Rev Bras Cir Cardiovasc 2008; 23(3): 396-399

Critical reading of the statistical data in scientific studies

Leitura crítica dos dados estatísticos em trabalhos científicos

Mário José da CONCEIÇÃO, TSA[1]

RBCCV 44205-1006

Abstract

Objectives: **Statistics are a valuable tool that validates the conclusions of scientific works. The objective of this review was to present some concepts related to statistic calculations that are fundamental for the critical reading and analysis of medical literature.**

Contents: **In general, authors present the results of their studies as charts, boxes, and tables with quantitative data, along with descriptive statistics (means, standard deviations, medians), and almost always mention the statistic tests used. After reviewing several studies, it was difficult to find the value attributed to the statistical test. Thus, it is up to the reader to evaluate the adequacy of the information, and to search for evidence that contradict possible mistakes that could threaten the validity of their conclusion.**

Conclusions: **Examining the design of the studies one observes that, in many of them, excessive importance is given to statistical calculations as definitive factors, irrefutable evidence of arguable, or equivocal, conclusions.**

Descriptors: **Statistical analysis. Data interpretation, statistical. Statistical methods and procedures. Research design/ statistics & numerical data.**

Resumo

Objetivos: **A Estatística é ferramenta valorizada no testemunho da validade das conclusões dos trabalhos científicos. O objetivo dessa revisão foi apresentar alguns conceitos relacionados com os cálculos estatísticos que são fundamentais para a leitura e o pensamento críticos diante da literatura médica.**

Conteúdo: **Em geral, os autores apresentam os resultados de seus estudos na forma de gráficos, quadros e tabelas com dados quantitativos, acompanhados de estatísticas descritivas (médias, desvios-padrão, medianas) e quase sempre mencionando os testes estatísticos realizados. Após revisão, em inúmeros desses estudos, será difícil encontrar valor atribuível ao teste estatístico. Assim, fica ao leitor a tarefa de avaliar a adequação das informações e buscar as evidências contrárias aos possíveis erros que poderiam ameaçar a validade das conclusões.**

Conclusões: **Muitas vezes, pelo exame do desenho do estudo, observa-se o excessivo peso dado aos cálculos estatísticos como fatores definitivos, provas irrefutáveis, de conclusões discutíveis, quando não equivocadas.**

Descritores: **Análise estatística. Interpretação estatística de dados. Métodos e procedimentos estatísticos. Projetos de pesquisa /estatística & dados numéricos.**

INTRODUCTION

Statistics, or Biostatistics, as it is conventionally called when applied to biological sciences, is a valuable tool to validate the conclusions of scientific works. In general, authors present the results of their studies as charts, boxes, and tables with quantitative data, along with descriptive statistics (means, standard deviations, medians) and almost always they mention the statistical tests used in the analysis. Results

of those tests are presented as values of “p”. After reviewing several studies, it is difficult to find the value

1. Professor of Surgical and Anesthetics Techniques – FURB – Blumenau – SC; Member of Editorial Board of Revista Brasileira de Anestesiologia, Pediatric Anesthesia and Regional Anesthesia and Pain Medicine; Co-Responsible by the CET Integrado of SESSC – Florianópolis, SC - Brazil.

Correspondence address: Dr. Mário José da Conceição
Rua Germano Wendhausen, 32/401
88015-460 - Florianópolis, SC. Brazil.
E-mail: marioconceicao@uol.com.br

Article primarily publish on Revista Brasileira de Anestesiologia Vol. 58, No 3, May-June, 2008.
Reproduced with the authorization of the Publishers.
Descriptors and References adapted to the Norms of RBCCV (BJCVS).

given to the statistical tests used. Therefore, it is up to the reader to evaluate the adequacy of the information presented and look for evidence that contradict the possible mistakes that could threaten the validity of the conclusions.

Thousands of scientific works dedicated to the divulgation of studies in the field of anesthesia, and correlated fields, are published every year in hundreds of journals. Biostatistics is used by the majority of those studies, including both basic sciences and clinical studies, to validate their conclusions.

Examining the design of the study, one observes that excessive importance is given to statistical calculations as definitive factors, irrefutable evidence of arguable, and even mistaken, conclusions. The objective of this review was to present some concepts related with statistical calculations that are fundamental for the critical reading and analysis of medical literature.

The mistake of the equivalent test

Browsing through the pages of medical journals one will find on the Methods section of several articles the insistent presence of $p > 0.05$ or $p < 0.05$, which mean statistically non-significant and significant, respectively. On finding a $p > 0.05$ or < 0.05, the author fundaments all the importance of his/her study on the result of this calculation, and it is done in a way he/she considers brilliant, and concludes that the

phenomenon or fact being studied exists (or does not).

Apparently, this problem has worsened with the advent of computers and its charts, which include several statistical programs, facilitating considerably those calculations. This has made several statistical analyses available to authors. However, not every author is prepared to use them properly. Morphine is a potent dose-dependent respiratory depressant, regardless whether it is administered intravenously or in the neuroaxis; an unquestionable truth, at least up to now.

As an example, consider a study in which the summary of its method is as follows: two groups of patients treated with fixed doses of morphine administered in the neuroaxis.

Postoperatively, they are transferred to two different places: one group is transferred to the regular ward while the other group goes to the intensive care unit. The objective of the study was to evaluate the development of respiratory depression in patients treated with morphine and the difference between both groups. The study presented a result of $p > ?0.05$, i.e., without statistically significant differences between both groups. Based on this result, the authors concluded that patients treated with morphine administered in the neuroaxis, are not at risk for respiratory depression. A $p > 0.05$, "without statistically significant differences" suggests lack of evidence of an effect.

When one reads "statistically significant differences were not observed between both groups", one is not facing the complete information. The high value of p does not mean absence of an effect, as the authors wrongly concluded. It only means that the data was not enough to establish the need of postoperative observation of those patients. In other articles (very common on articles in English), for space purposes or any other reason, the authors omit the word "statistics" and write: "differences between groups were not observed" or "significant differences were observed between both groups". Differences of 5% can be clinically significant, but not statistically significant.

Going back to the example of morphine, if only one patient had developed respiratory depression requiring ventilator support, clinically it is highly significant, for obvious reasons. When reading a conclusion based on values of "p" (higher or lower), the reader should interpret only "statistical differences between groups". Incorrect, if not dangerous, would be to assume that there was equivalency between groups for a certain clinical occurrence observed [1].

Power of the sample

Since it is not feasible to study all individuals affected by the same phenomenon, one uses a group of individuals chosen from said population to represent it. This is called sample. Very often "p" is greater than 0.05 simply because the number of individuals in the study (sample) is too small. How many times one has read: "thirty patients randomly etc…"

In fact, Brazilian authors love "randomized" and "randomization". Statisticians call this a type II error; i.e., when one does not detect, in a given sample, the phenomenon studied when it does exist1. Several post-graduate thesis work with small samples due to the short time available until the end of the post-graduation course and the amount of data that has to be gathered to write the thesis. The probability that a study will detect the phenomenon studied when it exists is called "power". Power depends on group variability, size of the sample, the true nature of the phenomenon being observed, and the level of significance.

A good clinical study should inform the calculated power of the sample, so the reader can evaluate "non-statistically significant" results. It would be reasonable to think that respiratory depression, after the administration of morphine in the neuroaxis, did not manifest in 30 patients, but it could have developed on the 32nd patient if the study sample had 35 patients. The power of the sample is defined by a percentage. A sample can be 40% or 99% reliable to detect a phenomenon. Do not trust large samples.

This is a common mistake in scientific studies: the author(s) think that a huge sample (for example, 5,000 cases)

allows him to infer absolute results. Bigger is not always better when it comes to sample size.

Therefore, the author should, before starting the study, carefully plan the size of the study sample, to make sure it is appropriate for his objectives. Gathering 10,000 cases of anything is absolutely inappropriate1; and the result of all this effort? None.

Choosing the wrong statistical program

Statistical packages available in the market, or those associated with the charts included in computers, cannot prevent the researcher from using the wrong model or indicate the limitations of the program. For example, how many times, in the medical literature, has the Bonferroni test been used to validate the Analysis of Variance (ANOVA)? The Bonferroni test, or the Dunn's test for multiple hypotheses, dispenses ANOVA and was not idealized for post hoc (after the fact) comparisons, but for a priori tests. The wrong program can generate $p < 0.05$.

When reading clinical studies, one must pay attention when complex statistical tests indicate certain effects that simpler tests reject. It is necessary to understand whether the author describes carefully the model used (and why) or simply refers to an automatic method of selecting variables. It is not enough to mention the parameters that fed his program without the guarantee that he verified whether they were allocated correctly [2].

Evidence originated by several studies

One single article is not enough to make a decision about a phenomenon. It is very common to find several studies on the same subject with different conclusions. One study might present statistically significant differences, attesting the existence of a specific phenomenon, while two or three other studies present the opposite conclusion.

Those observations might be a consequence of mistakes incurred. As mentioned before, values of $p > 0.05$ do not guarantee the equivalence, but it indicates the lack of evidence of a statistically significant difference. To infer that the number of studies, pro and against the evidence, define the problem can also be a mistake. A comparison among studies might align, on the same level, studies that are not appropriate or whose method was not properly planned. Multicenter studies, which combine data from different places, are more trustworthy.

Statistically speaking, the advantage of multicenter studies, when compared with a single study, lies in the reduced confidence interval for a phenomenon [2]. In this context, one can argue the power of metanalysis to validate clinical observations.

Experts diverge on this theme. However, a metanalysis of small samples is hardly the same as a large clinical assay resulting from a multicentric study. Besides, metanalysis do not substitute well-planned clinical observations.

Balance between control and study groups

Most clinical studies, in our field, begin the description of results by comparing basic characteristics between two groups: gender, age, weight, and physical status, which are called "demographic data". The intention of the author is to demonstrate to readers that both groups are balanced. Very often, the value of "p" is added to test the difference between both groups. But mistakes can still be made. There are differences among groups of patients that can interfere with the results [3]. For example, observe Table I, which was extracted from an analysis of the effects of neuromuscular blockers in children. The authors assume (and also induce the reader) that those groups are perfectly homogenous. However, nothing is mentioned regarding their nutritional state or hydration status. Here, a $p < 0.05$ was interpreted as undeniable proof of the homogeneity of the groups, and that other parameters can be discarded, regardless of the study model.

Table 1. Characteristics of patients who received mivacurium after Atracurium (Group AM), Cisatracurium (Group CM) or Mivacurium (Group MM)

	Group AM	Group CM	Group MM
Age (yr)	5.4 (2.3 -12.5)	6.0 (2.3 – 12.0)	5.8 (2.6 – 12.9)
Weight (kg)	20.0 (10.3 – 40.0)	21.0 (13.5 – 55.0)	23.0 (14.0 – 56.0)

Data are presented as median (ranges).
N = 15 per group.
There were no statistically significant between-group differences

It is curious that the reciprocal can be true. There are methods that use a value of $p < ?0.05$ to prove the need to include other parameters. Returning to table I, $p < 0.05$ attests that the distribution of parameters was not luck or arbitrary.

However, under methods, the authors stated that distribution was at random; thus, it was "by luck". The mistake here lies in the certainty of the authors that $p < 0.05$ determines parameters that should (or should not) be included in the model (gender, age, weight) and which ones should be safely ignored (nutritional state, hydration). In the case of neuromuscular blockers, the nutritional state of the children could have, undeniably, interfered with the results, but it is probable that gender could not. It is common, among authors, to think that it is enough to mention that patients "were randomly selected".

On the design of the method, some parameters could have been ignored [3]; on the other hand, models with too many parameters are difficult to interpret and use. However, the author must explain the impact on the results of the variables excluded. This is called "sensitivity analysis". Results become convincible when properly presented.

More rigorous Editorial Boards ask the author to send this information, including the list of parameters from where the results were extracted, which causes indignation in many of them.

CONCLUSIONS

If one intends to read a scientific article critically, he/she needs to know only the basic principles of statistics. However, the following questions should be answered:

• Did the author provide information regarding the mean baseline parameters of the study groups?

• Did the author use confidence intervals on the description of the results, especially when no evidence was found?

• Are there inconsistencies between the information presented on charts and boxes and those in the body of the text?

• Is the interpretation of "p" values correct?

• Did the author use adjustment tests (Newmann-Keuls, Dunnet, and other) for multiple comparisons?

• Did the author justify adequately the statistical model used?

Complex models are not necessarily correct. One should be attentive for the problem of multiple comparisons with many statistical tests.

Note: Articles consulted as example of mistakes, were not included in the references due to ethical consideration with the authors. Articles in English, Spanish, and Portuguese were reviewed. Besides, the experience of the author on reviewing articles for publication in three journals was used.

For the same reason, the article in which Table I was published was not mentioned on the references.

REFERENCES

1. Abramson JH. Survey methods in community medicine: epidemiologic studies. 5th ed. New York:Churchill-Livingstone;1999. p.311-25.

2. Avram MJ, Shanks CA, Dykes MH, Ronai AK, Stiers WM. Statistical methods in anesthesia articles: an evaluation of two American journals during two six-month periods. Anesth Analg. 1985;64(6):607-11.

3. Dawson B, Trapp RG. Bioestatística básica e clínica. 3a ed. Rio de Janeiro:McGraw-Hill;2003.