J Oral Res
oral and craniofacial sciences

Editorial

Keep it rich, keep it simple.

José de la Macorra.[1]

Affiliations: [1]Dean, School of Dentistry Complutense University. Madrid, Spain.

Corresponding author: José de la Macorra. Department of Conservative Dentistry (Estomatologia II), Faculty of Odontology, Complutense University, Plaza Ramon y Cajal s/n, 28040. Madrid, Spain. Phone: (91) 394 1996. E-mail: jcmacorra@gmail.com

Novel researchers often look for help to analyze data collected during their first research projects. They have usually collected a huge, chaotic collection of data and seek an appropriate statistical test that fulfills their needs.

This usually indicates a notable disconnect between the components of a research project, as choosing an appropriate data analysis methodology needs to be performed at the beginning. The analysis used to test an hypothesis should be defined at the start of a project as well as data defined -groups, type of variables, number of specimens- and these should be collected according to its needs. If one takes a look to the data before deciding which test to employ, then one will tend to choose one that will likely confirm previous assumptions, increasing the risk of a Type I error (finding a positive result when there is none).

Obviously, the test should be appropriate to the hypothesis to be tested also in its design: to be able to consider all groups at the same time, to look for a tendency if its presence is what one is checking, to compare means, proportions, ranks, changes, periods or testing the hypothesis needs.

Also a usual mistake is the way in which data are collected. Normally, a spreadsheet will be used, with cases in rows and variables in columns.

The selection of variables to be measured and collected is critical. A variable should grant valid and reliable information about the research question, be easy to collect minimizing errors, and should allow proper statistical analysis.[1] These traits are difficult to meet, and failure to do so hampers novel researchers' efforts.

As an example, inexperienced investigators get easily caught in the surrogate outcomes fallacy (a surrogate endpoint of a clinical or laboratory trial is a laboratory measurement or a physical sign used as a substitute for a clinically meaningful endpoint that measures directly how a patient feels, functions or survives[2] or a laboratory specimen performs). Such laboratory measurements or physical signs are more easily obtained than the measure of the complicated, multifaceted reality, are possibly the only ones at hand, but they do not complete, fully and accurately represent the real actual outcome.

Also, trainees in research often collect data needlessly losing its richness and decreasing their statistical power and meaning. Many results of a clinical or laboratory measure can be gathered numerically, in what is defined as a continuous variable, one that can take on any value between its minimum and its maximum values. Many measures, in the contrary, are specified in ordinal levels (high/moderate/low, grade I/II/III, positive/

Conflict of interests: None.

Acknowledgements: None.

Cite as: de la Macorra J. Keep it rich, keep it simple. J Oral Res 2017; 6(1): 6-7.
doi:10.17126/joralres.2017.001

ISSN Online 0719-2479 - www.joralres.com

irrelevant/negative), for the sake of simplicity and easiness of use. But information gathered in this manner is reduced. For example, grouping will consider all cases included as Grade I as the same, when it clusters values that are or may be very different, but have been defined as belonging to the same category. Obviously, information is lost and muddled here. And not only information is poorer, but also its statistical handling will be less informative and rich. It is always important to collect data in the most possibly informative level, so that the information can be processed using the most powerful analysis at hand. When appropriate, this information can be grouped when the results are delivered, if it is prudent to do so in order to increase or simplify its clinical meaning.

Additionally there are some additional basic rules that are not always met: codification and consistency.

Codification must be simple and efficient. Different possibilities to represent the values of categorical variables (male or female, yes or no, types A, B or C), should be coded in the simplest way, and designed from the beginning. For instance, use 0 and 1 for dichotomous variables (male/female, yes/no, dead/alive, positive/negative), or 1,2,3, for variables having three possible values, and so on. If a new variable appears during data collection, or clarification of an existing one is needed, add it as a new one instead of changing the originally defined ones. It is important to keep in mind that at the moment of the analysis one can group or combine variables to give them more meaning or refine them.

It is tempting to color cases to help identifying them, but abstain from using colors to code cases: it's more sensible to add a variable, because during data analysis, colors will not be useful for statistical analysis. For instance, if you want to identify valid cases to your study, use a new variable (valid will do) and enter a simple 0/1 code where appropriate.

Data ought to be consistent: What is supposed to be the same must indeed be the same. One must be careful when typing names, findings or other text, so that when you select similar cases or order your database using different text variables, no cases get excluded or wrongly assigned (due to these type or errors). It is always judicious to revise your data looking for misspellings, inappropriate capital letters, hidden spaces and other differences generated during data input that may result in posterior analysis errors. All spreadsheets have simple tools to check for these errors.

A final remark learned from personal experience: keep your data safe and trustable. It is simple and inexpensive to keep a backup of your database(s), and to use sequential copies as working data. Prepare chronological backups when appropriate, but always have at hand an updated file to refer to, because errors happen and computers sometimes behave funny. Besides, at publishing time, reviewers or editors may ask for additional calculations or a redoing of existing ones, or of tables or graphs, and it's a nightmare not to have a referenced and documented database to work with. And, finally, if an old dataset is needed to compare new data to, you will appreciate to have within reach a reliable, efficient and documented record of your research.

REFERENCES.

1. Thuissard-Vasallo IJ, Sanz-Rosa D. Manejo de variables en investigación clínica y experimental IV Jornadas de Investigación COEM-Universidades; Madrid, España. 2006.

2. Temple RJ. A regulatory authority's opinion about surrogate endpoints. In: Clinical Measurement in Drug Evaluation. New York, USA: John Wiley & Sons. 1995;3-22.