Preventive Veterinary Medicine 113 (2014) 298–303

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed

The data – Sources and validation

Ulf Emanuelson*, Agneta Egenvall

Department of Clinical Sciences, Swedish University of Agricultural Sciences, POB 7054, SE-75007 Uppsala, Sweden

ARTICLE INFO

Article history:
Received 17 February 2013
Received in revised form
19 September 2013
Accepted 1 October 2013

Keywords:
Database
Observational studies
Secondary data
Validation

ABSTRACT

The basis for all observational studies is the availability of appropriate data of high quality. Data may be collected specifically for the research purpose in question (so-called "primary data"), but data collected for other purposes (so-called "secondary data") are also sometimes used and useful in research. High accuracy and precision are required (irrespective of the source of the data) to arrive at correct and unbiased results efficiently. Both careful planning prior to the start of the data acquisition and thorough procedures for data entry are obvious prerequisites to achieve high-quality data. However, data should also be subjected to a thorough validation after the collection. Primary data are mainly validated through proper screening, by using various descriptive statistical methods. Validation of secondary data is associated with specific conditions – the first of which is to be aware of the limitations in its usefulness imposed by procedures during collection. Approaches for validation of secondary data will be briefly discussed in the paper, and include patient chart review, combining with data from other sources, two-stage sampling, and aggregated methods.

1. Introduction

Observational studies obviously rely on the availability of observations. Data based on such observations need to be of sufficient quality to avoid or minimize the risk of drawing false conclusions, i.e. falling victim to the "garbage in, garbage out" (GIGO) trap. This term was first used in 1963 within computer science (Anonymous, 2013) where computers were seen as unquestioningly processing the most nonsensical of input data ("garbage in") and therefore producing nonsensical output ("garbage out"; Fig. 1).

Nowadays, GIGO is also used in other areas and is equally applicable to, for instance, the field of analytical epidemiology – where we also can experience its counterpart where the modeling is miss-specified (Fig. 1; but that is a topic for other contributions of this special issue of Preventive Veterinary Medicine). For this paper, it is pertinent to realize that all data have errors – full stop! The task at hand is first and foremost to minimize data errors as much as possible, but also to identify errors so that their effects on the output can be identified and (hopefully) reduced. This paper is a summary of a presentation given at the Schwabe Symposium in December 2012 honoring the lifetime achievements in epidemiology and preventive veterinary medicine of Dr. Ian Dohoo; we briefly review's steps that can be taken to minimizing errors.

2. General recommendations

General recommendations on how to avoid systematic errors (i.e. reducing bias) and random variation (i.e. increasing precision) in data to be used in observational studies (and, indeed, most other types of studies) can be grouped into pre-execution, during-execution, and post-execution actions. Most such steps should be obvious to all who have a basic training and understanding in research, but can easily be overlooked and therefore are worthwhile to identify.

Prior to executing a project, the most important issue – but one sometimes forgotten or at least unheeded – is to formulate a clear hypothesis that is parsimonious and

* Corresponding author. Tel.: +46 0 18 67 18 26; fax: +46 0 18 67 35 45.
E-mail address: ulf.emanuelson@slu.se (U. Emanuelson).

0167-5877/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.prevetmed.2013.10.002

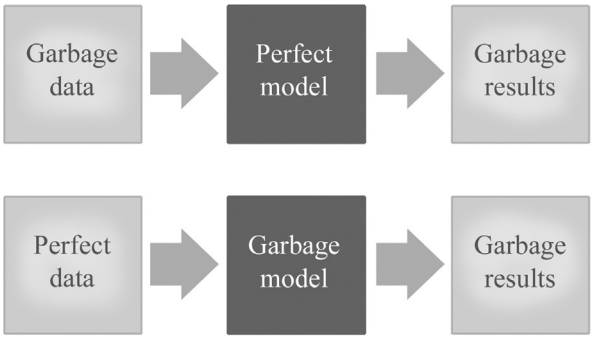

Fig. 1. The garbage in, garbage out paradigm.

testable. The hypothesis is a foundation for: identifying the most suitable study design; deciding what data to record and how to record them (to make it possible to test the hypothesis); and calculating an appropriate sample size. A proper hypothesis is therefore crucial for the validity of a research project. A next step in planning a project is to clearly define all observations that will be made in terms of unit of observation, types of variables (continuous, categorical, etc.), precision of measurements, etc. prior to the actual recording. Much can be said about how to record observations, but that is outside the scope of this paper. However, it is worth emphasizing that when data are collected through questionnaires, it is important to validate the questionnaire thoroughly prior to execution, e.g. for use with different languages (Dufour et al., 2010). A possible first step could be to conduct a pre-pilot test, which can be performed on a convenience sample of subjects (or their owners) or others that might not necessarily be part of the target population (e.g. colleagues, experts in the field). A proper pilot test should definitely be performed on a reasonable number of subjects who are representative of the target population for the questionnaire – but that will not be included in the actual sample. An appropriate piloting process allows the investigator to identify questions that are confusing or where there would be no variation in the answers. Finally, in all research it is important to ensure that all persons involved in the gathering of data understand their role in the project and are properly trained for their task – and perhaps continuously updated to avoid drift in data recording.

A continuous monitoring of the data that are recorded during the execution of a study is good practice. In some cases, missing values can be updated immediately when data handling takes place almost immediately after data have been compiled. Errors (or deviations e.g. in diagnostic tests) can also be discovered in time – but corrections are not easy to make if errors are discovered only after the completion of the study. Data-entry procedures should therefore also be in place during the execution phase, and these preferably could be designed so as to minimize the risk of errors at data transfer or typing errors. Finally, data should be validated after they have been gathered – which is the topic for the remainder of this paper.

3. Primary data

Primary data are data that have been collected with a specific research question (hypothesis) in mind. Their validation starts when the information is recorded. Ideally (and as we already pointed out), this should have already started before or during the execution of the study. Observations might be recorded in paper form – but at some point, all information will be put in a computerized format. Spreadsheets are convenient to use for that purpose, and therefore quite commonly are used – but they must be used with caution because it is possible to sort individual columns (and thus completely destroy the data), and their "seeming credibility" may cause unwanted (and unnoticed) changes of data. It is also more difficult to trace data edits using spreadsheets. A much better option would be to use a general-purpose database manager, of which there are several commercial alternatives (e.g. MS Access) but also within the public domain (e.g. OpenOffice, EpiData). Not only are the database managers not prone to the same errors as spreadsheets, but also they allow some error checking to be done at data entry (e.g. by using input masks or consistency checks). It is also worthwhile to consider a relational database when data are hierarchical in nature, because entering some of the information in duplicate can be avoided (and this minimizes the risk of inconsistencies). To reduce typing errors, data should be entered twice, with an automatic comparison between the entries. An alternative is to proofread all or parts of the data against original records. Data might also be scanned and parts checked manually (Murray et al., 2010). Software systems that read data from practice records have been used to structure clinical data (Lam et al., 2007); these offer potential for directly using clinical data.

Irrespective of the method of data recording, it is almost equally important also to collect and record metadata with the file (i.e. data about the data). Such data should contain a general description of the database and how the data were collected and by whom – but also should include definitions of variables (columns), units of measurements (e.g. kilogram, liters, optical density), precision of measurements, date of recording (including time-zone information), interpretation of codes, etc. Metadata might not be necessary for the immediate validity of primary data – but are absolutely crucial for their long-term preservation and use. Recommendations on stewardship of data and other aspects of databases can be found in a report from the US National Academy of Sciences (2009).

Validation of primary data either post-execution or during execution, is done by intelligent use of descriptive statistics. Variables measured quantitatively (on either a continuous or a discrete scale), could be evaluated by identification of possible outliers and illustrated e.g. by using a boxplot (also known as a "box-and-whiskers diagram"). Variables measured qualitatively will take only particular values and may be evaluated by using frequency tables to identify "illegal" (or at least, non-plausible) categories or unexpected distributions. Stratified analyses should be used for the evaluation of both quantitative and qualitative variables. Further exploration of the recorded data could be done by making use of graphical illustrations