20.7: Preparing data for analysis

The ‘raw materials’ for data analysis are the data files created by the data management process. However, the variables, as recorded in the questionnaire and entered into the database as raw data, are not always the ones directly suitable for data analysis. Recoding and creating of new variables is likely to be necessary. It is generally also necessary to combine information from different data files.

When preparing the data for analysis, it is good practice to create a new data set with a different name to separate it from the original study data. Also, it is advisable to keep a copy of the commands used to prepare the data (either the program that was used or the ‘log’ files), in case it is necessary to re-create the file from the raw data.

7.1 Data dictionary

The data dictionary is part of the metadata and is the link between the questionnaire and the data files. It typically contains the name and a description of each variable, with additional information such as the data type (for example, numeric or text), coding (for example, 0 = No, 1 = Yes), and the questionnaire section and question number to which the variable relates. The data dictionary is essential for understanding how the data are structured and is used in preparing for data analysis.

7.2 Creating new variables

Sometimes, it is necessary to create a new variable from two or more existing variables, since this new variable may be more meaningful than the ones on which data were collected directly. For example, body mass index (BMI, defined as weight in kilograms/ height in metres2) or weight-for-age may be better markers of nutritional status than weight on its own. Such composite variables may be calculated directly from the raw data or be obtained by comparison with a given
standard (as in the case of weight-for-age).

Variables related to time, such as the length of residence or the duration of exposure to a risk factor, present a special case. Depending on the characteristics of the variable and of the population under study, it may be preferable to record relevant dates on the questionnaires and to subtract them during the analysis stage to compute the duration of residence, exposure, etc. These calculations can be done, without difficulty, with any statistical package.

After creating a composite variable, it is useful to check that the distribution of the new variable seems reasonable. It is also appropriate to check the range of the new variable, as data errors may only show up at this stage. For example, negative ages or extreme weights-for-age may result from errors in the date of birth (or date of interview) in the questionnaire, though such errors should have been detected through consistency checks at an earlier stage.

### 7.3 Coding and re-coding

Before beginning the analysis, it is usually necessary to re-code some variables, so that they can be grouped into categories. Since it is advisable to look at cross-tabulations of data before moving on to regression methods, re-coding is generally needed for quantitative variables. Grouping makes it easier to understand the data and, in particular, to look for non-linear associations. But re-coding may also be necessary for categorical variables with large numbers of categories, or few observations in some categories.

When re-coding quantitative variables, one strategy is to divide the range of the variable into quartiles or quintiles, giving four or five groups with equal numbers of observations in each group. Alternatively, cut-off points may be chosen on the basis of established standards. For example, when grouping age, it is more natural to use 5- or 10-year age bands (for example, 20–29, 30–39, etc.), rather than base the categorization on quartiles. Similarly, there are recognized international cut-points for variables such as BMI (less than 18.5 is considered underweight) or weight-for-age (less than −2.0 is considered stunted). A histogram of the data is often a good way of deciding how to categorize a quantitative variable with no standard cut-points.

With categorical variables, it may be necessary to combine groups if there are very few observations in some groups. When combining groups, an important principle to remember is that, for combining to be appropriate, the risk of the outcome should be similar in each of the combined groups. For example, in a study of child malnutrition, it may not be appropriate to group mothers with no schooling with those with primary school education.

The number of groups to use also depends, in part, on how the variable will be used in the analysis. If the variable is an exposure of interest, where it is planned to examine the pattern of dependence of the outcome on the amount of exposure (for example, a dose–response), it is important to use enough groups to get a reasonable picture of the relationship. For example, to examine the effect of alcohol intake during pregnancy on birthweight, one group might be non-drinkers, and there could be four or five groups for different levels of alcohol intake.

After deciding if and how each variable should be grouped, the different categories should be assigned ‘labels’ to describe them. These labels should be saved in the data set, which will eliminate the need to return to the questionnaires or code lists during the analysis. When a variable is re-coded, it is important to create a new variable and allocate it a different name, so as to preserve the raw data. Thus, the variable ‘AGE’ might be grouped and allocated to another variable called ‘AGEGP’. 
7.4 Merging and linking data

The data required for a particular analysis may need to come from several different data sets (for example, questionnaire data on an individual’s recent sexual behaviour may need to be linked to laboratory results, demographic data collected previously, and household-level data on the socio-economic status). If complete data tables are extracted for analysis, merging of the data may be more easily managed in the statistical package used for the analysis.

Many data management packages allow the construction of complex views of the data and can be used to extract merged data for analysis. The data analyst can specify the variables for analysis, and these can be extracted from the database, using standard data management tools, thereby maintaining the confidentiality of the data. It also enables simple data extraction programs to be used at regular intervals for longitudinal data, giving regular snapshots of the data for analysis.