Biostatistics Unit

Biomedical Research Institute of Lleida

Objectives and operation

APRIL 14th, 2020

1 Introduction

- The Biostatistics Unit (UBiostat) has the **mission** of contributing to generate knowledge to improve health. Statistical methods are essential to design studies, to analyze data, and to interpret results. At the UBiostat, we work with researchers to convert data into useful information, with the goal of improving the research. It is important to make aware researchers of the importance of contacting the Ubiostat at the very begining of their research, for the sake of better designing the experimental/observational research and planning the further analyzes.

- The UBiostat **depends** on the IRBLleida Director. Hence, the projects and the work developed by the UBiostat are presented to the Director.

- The UBiostat **aims to finance itself** with the income from the work carried out and the **financial contribution** of IRBLleida. For this reason, financing mechanisms have been established, which are described in section 4 of this document. All groups/researchers, emerging and consolidated, that don't have financial resources for the statistical section, will be guaranteed initial support from UBiostat to be able to prepare research projects or analyse data that will justify future funding.

- Ubiostats members can be contacted in the IRBLleida website. In the scientific-technical services section, in the UBiostat page, a online forme can be found for this purpose.

2 Objective and operation

The **main objective** of UBiostat is to improve the efficiency and quality of the research undertaken by the different groups at the IRBLleida. This objective is implemented as follows:

1. By providing advice and methodological support for the design and data analysis of studies from IRBLleida research groups.

2. By participating in multidisciplinary teams to improve the processes of data collection, validation and integration for IRBLleida research groups.

3. By participating actively in Biostatistics networks in Spain and abroad.

4. By organizing and participating in training courses in Statistics and Research Methodology in the context of training programs at IRBLleida.

5. By providing advice and methodological support for the design, data analysis and publication of studies to external entities through the establishment of collaboration agreements.

6. By training master's and doctoral students to collaborate with researchers from UBiostat and IRBLleida.

3 Types of service

1. **Scientific collaborations:**

   Participation in the design, statistical analysis, interpretation of the results and writing of articles. It includes the assessment of the budget to request financing for the tasks of the UBiostat.

2. **Initial statistical support:**

   Advice on the preparation of research projects for emerging and consolidated groups.

3. **Internal consulting** It allows solving specific doubts about some of the aspects of the statistical analysis of a project.

4. **Training**

   It consists of improving the training of researchers in Statistics to be more autonomous. Courses in Basic and Advanced Statistics and the use of the R program for data analysis are provided.

5. **Agreements**

   The aim is to establish collaborations between IRBLleida and different entities for studies or programs of bilateral interest.

6. **External consulting**

   The UBiostat is open to requests for statistical advice, data analysis and writing reports or articles from external entities.
   Service requests can be made online. After requesting the service, a meeting will be arranged with the user, a budget will be prepared and an estimated date of completion of the study will be given to the user.

Attached documents

This document and the annexes specified below are available on the IRBLleida website, in the specific space of the UBiostat. These documents will be updated periodically in order to streamline and improve communication between researchers and UBiostat.

A.1 Service request sheet (online).

A.2 Document of work plan, deadlines and budget.

A.3 Ethical aspects and confidentiality of the data.

A.4 Recommendations for the design and coding of data files from research studies.

A.1 Service request sheet

This document can be completed online on the UBiostat page.

Applicant details

- Name
- Department/ Service
- Institution
- Telephone
- E-mail

Type of study

- With financing - Statistical support included
- With financing - Statistical support not included
- Without financing
- Request for new projects
- Others (especificar)

The purpose of the study

Specify the question the study intends to answer.

Service

Indicate what type of statistical service or methodological support from UBiostat is required: (Check all the necessary sections)

- Competitive project request
- Study design
- Database design
- Analysis of data
- Others (specify

Commentaries

Free Text

A.2 Document of work plan, deadlines and budget

This document will be prepared jointly by the service applicant and the UBiostat.

Applicant details

- Applicant
- Principal investigator
- Department /Service
- Institution
- Name of the related research project
- CEIC approved?

Study Objectives

The person or group requesting the service must describe in detail the objectives of the study or the questions they want to answer.

Work plan

The UBiostat will prepare the Work Plan together with the applicant. The Work Plan must detail the tasks that will be carried out at UBiostat to respond to each of the objectives. It should also detail, if possible, the tables and figures that the results section should include.

Delivery time

The UBiostat will make a proposal for delivery dates according to the work volume of the UBiostat and the type of request.

Budget

The UBiostat will prepare a budget based on the workload involved in the request.

Firm

Date and signature of both parties

A.3 Ethical aspects and data confidentiality

A.3.1. Ethical aspects

The UBiostat will consider the ethical aspects, the quality and the viability of each scientific project, in accordance with current legislation:

- Law 14/2007 on biomedical research
- Organic law 15/1999 on the protection of personal data
- Law 41/2002 regulating patient autonomy and rights and obligations regarding information and clinical information.
- ICS Good Practice Guide to Health Science Research. Institut Català de la Salut, July 2105

A.3.2. Confidentiality of data

- UBiostat recommends removing the names of patients or identifiers such as the CIP or NHC of patients included in the databases that are provided to UBiostat staff. All data files must include a numerical patient identification code.
- The main body that provides clinical data is responsible for having the informed consent of the patients participating in their research project.
- All clinical information will be anonymous and confidential. The UBiostat will watch to protect the integrity and confidentiality of all the data files that the researchers provide it.

A.4 Design recommendations and encoding files data from research studies

This document aims to establish some recommendations that guarantee quality in the design of databases and the collection of information. If the data are recorded well, the statistical analysis is faster and the risk of making mistakes decreases.

A.4.1. Confidentially of data

The UBiostat recommends deleting the names or identifiers, such as the CIP or NHC, of the patients included in the databases provided to the unit. It's necessary to create the identifier (ID) variable that will serve to relate all the data tables that need to be linked. UBiostat staff can advise on the creation of this identifier.

A.4.2. Design and management of databases with the REDCap application

In order to correctly manage the collection of data -especially when the volume is important-, it is recommended to use web applications designed for this purpose, such as REDCap (Research Electronic Data Capture).

REDCap is a secure web application that is used for the construction and management of online surveys and databases. REDCap can be used to collect practically any type of data. It is especially oriented to data collection for research studies. The REDCap Consortium is made up of 2.213 active institutional partners in 108 countries, who use and support REDCap in different ways. Currently the application is used by more than 370.000 projects with more than 475.000 users covering numerous areas of research interest throughout the consortium.

Despite requiring some computer and statistical support, its use is recommended to minimize errors in the data and thus speed up the statistical analysis. Therefore, we encourage you to contact UBiostat before launching a new study.

A.4.3. Design and management of databases without web applications

If any web application that guarantees a minimum quality of the data collected is used, it's necessary to take into account certain indications - some more general and therefore common to all data sets, and others more specific depending on the study design and the nature of the data to be collected.

- Data file dictionary. The dictionary of the data file will be the document that will exhaustively collect all the information of the study variables and that will help the statistician to analyze the data correctly.
- General indications:
  - Each row of the database will refer to a record (individual or visit) and each column to a variable.
  - Each record will have a unique identifier. In the event that each individual has a single record, the identifier will be based on a variable (for example, "Pac.ID" in Table A.5.1). If we have multiple records for the same individual, the identifier will be based on a combination of variables (for example: "Pac.ID", "Data.visit" in table A.5.1).
  - The same name will not be used for two variables. In case of measuring the same variable different times (for example: different visits, pre-post, ...) he multiple record format should be used for the same individual with the unique identifier as indicated in the previous point. The date of each of the measures must be recorded.
  - If it is a single variable measured twice, you can choose the format of a single record per individual. In this case, the two measures are collected in two different variables (for example: "tas_basal", "tas_v1" to Figure A.5.1): the first part of the name will have to be coincident between all and use the

same separator (“.” o “_”) enter the name of the measure (“tas”) and the identifier of the moment of the measurement (for example: “basal”, “v1”; “pre”, “post”).

- The variable names must not include spaces or special characters and must reflect the content of the variable in abbreviated form (for example: “IMC” to Figure A.5.1). The characters “.” o “_” can be used. (for example: “Pac.ID”, “Data.visit” to Table A.5.1)

- All the values of the same qualitative variable must be measured with the same units, which must be identified in the database dictionary (for example: “tas_v1” in mmHg in Figure A.5.1).

- In case of censored quantitative measurements (for example: “Hours.observation” = 2, 17, 12, 5, “>24”, : : :) you cannot mix the real numerical values with the categories (> 24). Censored values will be reported with specific numerical codes whenever possible (in no case will numerical codes be used that can be confused with real uncensored quantitative values), or with an auxiliary variable that indicates the censored observations of the original variable.

- All qualitative variables must be recoded numerically and previously. All recoding will be noted in the database dictionary (see Figure A.5.1). The use of accents and special characters (ñ, ç, º, ª,%) should be avoided. (for example: Municipality 1=Lleida, to Table A.5.1)

- Only in the case of open, non-recordable questions (for example, text variables such as comments) can they be registered as alphanumeric variables, which in no case will be analysed.

- Missing values will be recorded with specific numeric codes (blank space should be avoided). In order to collect the reason for the missing, it is recommended to use specific numerical codes according to the cause of the missing. These codes will be registered in the dictionary of the database (see Figure A.5.1). The code used to indicate the missings must be the same for all the variables in the database and cannot correspond to any possible recoding value. (for example:

in the variable “Cig_dia”, 998 can be used as a non-applicable value in the case of non-smokers, while 999 refers to a missing value (missing value))

- The calendar dates must be recorded (date of birth, date of visit, date of the event, ...) from which it will be possible to obtain the exact value of the measures of interest (inclusion age, time between visits, survival time, time until recurrence,...) in the appropriate temporary units.

- The database file must be written in one of the following formats: Excel (.xls, .xlsx), SPSS (.sav) or .csv.

Table A.5.1: Example of the structure of a general database

Pac.ID	Data.visit	Municipality	Sex	Age	Smoker	Cig_day
1	08/08/2016	1	0	41	0	998
2	10/08/2016	2	1	54	1	3

Figure A.5.1: Database Diccionary

Grupo	Nombre variable	Descripción variable	Tipo	Niveles/Rango				Unidades
Visit basal	age	Patient age	Numerical	≥ 18				
Visit basal	sex	Patient gender	Categorica	0=Man	1=Woman			
Visit basal	menopause	If have the menopause	Categorica	0=No	1=Yes			
Visit basal	tas_basal	Basal systolic blood pressure	Numerical					mmHg
Visit basal	tad_basal	Basal diastolic blood pressure	Numerical					mmHg
Visit basal	epworth	Daytime sleepiness index	Numerical	[0,24]				
Visit 1month	IMC	Body mass index	categorical	0=Under weight	1=normal weight	2=Overweight	3=Obesity	
Visit 1month	hospital	Hospitalizations during the year	Categorica	0=No	1=yes			
Visit 1month	hospital_obs	Hospitalizations observations	Chain					
Visit 1month	tas_v1	Systolic blood pressure 1 month	Numerical					mmHg
Visit 1month	tad_v1	Diastolic blood pressure 1 month	Numerical					mmHg
...	...	...	...	...	...	...	...	...

A.4.4. Defining the database variable

In order to correctly interpret the information and facilitate the analysis of the data, it is necessary to thoroughly understand each of the variables. The

database dictionary must be the document that contains the necessary information.

- **Group:** In case the variables are structured in blocks, indicate which of the blocks each variable belongs. (for example: basal visit)
- **Variable name:** to name each of the variables in the database.
- **Variable description:** to describe each of the variables in the database. (for example, for “IMC” the description should be “Body Mass Index (kg/m2)”)
- **Type:** to specify the type of variable. (quantitative, qualitative (with always predefined numeric codes), or alphanumeric without accents or special characters).
- **Levels /Range:** to specify each of the levels of the variable, if it is a factor (for example: “level_studies”, 0=no studies; 1= primary studies; 2= secondary studies; 3=university studies). To specify the minimum and maximum possible values of the variable, if it is quantitative (for example: “age” 18, “ep-worth” 2 [0; 24]).
- **Units:** To specify the units in the variable description, if applicable. (for example, for “IMC” the description should be "Body Mass Index (kg / m^2)