Frequency of the Adequate Use of Statistical Tests of Hypothesis in Original Articles Published in the Revista Brasileira de Anestesiologia between January 2008 and December 2009

Fabiano Timbó Barbosa [1], Diego Agra de Souza [2]

Summary: Barbosa FT, Souza DA – Frequency of the Adequate Use of Statistical Tests of Hypothesis in Original Studies Published in the Revista Brasileira de Anestesiologia between January 2008 and December 2009.

Background and objectives: Statistical analysis is necessary for adequate evaluation of the original article by the reader allowing him/her to better visualize and comprehend the results. The objective of the present study was to determine the frequency of the adequate use of statistical tests in original articles published in the Revista Brasileira de Anestesiologia from January 2008 to December 2009.

Methods: Original articles published in the Revista Brasileira de Anestesiologia between January 2008 and December 2009 were selected. The use of statistical tests was deemed appropriate when the selection of the tests was adequate for continuous and categorical variables and for parametric and non-parametric tests, the correction factor was described when the use of multiple comparisons was reported, and the specific use of a statistical test for analysis of one variable was mentioned.

Results: Seventy-six original articles from a total of 179 statistical tests were selected. The frequency of the statistical tests used more often was: Chi-square 20.11%, Student *t* test 19.55%, ANOVA 10.05%, and Fisher exact test 9.49%. The frequency of the adequate use of statistical tests was 56.42% (95% CI 49.16% to 63.68%), erroneous use in 13.41% (95% CI 8.42% to 18.40%), and an inconclusive result in 30.16% (95% CI 23.44% to 36.88%).

Conclusions: The frequency of inadequate use of statistical tests in the articles published by the Revista Brasileira de Anestesiologia between January 2008 and December 2009 was 56.42%.

Keywords: ANESTHESIOLOGY; publication; STATISTIC: data interpretation; SCIENTIFIC METHODS: statistic.

[Rev Bras Anestesiol 2010;60(5): 528-536] ©Elsevier Editora Ltda.

INTRODUCTION

Readers of scientific journals should make a critical interpretation of the design and conduction of a study as well as the statistical analysis of the tests used in each study to interpret its results [1]. The literature has emonstrated that clinicians especially those who do not have a formal epidemiology and biostatistics education have a poor understanding of statistical tests and a limited ability to interpret the results of studies published in original articles [2].

A statistical analysis of the original article is necessary so the reader will have conditions to better visualize and understand the results, as well as understand how the data of the study were treated, although it is not always obligatory since some original articles are the result of qualitative or merely descriptive investigations. It is important that the statistical analysis be adequately selected and used in order to validate the results of each study. Other scientific journals have already performed the analysis of their material, and editors have an interest in improving their publications [3-6].

The objective of the present study was to determine the frequency of the adequate use of statistical tests in original articles published by the Revista Brasileira de Anestesiologia between January 2008 and December 2009.

METHODS

This study was submitted to the Ethics on Research Commission of the Universidade Estadual de Ciências da Saúde de Alagoas that consider an evaluation not necessary since this study involves public domain data. The informed consent does not apply. The expenses of this study were responsibility of the author. This is an observational transversal study undertaken from January to March of 2010.

Received from Universidade Estadual de Ciências da Saúde de Alagoas.

1. Master's Degree in Sciences from Universidade Federal de Alagoas, Professor of Basis of Surgical and Anesthetic Techniques of Universidade Federal de Alagoas
2. Medical student of Universidade Estadual de Ciências da Saúde de Alagoas, PhD student

Submitted on May 4, 2010.
Approved on May 16, 2010.

Correspondence to:
Dr. Fabiano Timbó Barbosa
Comendador Palmeira, 113, ap. 202
Farol
57051-150 – Maceió, AL, Brazil
Tel: (82) 9983-2054
E-mail: fabianotimbo@yahoo.com.br

The inclusion criterion was studies published in Revista Brasileira de Anestesiologia between January 2008 and December 2009. Studies other than original article, such as review articles, clinical information, case reports, miscellaneous articles, editorials, and letters to the editor were excluded. The study was considered original when it presented in its description the report of an investigation method only one or a set of results, and the interpretation and discussion of the results observed. The period from 2008 to 2009 was chosen since it includes the most recent original articles.

The primary variable of this study was the frequency of the adequate use of statistical tests of hypothesis used in the evaluation of the results. Secondary parameters included the frequency of: the use of statistical tests, the report of the exact value of "p" in the results, presence of descriptive statistics (mean, mode, median, standard deviation, amplitude, variance, standard error, percentile, and quartile), use of analysis in contingency tables (Chi-square, Fisher exact, McNemar, and Z tests), use of advanced statistical tests (logistic regression, Cox regression, univariate and multivariate linear model), frequency of original articles with the correct use of statistical tests, frequency of the use of confidence interval, description of the hypothesis, and description of the calculation of the sample size.

The use of statistical tests was considered adequate when:

- The selection of the tests was adequate for continuous and categorical variables and parametric and non-parametric tests.
- A description of the correction factor was present when the use of multiple comparisons was reported.
- The specific use of a statistical test for analysis of a variable was mentioned.

Analysis of the tests was inconclusive when:

- It was not possible to evaluate whether the distribution of continuous variables was normal or asymmetrical.
- Values of "p" were reported, but it was not specified which tests were used for each variable of the study.
- The use of tests and alpha value were mentioned, but in the results neither the value of "p" nor the tests were mentioned.

If the data had a normal distribution a parametric test was considered to be used correctly, but when this criterion was not achieved the use of a non-parametric test was considered adequate. The distribution of the data was considered normal when the author of the original article reported that the variable assumed a normal distribution; when the Kolmogorov-Smirnov, Shapiro-Wilk, and the D'Agostino-Pearson normality tests were used in the analysis of the data of the variable; by the observation of the relationship between mean and standard deviation; realization of the calculation of the variation coefficient; and by the analysis of charts demonstrated in the studies. The linear regression model was considered appropriate when used for continuous variables. The use of non-parametric tests was considered adequate for categorical variables.

Calculation of the size of the study population revealed the need to analyze 76 original articles considering a frequency of the adequate use of statistical tests of 70%, an absolute precision of 10%, and a level of significance of 5%[7]. Descriptive statistics, by means of simple frequency and 95% confidence interval for each estimated point, was used.

RESULTS

Seventy-six articles were selected and analyzed from volumes 58 and 59 of Revista Brasileira de Anestesiologia. Those two volumes contained a total of 179 statistical tests of hypothesis. Tables I and II show the results of primary and secondary variables.

Descriptive statistics was present in all articles reviewed. Only 10.52% (8/76) of the studies used only descriptive statistics.

Considering each study, 30.26% (23/76) used adequately all statistical methods, 22.36% (17/76) used incorrectly all statistical tests, and 28.94% (22/76) had inconclusive data. It also should be mentioned that 0.39% (3/76) of the studies considered correct were associated with inconclusive statistical tests, and 0.39% (3/76) with incorrect and inconclusive tests.

Table I – Frequency of the Use of Statistical Tests

Frequency of the tests, statistical methods, and regression methods		
	Percentage (%)	Absolute
χ^2 test	20.11	36
t test	19.55	35
ANOVA	10.06	18
Fisher	9.50	17
Mann-Whitney	7.82	14
Kruskal-Wallis	6.70	12
Wilcoxon	3.91	7
Kolmogorov-Smirnov	3.35	6
Linear multiple regression	2.79	5
Spearman correlation	1.68	3
ANOVA rep	1.68	3
Logistic regression	1.12	2
Tukey	1.12	2
Learning curve	1.12	2
Cronbach's alpha	1.12	2
Scheffé	1.12	2
Student-Newman-Keuls	1.12	2
Mood	1.12	2
Friedman	0.56	1
Simple linear regression	0.56	1
Kaplan Meier	0.56	1
CUSUM curve	0.56	1
Shapiro-Wilk	0.56	1
Bartlett	0.56	1
Kappa	0.56	1
L test	0.56	1
Log-Rank	0.56	1

Table II – Results of Primary and Secondary Variables: use of statistical tests of hypothesis

	Absolute value	Relative value	95% CI
Frequency of use of the statistical tests of hypothesis			
Adequate	101	56.42%	49.16% – 63.68%
Inadequate	24	13.41%	8.42% – 18.40%
Inconclusive	54	30.16%	23.44% – 36.88%
Description of the calculation of the size of the study population			
Yes	20	26.32%	16.42% – 36.22%
No	56	73.68%	63.78% – 83.58%
Description of the hypothesis of the study			
Yes	8	10.53%	3.63% – 17.43%
No	68	89.47%	82.57% – 96.37%
Description of the value of "p"			
Yes	63	82.89%	74.42% – 91.36%
No	13	17.11%	8.64% – 25.58%
Use of the CI			
Yes	10	13.16%	5.56% – 20.76%
No	66	86.84%	79.24% – 94.44%

DISCUSSION

The three steps to be considered as definition of the best test to be used in a statistical analysis include: analyze the question contained in the study, determine the level of data measurement, and define the best study design to elucidate the phenomenon or the data of the population of interest [7]. When a statistical test is erroneously used the results obtained may not be reproducible.

The classification of the original articles, taking into consideration the calculation of the size of the study population, demonstrated that 73.69% of the studies analyzed did not describe this calculation. The size of the sample has an inverse relationship with the value of "p" and vice-versa; therefore very large populations have a tendency for lower "p" values while very small populations might not indicate statistically significant differences [8]. The adequate size of the study population also allows to estimate expenses and minimize the use of interventions in a higher number than necessary to prove the study hypothesis [9]. The authors of the present study did not evaluate the effect of the results reported by the original articles in clinical practice, but the adequate use of statistical test for the variables presented by the authors. Readers should judge the validity of the results reported by the articles, but calculation of the sample size is an item that shows the quality of the study; therefore, when present, the results of the study gain more credit. Not reporting the calculation of the sample size should not be mistaken by inadequate use of statistical tests. When a study reports results without statistical significance that does not necessarily mean that the clinical effect investigated does not exist, but that the study might not have had enough statistical power to demonstrate it; for this reason, oftentimes studies from different areas of knowledge have phrases that focus indirectly on the importance of this calculation, such as "further studies are necessary" or "the study population was small to determine the difference".

The adequate use of statistical tests in the study population did not surpass the 70% assumed in the hypotheses of the present study and which was based on the international medical literature[7]. This finding can be justified by the fact that the majority of the mistakes in the use of tests observed in the present study was due to the use of the Student *t* test for small samples, in which the authors of the study did not consider that the data had a normal distribution, and by using a parametric test when a non-parametric test would have been more adequate. The results of the present study does not take away its credit for the scientific community, since the adequate use in international journals might not reach a mean of 30% [3-6].

Analysis of the frequency of the use of statistical tests demonstrated that the Student *t* test was the parametric test used more often. Besides, it made it clear that descriptive analysis was present in all studies. Those results corroborate other studies within and outside the intensive care field, which demonstrated that the Student *t* test and descriptive statistics are used more often in the studies [3-7,9]. Analysis of two independent groups is common in studies in the medical field and it might justify the greater frequency of the Student *t* test[10]. Descriptive statistics organizes and summarizes the data and it represents the final point of descriptive studies and the initial point of some studies before analytical tests of hypothesis are performed [11]. Descriptive statistics helps characterize the study populations and facilitates the perception of the reader regarding differences or similarities.

Analysis of the frequency of the use of statistical tests demonstrated that the most common tests used were the Student *t* test and Chi-square test. The Student *t* test is a parametric test that evaluates the mean of two groups when the data assumes a normal distribution [10]. The Chi-square test is used to evaluate proportions [7]. A limitation of the analysis of the adequate use of tests for contingency tables observed in this study was the difficulty to see in which situation the Chi-square and Fisher exact tests were used, since some studies described the use of both of them, but the results did not express in which variable one and the other test was used. Descriptions like “the Chi-square test was used” or “Fisher exact test was used whenever appropriate” made it impossible to analyze the adequate use of those tests. Authors should be encouraged to give a more clear description regarding the use of each test because it would make it easier for readers to interpret the results as well as the perception about validation of the data.

The real value of “p” was present in 81.57% of the original articles that used statistical tests. The value of “p” demonstrates the magnitude of the statistical significance; however, the investigator should demonstrate the clinical importance of the results observed [9,12]. Using just the reference value of “p” described in the “methods” section to report the results of a study hinders the critical analysis of said study; therefore, results followed by the expressions $p > 0.05$ or $p < 0.05$ should be avoided.

The description of the confidence interval was present in 13.15% of the original articles analyzed. It is more practical to present statistical samples as estimates of the result that should have been obtained if the entire population had been investigated; however, the lack of precision that results from the degree of variability of the factor under investigation and the limited size of the study population might influence the results [13]. A better estimate of the result could be demonstrated by the confidence interval [13]. This interval could be seen as a summary of the results, for some statistical tests, and it has proven to be more informative than the result regarding the null hypothesis [14]. The confidence interval presents the advantage of having statistical significance, demonstrating a band of values in which the true populational value may take into consideration a certain level of confidence [13,14]. It is more advantageous to the reader to present the results of “p”, as well as the confidence interval, than to present just one of those measurements, making interpretation of the results more logic.

A study published in the decade of 1980 demonstrated that approximately half of the studies published in the medical field used statistical tests erroneously, and the Student *t* test was responsible for the majority of the mistakes [15]. Some rules have been stipulated so readers can estimate whether statistic methods were used adequately: know the difference between standard deviation and standard error of the mean, understand the meaning of “p”, and recognize a common error in the use of the *t* test. Standard deviation shows how distant the values observed are from the mean, since adding or subtracting the value of a standard deviation from the mean, one has the distribution of 68% of the data. The use of standard error demonstrates the homogeneity of the data that might not be real. The value of “p” represents the probability of a result having occurred by chance, even if it is not present in the population the sample originated from. The *t* test should be used to compare two means and not for double means, since this increases the chances of finding clinically important results.

The frequency of the adequate use of statistical tests in original articles published in Revista Brasileira de Anestesiologia between January 2008 and December 2009 was 56.42%.

REFERÊNCIAS / REFERENCES

01. Windish DM, Hout SJ, Green ML – Medicine residentes' understanding of the biostatistics and results in the medical literature. JAMA, 2007;298:1010-1022.
02. Wullf HR, Anderson B, Brandenhoff P et al. – What do doctors know about statistics? Stat Med, 1987;6:3-10.
03. Avram MJ, Shanks CA, Dykes MH et al. – Statistical methods in anesthesia articles: an evaluation of two American journals during two six-month periods. Anesth Analg, 1985;64:607-611.
04. Hokanson JA, Luttman DJ,Weiss GB – Frequency and diversity of use of statistical techniques in oncology journals. CancerTreat Rep, 1986;70:589-594.
05. Cardiel MH, Goldsmith CH – Type of statistical techniques in rheumatology and internal medicine journals. Rev Invest Clin, 1995;47:197-201.
06. Huang W, LaBerge JM, Lu Y et al. – Research publications in vascular and interventional radiology: research topics, study designs, and statistical methods. J Vasc Interv Radiol, 2002;13:247-255.
07. Kurichi JE, Sonnad SS – Statistical Methods in the Surgical Literature. J Am Coll Surg, 2006;202:476-484.
08. Cavalcanti AB, Akamine N, Sousa JMA – Avaliação Crítica da Literatura. em: Knobel E – Condutas no Paciente Grave. 3ª ed. São Paulo, Atheneu, 2006;2635-2647.

09. Barbosa FT, Jucá MJ – Avaliação da qualidade dos ensaios clínicos aleatórios em anestesia publicados na revista brasileira de anestesiologia no período de 2005 a 2008. Rev Bras Anestesiol, 2009;59:223-233.
10. Gaddis GM, Gaddis ML – Introduction to biostatistics: part 4, statistical inference techniques in hypothesis testing. Ann Emerg Med, 1990;19:820-825.
11. McHugh ML – Descriptive statistics, part I: level of measurement. J Spec Pediatr Nurs,, 2003;8:35-37.
12. Gonçalves GP, Barbosa FT, Barbosa LT et al. – Avaliação da qualidade dos ensaios clínicos aleatórios em terapia intensiva. Rev Bras Ter Intensiva, 2009;21:45-50.
13. Gardner MJ, Altman DG – Confidence intervals rather than P values: estimation rather than hypothesis testing. BMJ, 1886;292:746-750.
14. Thompson WG – Statistical criteria in the interpretation of epidemiologic data. Am J Publ Health, 1987;77:191-194.
15. Glantz SA – Biostatistics: how to detect, correct and prevent errors in the medical literature. Circulation, 1980;61:1-7.

Resumen: Barbosa FT, Souza DA – Frecuencia del Uso Adecuado de los Test Estadísticos en los Artículos Originales Publicados en la Revista Brasileña de Anestesiología entre enero de 2008 y diciembre de 2009.

Justificativa y objetivos: La realización de un análisis estadístico se hace necesario para una evaluación pertinente del artículo original por parte del lector, ayudándolo a obtener una mejor visualización y comprensión de los resultados. El objetivo de esta investigación fue determinar la frecuencia del uso adecuado de los test estadísticos de hipótesis presentes en los artículos originales publicados en la Revista Brasileña de Anestesiología, entre enero de 2008 y diciembre de 2009.

Métodos: Se seleccionaron artículos originales publicados en la Revista Brasileña de Anestesiología entre enero de 2008 a diciembre de 2009. El uso de los test estadísticos se evaluó como apropiado cuando: la selección de los test fue satisfactoria para las variables continuas y categóricas y para el test paramétrico y no paramétrico; hubo una descripción del factor de corrección cuando se relató el uso de múltiples comparaciones; fue mencionado el uso específico de un test estadístico para el análisis de una variable.

Resultados: Se seleccionaron 76 artículos originales, con un total de 179 test estadísticos de hipótesis. La frecuencia de los test estadísticos más utilizados fue: 20,11% para el Chi-Cuadrado, 19,55%, para el test *t* de *Student*, 10,05% para el test de ANOVA y 9,49% para el test exacto de Fisher. La frecuencia de uso adecuado de los test estadísticos de hipótesis fue de un 56,42% (IC 95% 49,16% a 63,68%), de uso inadecuado, 13,41% (IC 95% 8,42% a 18,40%), con un resultado sin conclusiones en un 30,16% (IC 95% 23,44% a 36,88%).

Conclusiones: La frecuencia del uso adecuado de los test estadísticos utilizados en los artículos originales publicados en la Revista Brasileña de Anestesiología entre enero de 2008 y diciembre de 2009, fue de un 56,42%.