
3. When a batch has been entered, record the date, the number of questionnaires and the number with errors.
4. Note the date that corrections are made.
5. Each time error reports from a batch are produced, note the number of documents with errors and the date.
6. When the main database is updated with the batch, note the date and number of records added.

Data from the questionnaire are entered for each household in the order they are collected: the household data from the listing page, the data on all children on the child roster, the water and sanitation and salt iodization modules, education data for each child over school-entry age, tetanus toxoid module for each mother, care of acute respiratory illness (CARI) module for each mother and, last, child health modules for each child. Once all the data for one household are entered, the clerk moves on to the next household.

**Figure A3.1 Data processing record-keeping form**

| OFFICE | | | DATA ENTRY | | | 1ST EDIT | | | 2ND EDIT | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cluster | No. of question-naires | Date given to data entry | Date entered | No. of question-naires | No. of errors | Date given for manual correction | Date file updated | No. of errors | Date given for manual correc-tion | Date file updated | No. of errors |
| | | | | | | | | | | | |
| | | | | | | | | | | | |
| | | | | | | | | | | | |
| | | | | | | | | | | | |
| | | | | | | | | | | | |
| | | | | | | | | | | | |
| | | | | | | | | | | | |
| | | | | | | | | | | | |
| | | | | | | | | | | | |
| | | | | | | | | | | | |
| | | | | | | | | | | | |
| | | | | | | | | | | | |
| | | | | | | | | | | | |
| | | | | | | | | | | | |
| | | | | | | | | | | | |
| | | | | | | | | | | | |

The data entry can be done by a number of data entry operators. Error reports are produced at the end of each cluster's data entry with a description of any inconsistent items.

Corrections, written on the error sheets, are used to correct the appropriate records on diskette. When no more errors that can be corrected are encountered, the data should be backed up to diskette.

## HOW TO INSTALL AND USE THE DATA ENTRY AND ANALYSIS PROGRAMS

### EPI INFO PACKAGE FOR MID-DECADE GOAL SURVEYS: INSTALLING THE PROGRAMS

All the files for data entry and analysis should be placed in a separate directory on the hard drive. To install the software, first create a new directory on the hard drive by typing at a DOS prompt:

```
MD C:\MIDDECAD
```

Then copy all of the files on the diskette into this directory by typing:

```
COPY A:*.* C:\MIDDECAD
```

It is assumed that EPI INFO v.6 is already installed on the hard drive. If not, follow the procedures described in the EPI INFO manual for installation. The \EPI6 directory should be included in the computer's PATH. You can verify that it is by typing `PATH` at a DOS prompt. If \EPI6 is not included in the PATH, modify your AUTOEXEC.BAT file to include it.

### RUNNING THE PROGRAM

The software should always be run from the \MIDDECAD directory. First switch to this directory by typing:

```
C:
CD \MIDDECAD
```

Then to run the programs, type:

```
MENU
```

This will bring up a menu of choices including:

1. Data Entry
2. Consistency Check
3. Variable Frequencies
4. Data Analysis

B. Data Backup

Q. Quit to DOS Sub-directory

**Data Entry**

Choose option 1 from the menu to either enter new data or to edit already existing data. Choosing this option will bring up the first data-entry screen for entering data from the household information sheet. A new questionnaire may be immediately input. To edit an existing questionnaire, press <CTRL F> and enter the CLUSTER and HOUSEHOLD numbers of the desired questionnaire. Then press <F3> and <ENTER> to bring up this questionnaire for editing. Pressing <CTRL N> will again allow new questionnaires to be input. For further details on moving about the data-entry screens within EPI INFO, refer to the EPI INFO manual on data entry.

Data entry is organized into several modules, including:

1. Household information (single occurrence)
2. Mother and child roster (multiple occurrences)
3. Water and sanitation (single occurrence)
4. Salt iodization (single occurrence)
5. Education (multiple occurrences)
6. Tetanus toxoid (multiple occurrences)
7. Care of acute respiratory infection (multiple occurrences)
8. Child health (multiple occurrences)

This organization allows the data-entry staff to complete all data entry for a page of the questionnaire before moving on to the next page, and thus to eliminate flipping back through the questionnaire.

Entry into each module is controlled at the bottom of the household screen. To begin entering data from the mother and child roster, simply press <ENTER> next to GoToChild. This will bring up a screen to enter the first line in the roster. When the data are entered and written to disk, the same screen is repeated, since multiple lines are allowed. After entering all lines from the roster, press <F10> (RETURN) and the household screen will reappear. Press <ENTER> next to GoToWater to go to the Water and Sanitation Module. Since only a single occurrence of the Water and Sanitation Module is allowed for each household, the household screen is immediately reentered upon saving the water record. The same procedures are applied for the remaining modules. If there are no data for a module, the module can be skipped by moving the cursor beyond the module before pressing <ENTER> or by simply pressing <F10> immediately upon entering the module. When data have been entered for all modules, pressing <ENTER> next to Finish will end the data entry.

The cluster and household numbers are concatenated to form a HOUSEID which serves as a key to link all of the files together. The HOUSEID must uniquely identify the household and cannot be repeated. Additionally, the mother's or child's line number serves as a key to link the individual's information to the appropriate line in the mother and child roster. It is important that this line number be entered correctly as many of the indicators rely on linking data from the roster and the specific modules.

***Notes on the specific modules:***

*Household information panel:* The date of interview must be entered because all ages and dates of birth must be verified against the date of interview.

*Mother and child listing form:* Line numbers for mothers or caretakers must end in a 0 (i.e., 1-0, 2-0, etc.). Date of birth is not entered for them. Line numbers for children must end in a number from 1 to 9 (i.e., 1-1, 1-2, etc.)

*Tetanus toxoid:* The mother's line number (ending in 0) must be entered here to be used to link the mother to her youngest child. (The key for linking children is based on the line number plus 1, since dates of immunization are to be compared against the youngest child's date of birth.) Data for question 6 need not be entered, since the answer to question 6 is simply a total of questions 2 and 4.

*CARI:* The mother's line number (ending in 0) must be entered, although the key for linking is based on the line number plus 1.

*Children under age five years of age:* Because all of the modules for children under five years of age (diarrhoea, vitamin A, breastfeeding, immunization, and anthropometry) are to be filled out for one child before going on to the next child, they are treated as a single module for data entry. Data on the date of interview and the child's date of birth must be reentered in this module because EPI INFO will calculate anthropometric indices at the time of data entry.

**Consistency Checks**

Data for each cluster should be checked for internal consistency as soon as all the data have been entered for the cluster. Choose option 2 to perform consistency checking. You will be prompted as to whether you want to send the results directly to the printer or to a file for future examination. You will also be prompted as to what range of clusters are to be checked. If you are only interested in cluster 101, type `101` as the starting and as the ending cluster numbers. If you would like to run consistency checks for the entire file, type `000` as the starting cluster and `999` as the ending cluster.

Consistency errors will be written to a file called CHECK.TXT or to the printer directly. Each error statement shows: (1) the household identification number, (2) the questionnaire module that contained the error, (3) a description of the error and (4) the individual number, if applicable.

**Variable Frequencies**

When all the data have been entered, it will be useful to examine a frequency listing of all the variables in the data file. Choosing option 3 will produce a frequency listing of all the variables with the exception of HOUSEID and CHILDID.

**Data Analysis**

Choose option 4 to analyse the data. This option runs the set of commands contained in INDICATR.PGM, which produces the results for the indicators, according to the specifications laid out in Appendix 4. You can obtain a full listing of the .PGM files by simply printing the file, either to the screen or to your printer. The

output is limited to looking at the specific indicators for the entire sample. The program can be modified to look only at specific subgroups of the population, and then rerun for that subgroup. Instructions on modifying the program are given in the section "Notes to the Programmer" below.

The analysis program does not supply standard errors for the indicators. These may be calculated using the EPI INFO v.6 CSAMPLE option. See below for instructions on how to use CSAMPLE.

**Data Back-Up**

Data should be backed up to diskette on a daily basis, using different diskettes on different days of the week. Choosing option B will copy the appropriate files to the diskette. This procedure assumes that the size of the data files will not exceed the capacity of a single diskette. In the event that the data files become too large for a diskette, the data manager will need to find an alternative system for backing up and modify the MENU.BAT file accordingly. Systems which might be considered include PKZIP, BACKUP (DOS), or PCTOOLS.

**Enter EPI INFO**

Should there be a need to work directly with the EPI INFO package, choosing Option E will take you to the EPI INFO v.6 startup screen. All the EPI INFO programs and facilities can be run from this screen. When exiting EPI INFO, you will return to the menu. (Option E does not actually appear on the menu list, since data-entry staff will not regularly need to enter EPI INFO directly. The option is functional, however.)

**Quit to DOS Sub-directory**

Choose option Q to exit the menu system.

**RUNNING CSAMPLE TO OBTAIN STANDARD ERRORS FOR CLUSTER SURVEYS**

To estimate correct standard errors for the indicators in EPI INFO, a data set must be on disk that has the created variables for the indicator of interest. For example, suppose we are interested in obtaining standard errors for Indicator 11.1, the per cent of under-fives with low weight-for-age. To create such a data set, add two lines to the INDICATR.PGM file after the line "report waz.rpt":

```
ROUTE TEMP.REC
WRITE "RECFILE" CLUSTER LOW2WAZ
```

and run the analysis. This will create a temporary file called TEMP.REC, which contains only the cluster variable and the variable on whether or not the child has low weight-for-age.

Then choose the CSAMPLE option from the PROGRAMS menu option in EPI INFO v.6. Type `TEMP` as the file name. A screen of options will appear. Under Main, type `LOW2WAZ`, the variable for which you need standard errors. Specify the strata variable only if the sample design was stratified. Specify the PSU (in chapter 4, the term "small area" refers to these Primary Sampling Units) variable, usually given the name `CLUSTER`. Weight is not needed unless the sample design requires weights. Results may be viewed immediate Specify a Crosstab variable only if you need to examine standard errors within population

subgroups. ly on the screen, printed directly to a printer or stored in a file for future use. If the data are not already sorted by CLUSTER, then choose SORT. Choose TABLES, and the analysis will be performed.

The output gives the standard error for the indicator and its 95 per cent upper and lower confidence intervals. The program also computes the design effect obtained in the survey for the variable of interest. For instance, in analysing weight-for-age, the output might look like the example shown in Figure A3.2, in which the percent (rounded to the nearest tenth) with low weight-for-age is 21.0 with a confidence interval of (19.3, 22.7).

**Figure A3.2 Example of CSAMPLE output when analysing weight-for-age**

**CTABLES COMPLEX SAMPLE DESIGN ANALYSIS**

```
 Analysis of LOW2WAZ
LOW2WAZ
|            |Total      |
|------------|-----------|
|0           |           |
| Obs        |       2168|
| Percent   V|     79.009|
| SE%        |      0.887|
| LCL%       |     77.271|
| UCL%       |     80.746|
|------------|-----------|
|1           |           |
| Obs        |        576|
| Percent   V|     20.991|
| SE%        |      0.887|
| LCL%       |     19.254|  LCL = lower confidence limit
| UCL%       |     22.729|  UCL = upper confidence limit
|------------|-----------|
|Total Obs   |       2744|
|------------|-----------|
|Design eff. |      1.300|
|------------|-----------|

  Sample Design Included:
  -----------------------
  Sampling Weights--None
  Primary Sampling Units from CLUSTER
  Stratification--None

  0 records with missing values
```

## NOTES TO THE PROGRAMMER

Some modifications to the programs will be required in every country where the package is used. For example, the definition of "safe and convenient water" is country-specific, and therefore, the analysis program INDICATR.PGM must be edited to reflect the definition used in a specific country. The definition of school-entry age varies from country to country. Also, alternative modules exist for vitamin A as well as for immunization coverage. A programmer who is familiar with EPI INFO must make the appropriate country-specific modifications to the program. It is best that the programmer fully understand the structure of the data base prior to making any modifications. This is especially true if there are to be many changes to the questionnaire.

Any changes to a .QES or .CHK file require that the corresponding .REC file be recreated. This can be accomplished by choosing ENTER in EPI INFO v.6 and choosing option 2, "Create new data file from .QES file." Since the data are contained in the .REC file, it is important that all changes to .QES and .CHK files be made before data entry begins. If it is necessary to make some modification to these files after data have been entered, consult the EPI INFO manual on "Revising the Structure of a Data File (Menu Choice 3)," in chapter 8, "Entering Data."

### Removing Modules

Any modules in the sample questionnaire that a country chooses not to implement can easily be skipped. If the module to be deleted is one of those listed above on page A3.6, a simple deletion in the HOUSEHLD.CHK file will prevent ENTERing the module. The variable in the HOUSEHLD.QES for that module should also be deleted.

If only a portion of a data-entry module is to be deleted (for example, the breastfeeding questions within the child health module), then the lines referring to those questions should be deleted from the .QES file and the .CHK file. The alternative vitamin A and immunization modules which are not implemented should be removed in this fashion. The appropriate number of blank lines should be added or deleted to the .QES file to ensure that each page breaks on the screen at a logical place.

### Adding Modules

If additional modules are to be added to the questionnaire, it will be necessary to develop new .QES and .CHK files. The existing files can serve as templates of how to do this. In most cases, additional questions could be added to the existing modules, in which case, modification of the .QES and .CHK files should suffice.

### Changing the Wording of Questions or Response Categories

Changes to the wording of questions (including translation into languages other than English) simply requires that the screens defined in the .QES files be modified to show the appropriate wording. The original variable names which are used in the .QES, .CHK and .PGM files need not be changed. However, if the variable names are changed, the changes must be made in all of these files.

If the categories of acceptable responses to questions are changed, appropriate changes must be made to the .CHK file (to allow all valid responses but not allow invalid responses) and to the .PGM file (to ensure that the indicators are calculated using the correct categories). The package does not explicitly assign labels to the responses, so that there is generally no need to type in labels to the responses.

**Changing the Analysis/Consistency Checks**

Both consistency checking and calculation of the indicators are accomplished through .PGM files. These files can be edited to remove references to any modules not implemented and to specify the definition of certain indicators (e.g., safe and convenient water, school-entry age and frequency of vitamin A capsule use). The indicators are usually desired for other specific subgroups, such as region-specific indicators. The indicators are usually desired for other specific subgroups, such as region-specific, or sex-specific, indicators.

☞ Obtaining estimates for sub-groups: To calculate each indicator for specific subsets of the data, the easiest modification would be to select only those observations in the subset, and rerun the analysis program. For example, to calculate the indicators for males only, select the male observations:

```
SELECT SEX = 1.
```

Since the select statement must be executed every time a new data set is read, the INDICTR.PGM includes a subroutine at the end called ":SELECTSEX" to perform the analysis for whatever subsets of the data are required. Edit this subroutine and then run the revised program to calculate the indicators for boys only. For girls only, revise the subroutine with the line:

```
SELECT SEX = 2
```

and rerun the INDICATR.PGM file. The select statement can also be revised to consider only observations in a specific REGION, AGE group or URBAN/RURAL category. Of course, the programmer should make sure that the variable of interest is always available by issuing an appropriate RELATE command before the selection.

☞ Additional .PGM files can be created for other purposes—for example, to compute the mean duration of breastfeeding or to examine the age distribution of children. Examination of the existing .PGM files can be helpful in getting started on new analysis.

☞ All the EPI INFO analysis programs are written assuming that the sample in the country is designed to be self-weighting (i.e., the weights are all 1.0). However, a subroutine is included to make weighted analysis as straightforward as possible. The programmer will simply need to edit the subroutine ":WEIGHT" at the end of the INDICATR.PGM file, to include assignment of weights as described in chapter 7.

### Merging Data Sets from Several Computers

In some cases, it will be necessary to enter questionnaire data on several computers and then merge the data together for analysis. This can be accomplished through the MERGE program in EPI INFO. The .REC files from one computer should be copied onto the other, but placed in a separate subdirectory. The MERGE program, using the CONCATENATE option, will create a new data file with data from both files. The procedure can be repeated if more than two computers are used. The MERGE must be completed separately for each module.