Allergol et Immunopathol. 2009;**37(2):**93-7

ELSEVIER DOYMA

Allergologia et immunopathologia

www.elsevier.es/ai

SERIES: BASIC STATISTICS FOR BUSY CLINICIANS (II) EDITOR: V. PÉREZ-FERNÁNDEZ

Design and debugging databases for statistical analysis

M.M. Rodríguez del Águila[a,]* and P. Garrido-Fernández[b]

[a]*Servicio de Medicina Preventiva y Salud Pública, Hospital Virgen de las Nieves, Granada, Spain*
[b]*Fundación para la Investigación Biosanitaria de Andalucía Oriental (FIBAO), Hospital Torrecárdenas, Almería, Spain*

KEY WORDS
Data file;
Debugging;
Statistical data analyses

Abstract
Any form of data analysis requires the prior creation of a database to house the study information collected in one format or other (questionnaire, clinical history, etc.). The design of such databases should be optimised to allow adequate statistical analysis without the drawing of wrong conclusions. In addition, prior to analysis, debugging or filtering of the variables is required in order to avoid doubling the effort made in extracting the results. The present study offers a series of suggestions for database design and debugging, to ensure that the later statistical analyses are based on the revised data.

Introduction

Prior to any type of investigation, a study protocol must be established, contemplating methodological and technical aspects which will be applied in the course of the study. To this effect, it is very important to establish an epidemiological design adapted to the planned study objectives, and to adopt a rigorous and explicit methodology.

It is advisable for the protocol to include a case report form to be used in the measurement and collection of the information from clinical or case histories, or by means of some other kind of procedure. The case report form should specify each and every one of the fields which the investigator considers necessary for the study, with specifications where possible of the coding system used in each case.

Once the information corresponding to the case sample or series specified in the protocol has been obtained, computer-based data input is required[1]. Good data input is a decisive element for ensuring that the subsequent analyses are correct. It is common practice to design databases in which the variables have not been well defined, or where the information has not been entered with uniform or homogeneous criteria. This results in the need for later restructuring of the data files, with the associated waste of time and resources.

The present study aims to offer investigators the guidelines needed to design databases for any type of study, and explains the procedure required to filter or debug the information prior to the final statistical analysis[2].

Designing a database

In designing a database, it is very important for investigators to select software with which they are familiarized. The

*Corresponding author.
E-mail: mmar.rodriguez.sspa@juntadeandalucia.es (M.M. Rodríguez del Águila).

0301-0546/$ - see front matter

section below describes some of the software applications that can be used.

In general, the data are entered in a box format obtained by crossing rows and columns. Each study variable conforms an individual column or field in the database, and all the data of a given subject are contained in one same row. Thus, if patient age is the first variable, gender the second variable, and marital status the third variable, the values of these three variables corresponding to a given individual will be contained in individual boxes aligned in a single row.

A series of recommendations for efficient designing of a database are provided below. The present study only addresses non-relational databases, i.e., those in which the data are contained in a single table or file. In comparison, relational databases comprise complex structures[3] in which different data tables are related though key fields.

Data file structure

The first step is to clearly establish the variables to be included in the file, in order to define the data fields according to the type of variable involved.

1. Quantitative variables (i.e., those measuring amounts) are to receive an adequate numerical format, indicating whether they correspond to integers (without decimals) or real numbers (with decimals). Whenever possible, these variables should be entered in numerical form, not grouped into intervals, since the latter approach gives rise to information losses.
2. Qualitative variables (i.e., those measuring categories) are generally coded to allow faster and more homogeneous processing. As an example, in relation to subject marital status, it is general practice to code the different categories as follows: 1 = single, 2 = married / partner, 3 = separated/divorced, and 4 = widowed. In this context, it is always easier to enter number codes than category labels, since the latter approach entails a risk of error (e.g., entering uppercase or capital letters in one case and lowercase in another). In most statistical programs it is possible to define these category labels *a posteriori*, and it is convenient to keep them at hand in order to identify the different categories.
   For variables of this kind it is advisable to generate a coding manual with the codes corresponding to all the categories, so that each time the database is used a document is available to identify the content of each variable and the way in which it is measured (i.e., independent of the person creating the database). This document is usually included in the established study protocol, in the section corresponding to measurement variables, or as an annex identifying the case report form.
   When qualitative variables contain many categories, it is sometimes preferable to enter the data as chains, followed by grouping and coding — although this may be time consuming. Emphasis is placed on the convenience of entering the data in coded form, as described above.
3. Data relating to dates are to be defined as such in the database, in short format (day/month/year, dd/mm/yy). Data collected as hours and minutes should be transformed to decimal format for analytical purposes, dividing the minutes by 60 and adding them to the recorded hours.
4. Dichotomic variables (with only 2 possible values) should be coded as 0 and 1 (where 0 = absence and 1 = presence). In the case of the variable sex or gender, 0 and 1 should be taken to indicate females and males, respectively. In general, such coding will depend on the established objective. Thus, if the aim is to determine risk factors, 1 (or the highest number code) should be used in reference to the category favouring appearance of the event of interest.
5. Those variables that allow multiple and non-mutually excluding responses are to be defined as different fields (one for each variable). As an example, when considering a series of symptoms classified as *pruritus*, *rhinitis*, *asthma* and *others*, one variable is defined for each symptom — followed by input as 0 if the subject does not have the symptom in question, or as 1 if the subject has the symptom. It is very common to find databases in which all these data have been entered under one same variable — thereby making analysis impossible. As an example, if a subject presents *pruritus* (coded as 1) and *asthma* (coded as 3), it is common to find both variables merged into one as "1,3" — when in fact the value 1 should be entered in the column corresponding to *pruritus*, with the value 1 in the column *asthma*, and values of 0 in the columns *rhinitis* and *others*, since the subject in question does not have these latter symptoms.
6. When repeated measurements of one same subject are made, they should be reported as independent variables. As an example, different spirometric measurements over time are to be entered as different variables (one for each measurement made).
7. Missing values (i.e., values not obtained, lost values, data collecting errors, etc.) are to be coded in blank or using a code outside the range for the variable in question — with later indication in the program of the value represented by the mentioned lost code, in order to prevent such data from being included in posterior analyses. Whenever possible, it is advisable to use the same scheme for the coding of these values, for all variables. As an example, if the parameter *age* has not been obtained for some patients, these values are usually left blank, or alternatively the value "−9" is entered — the program later indicates that this code represents a missing value (subject age cannot yield the value "−9").
8. All databases tend to contain an identification variable at the start. Such an identification code must be unique for each subject, guaranteeing data confidentiality, and making it possible to link each case to the questionnaire, clinical history or information of interest for the study.

Data input

In general, the data are entered by rows, each row corresponding to an observed case or subject. In some databases and spreadsheets, as well as in most statistical packages, drop-down menu fields can be used for qualitative data input — a fact that facilitates input processing.