Acta Odontologica Scandinavica. 2015; 73: 76–80

informa
healthcare

ORIGINAL ARTICLE

The statistical reporting quality of articles published in 2010 in five dental journals

HANNU VÄHÄNIKKILÄ[1], LEO TJÄDERHANE[1,2] & PENTTI NIEMINEN[3]

[1]*Institute of Dentistry, University of Oulu, Oulu, Finland,* [2]*Medical Research Center Oulu, Oulu University Hospital and University of Oulu, Oulu, Finland, and* [3]*Medical Informatics and Statistics Group, University of Oulu, Oulu, Finland*

Abstract

Objectives. Statistical methods play an important role in medical and dental research. In earlier studies it has been observed that current use of methods and reporting of statistics are responsible for some of the errors in the interpretation of results. The aim of this study was to investigate the quality of statistical reporting in dental research articles. ***Methods.*** A total of 200 articles published in 2010 were analysed covering five dental journals: *Journal of Dental Research, Caries Research, Community Dentistry and Oral Epidemiology, Journal of Dentistry* and *Acta Odontologica Scandinavica.* Each paper underwent careful scrutiny for the use of statistical methods and reporting. A paper with at least one poor reporting item has been classified as 'problems with reporting statistics' and a paper without any poor reporting item as 'acceptable'. ***Results.*** The investigation showed that 18 (9%) papers were acceptable and 182 (91%) papers contained at least one poor reporting item. ***Conclusions.*** The proportion of at least one poor reporting item in this survey was high (91%). The authors of dental journals should be encouraged to improve the statistical section of their research articles and to present the results in such a way that it is in line with the policy and presentation of the leading dental journals.

Key Words: *articles, journals, statistics reporting, review literature*

Introduction

Statistical methods play an important role in medical research. The use of statistical methods in medical research articles have increased over recent decades [1,2]. The same phenomenon has also been seen in dental research [3]. In consequence of the wide use of statistical methods, the risks of poor reporting and methodological errors have increased [4–6]. The existence of these problems is often regarded as evidence that poor reporting slips through the peer review process [7]. Studies with poor statistical analysis and reporting may lead to incorrect conclusions, artificial research results and waste of valuable resources. Misuse of statistical methods can also have serious clinical consequences because evidence supporting or opposing the research hypothesis is not correctly attained or reported [8].

Poor reporting and statistical errors are very difficult to define. Several studies have examined the acceptability of statistical methods in the medical literature [9–11], but there is no standard form or guideline to evaluate the poor reporting and misuse of statistical methods. To our knowledge, there has been only one study in dental literature which has been done regarding the prevalence of misused statistics or statistical errors [6]. Previous studies have mainly concentrated to misuse of statistics and have not evaluated the quality of reporting statistical aspects.

Our aim in this study was to assess the quality of statistical reporting in dental literature; not to judge the journals which were under consideration, but to draw attention to this important but sometimes neglected part of reporting research results.

Correspondence: Hannu Vähänikkilä, Institute of Dentistry, University of Oulu, Box 5281, FIN-90014, Finland. Tel: +358 50 3438266. Fax: +358 8 537 5560. E-mail: hannu.vahanikkila@oulu.fi

(Received 10 March 2014; accepted 1 August 2014)

ISSN 0001-6357 print/ISSN 1502-3850 online
DOI: 10.3109/00016357.2014.954612

Materials and methods

Sample of papers

Five dental journals were selected for the evaluation: *Journal of Dental Research* (JDR), *Journal of Dentistry* (JD), *Caries Research* (CR), *Community Dentistry and Oral Epidemiology* (CDOE) and *Acta Odontologica Scandinavica* (AOS). JDR is one of the most visible and cited dental journals (IF = 3.773), JD is a European journal that has regularly published research carried out using time-to-event methods (IF = 2.115), while CR has the highest impact factor of any journal specifically concentrating on caries research (IF = 2.926). CDOE represents an epidemiological approach to dental issues (IF = 2.328) and AOS is a European dental journal that is mainly directed at the Nordic countries (IF = 1.130). The journals were chosen to enable us to make valid inferences about the wide range of statistical reporting of dentistry. JDR, JD and CR are among 15 highest-ranking journals in dentistry, oral surgery and medicine according to the impact factor, while AOS clearly has a lower impact factor. As a journal concentrating on epidemiology, CDOE could be expected to have articles with better statistical analysis.

Several articles have reviewed medical papers and tried to find errors in the statistical procedures [1,12]. The reported proportion of erroneous articles has been ~30–50%. Presuming 40% error rate in dental papers, a sample size of 100 articles was calculated as a minimum number for the present purpose, allowing a maximum difference of 10% between the sample rate and true population rate at a 95% significance level. However, we anticipated that 20 articles per journal would be too low a number of papers to make comparisons between the journals and chose to double it to 40 articles per journal. Letters, brief reports, narrative reviews and editorials were excluded from this sample. The starting articles were chosen randomly, the only criteria being that there would be at least 39 eligible subsequent articles published that year in the journal in question. For example, CR published 56 articles during 2010, so there were 17 possible articles to start the evaluation to fulfil the goal. The total number of papers fulfilling criteria were for JDR 189, JD 119, CR 56, CDOE 58 and AOS 44. The total number of articles reviewed was 200 and each paper underwent careful scrutiny for the use of statistical methods and reporting.

Classification of study design

The papers were classified into three groups: experimental studies (including randomized clinical trials and laboratory works), non-experimental studies (including cross-sectional surveys, cohort studies and case-control studies) and others (including reliability, methodology and basic science studies).

Classification of poor reporting

The following aspects were recording as statistical error or poor reporting as defined by Lang and Secic [13], Altman [1] and Nieminen et al. [12]:

(1) Incomplete description of procedures (failure to specify all methods used, wrong names for statistical methods, misuse of technical terms such as quartile);
(2) Statistical software not reported;
(3) Sample size not reported;
(4) No justification for the sample size reported (power calculations, sample size calculations, budget restrictions);
(5) Inexact p-values reported (p = ns, $p > 0.05$, $p = 0.0000$);
(6) Problems in tables or figures (insufficient title or legends, numbers do not sum up, no percentages, scale problems, using pie chart in describing continuous variables); and
(7) Data dredging with multiple comparisons (number of p-values over 50 or number of p-values greater than the number of subjects in the study).

A paper with at least one poor reporting item was classified as 'problems with reporting statistics' and a paper without any poor reporting item was classified as 'acceptable'.

Evaluation of the papers and statistical analysis

All 200 papers were reviewed manually by one of the authors (HV) to assess their statistical methods used and reporting. The reliability of the evaluation method has previously been shown to be good (Kappa coefficient of reliability between raters ranged from 0.65–1.00) [3] and the use of several raters was, thus, considered unnecessary. SPSS for Windows version 18.0 (SPSS Inc., Chicago, IL) was used for data handling.

Results

Most of the articles were experimental studies; 24% were randomized clinical trials and 26% were laboratory works. Only two articles were meta-analyses. The study design of all other articles was non-experimental (cross-sectional = 22.5%, cohort study = 11.5%, case-control study = 4%) and other study design = 11.0% of the articles.

Of the 200 papers published in 2010, 182 (91.0%) had problems in statistical reporting and the proportion of the journals varied between 82.5–97.5% of the papers. The most common poor reporting item was the lack of justification for the number of cases

Table I. Percentage distribution of reporting items in five dental journals.

Item	AOS	CDOE	CR	JD	JDR	Total
Incomplete description of procedures						
Yes	25.0	17.5	32.5	42.5	45.0	32.5
No	75.0	82.5	67.5	57.5	55.0	67.5
Software						
Not reported	20.0	20.0	17.5	47.5	55.0	32.0
Reported	80.0	80.0	82.5	52.5	45.0	68.0
Sample size						
Not reported	7.5	0	0	15.0	20.0	7.5
Reported	92.5	100	100	85.0	80.0	92.5
Justification for number of cases						
No	77.5	70.0	70.0	95.0	87.5	80.0
Yes	22.5	30.0	30.0	5.0	12.5	20.0
Inexact *p*-values						
Yes	40.0	12.5	42.5	30.0	20.0	29.0
No	60.0	87.5	57.5	70.0	80.0	71.0
Problems in tables and figures						
Yes	7.5	10.0	15.0	0	7.5	8.0
No	92.5	90.0	85.0	100	92.5	92.0
Data dredging with multiple comparisons						
Yes	12.5	32.5	7.5	0	12.5	13.0
No	87.5	67.5	92.5	100	87.5	87.0

AOS, Acta Odontologica Scandinavica; CDOE, Community Dentistry and Oral Epidemiology; CR, Caries Research; JD, Journal of Dentistry; JDR, Journal of Dental Research.

(80.0%). If this item is left out, the proportion of at least one error decreased to 60.5%. About one third of the papers had an incomplete description of statistical procedures or the statistical software was not reported. Sample size was quite well reported in all journals (80–100% of the articles) and there were only a few problems with tables and figures (0–15% of the articles). The items with notable variation between the journals were the incomplete description of procedures (17.5–45%) and data dredging with multiple comparisons (0–32.5% of the articles). Inexact *p*-values were reported in 29% of the articles (Table I). The distribution of the number of different types of problems in each article is illustrated in Figure 1.

Discussion

The proportion of at least one poor reporting item in this survey was high (91%). However, if the justification for the number of cases is left out, the proportion decreased to 60.5%. Our finding of poor quality of statistical reporting is not directly comparable with a previous study in dental literature [6], because we

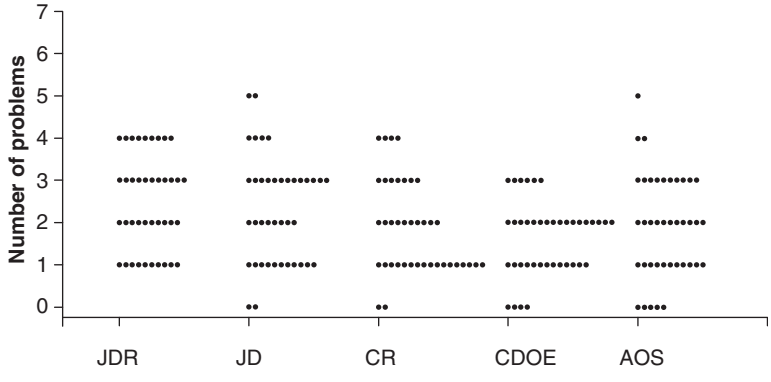

Figure 1. Distribution of the number of different types of problems in each article.

concentrated on reporting statistics, whereas their goal was to assess the level of statistical misuses and abuses in dental literature. The reason that we did not take statistical errors into account was that, if the original data is not available, it is difficult to define if the correct statistical methods are used or right statistical test is used. Instead we analysed how the methods and results were reported, which are not dependent on the availability of the data used.

In our study we have considered these reporting errors as equally serious. However, some of the errors could be judged to be a more serious reporting error (e.g. not mentioning the sample size vs not mentioning the statistical software used). Altman [14] has also discussed this topic and categorized statistical errors in study design, analyses, presentation and interpretation into three groups: definite errors, matter of judgements and only poor reporting. While the aim of our study was to assess the quality of statistical reporting, Altman [14] has more emphasis on study design, analysis and interpretation of the results. Thus, it cannot be used as such to categorize our findings, but could serve as guidelines for researchers when planning the study and reporting the results.

The statistical procedures of the evaluated articles were mainly described adequately. A sufficient description of the methods used is essential for a research paper to be used and cited frequently [15] and for it to be included in systematic reviews or meta-analyses. When a published study does not clearly report its methods, a reader attempting to evaluate its scientific validity may rely largely on the authors' reputation and style of writing or simply on the journal's reputation. Statistical tests and methods should, therefore, be identified in the methods section [13]. It is also useful to report the statistical software used in multivariate analyses because different software packages use different calculation algorithms [16]. The reporting of software also helps critical readers to evaluate and understand details that are specific to certain statistical software packages.

Sample size was usually reported in the evaluated papers. However, sample size calculations were seldom reported in the evaluated articles set. This problem is common in medical research. Altman [14] reported a review of 100 consecutive papers published in the *British Medical Journal* in 1991–1993 (excluding controlled trials) and noted that no information on sample size calculations was provided in any of them. This is an important issue, as p-values evaluating statistical significance are highly dependent on sample size and the validity of statistical inferences can also be affected by a small number of events per variable. To evaluate the findings, it is beneficial for the reader to know the justifications for the sample size selected. Mathematical justifications for the intended sample size using statistical power calculations are valuable [17]. Textbooks of medical statistics simply require that the sample size should be large enough (or as large as possible) [4].

Inexact p-values (e.g. $p < 0.01$, $p > 0.05$) were often represented in our review. This is not statistical error, but hides evidence from the reader. As more accurate information, it is advisable to provide exact p-values, say $p = 0.429$. When p-value is represented by $p > 0.05$, it implies that the p-value of the test is anywhere between 0.05–1.0. It would be important to report the exact level of significance for the reader so that the reader can also evaluate the statistical significance of the findings (as this would determine whether or not to reject or accept the null hypothesis at the given level [6]). There is no reason that the p-value must be degraded into this less-informative dichotomy [18].

Figures are visual means of conveying information and have a strong visual impact. Tables are commonly used for two purposes: to present background information related to methods and to present data. A well-structured table is perhaps the most efficient way to convey a large amount of data in a scientific paper [19]. In our material there were only a few problems in the figures and tables. The most typical problems in the figures were that there was an insufficient legend, there were no measurement units and there were scale problems. The most typical problems in the tables were that numbers did not sum up and there were no percentages represented. When the results are presented in figures or tables and statistical significances are evaluated, it is helpful for readers if these differences also are indicated in tables or figures. The mode of presentation should be selected according to the data and depends on the purpose and type of information being displayed. This way the reader can quickly understand the key point and find the outcome of interest [19].

The statistical analysis of studies often includes several significance tests, which increases the risk of making a type I error, such as saying that a treatment is effective when chance is a more likely explanation for the results. This is often referred to as the multiple testing problem. It may encounter when testing several baseline characteristics for differences between groups, performing multiple pair-wise comparisons, testing multiple end-points, performing secondary analyses, performing sub-group analyses not planned in the original study, comparing groups at multiple time points or performing step-wise regression analyses. These 'data dredging' analyses involve computing many p-values to find something that is statistically significant [20].

If studies generate hundreds of p-values, interpreting multiple p-values is difficult. If authors made many comparisons, it may be reasonable to expect some p-values to be small just by chance. We think that this is poor reporting and it is difficult to know how to deal with reports that do not concentrate on

the planned main outcomes. To make sense of this kind of study, readers need to look at the overall pattern of results and not interpret any individual *p*-value too strongly.

Most of the shortcomings in the reporting of statistical information in the articles reviewed here are related to topics included even in most introductory medical statistics books. Apparently these problems in one of the key issues in high-quality research are so common and wide-spread that actions should be taken to improve the statistical reporting. For example, the 'Instructions to Authors' should highlight the importance of the use of appropriate statistical methods and the accuracy of their description in the manuscript. The journals should also inform the authors of the guidelines developed for different kinds of study protocols [7], either describing the guidelines or providing the reference to the most recent version in the 'Instructions to Authors'. The reviewers could be separately asked to evaluate the quality and reporting of the issues dealing with the statistical analysis. The editors and editorial boards should also consider consulting a statistician either with every manuscript with statistical analysis or at least when the reviewer raises questions or doubts about the quality and/or reporting of the statistical analysis. We also propose that the manuscript should include a statement to clearly state who is responsible for the statistical analysis and the quality of the statistical reporting.

It must take into consideration that some journals (e.g. JDR) place page or word restrictions on manuscript submissions, which could contribute to some authors' decisions to minimize wording relative to statistical issues.

The authors of dental journals might apply these results to improve the statistical section of their research articles and to present the results in such a way that it is in line with the policy and presentation of the leading dental journals.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

[1] Altman DG. Statistics in medical journals: some recent trends. Stat Med 2000;19:3275–89.
[2] Horton NJ, Switzer SS. Statistical methods in the journal. N Engl J Med 2005;353:1977–9.
[3] Vähänikkilä H, Nieminen P, Miettunen J, Larmas M. Use of statistical methods in dental research: comparison of four dental journals during a 10-year period. Acta Odontol Scand 2009;67:206–11.
[4] Nieminen P, Miettunen J, Koponen H, Isohanni M. Statistical methodologies in psychopharmacology. Hum Psychopharmacol 2006;21:195–203.
[5] Baccaglini L, Shuster JJ, Cheng J, Theriaque DW, Schoenbach VJ, Tomar SL, et al. Design and statistical analysis of oral medicine studies: common pitfalls. Oral Dis 2010;16:233–41.
[6] Kim JS, Kim DK, Hong SJ. Assessment of errors and misused statistics in dental research. Int Dent J 2011;61:163–7.
[7] Altman DG. Poor-quality medical research: what can journals do? JAMA 2002;287:2765–7.
[8] Gardenier JS, Resnik DB. The misuse of statistics: concepts, tools, and a research agenda. Account Res 2002;9:65–74.
[9] Gore SM, Jones IG, Rytter EC. Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976. Br Med J 1977;278:85–7.
[10] Gardner M, Bond J. An exploratory study of statistical assessment of papers published in the British Medical Journal. J Am Med Assoc 1990;263:1355–7.
[11] Welch G, Gabbe S. Review of statistics usage in the American Journal of Obstetrics and Gynecology. Am J Obstet Gynecol 1996;175:1138–41.
[12] Nieminen P, Carpenter J, Rucker G, Schumacher M. The relationship between quality of research and citation frequency. BMC Med Res Methodol 2006;6:42.
[13] Lang T, Secic M. How to report statistics in medicine. Philadelphia: American College of Physicians; 1997.
[14] Altman DG. Statistical reviewing for medical journals. Stat Med 1998;17:2661–74.
[15] Nieminen P, Rucker G, Miettunen J, Carpenter J, Schumacher M. Statistically significant papers in psychiatry were cited more often than others. J Clin Epidemiol 2007;60:939–46.
[16] Okunade AA, Chang CF, Evans RD. Comparative analysis of regression output summary statistics in common statistical packages. Am Stat 1993;47:298–303.
[17] Friedman LM, Furberg CD, DeMets DI. Fundamentals of clinical trials. St Louis: Mosby; 1996.
[18] Rothman KJ. Epidemiology. An introduction. Oxford: Oxford University Press; 2012. p 152.
[19] Iverson C, Christiansen S, Flanagin A, Fontanarosa PB, Glass RM, Gregoline B, et al. AMA manual of style: a guide for authors and editors. 10th Ed. New York: Oxford Press; 1997.
[20] Motulsky H. Intuitive biostatistics. Oxford: Oxford University Press; 1995. p 117–26.