European Journal of Orthodontics 33 (2011) 22–25
doi:10.1093/ejo/cjq025
Advance Access Publication 14 July 2010

Appropriateness of reporting statistical results in orthodontics: the dominance of *P* values over confidence intervals

Argy Polychronopoulou*, Nikolaos Pandis** and Theodore Eliades***
*Department of Preventive and Community Dentistry, School of Dentistry, University of Athens, **Private Practice, Corfu and ***Department of Orthodontics, School of Dentistry, Aristotle University of Thessaloniki, Greece

Correspondence to: Theodore Eliades, 57 Agnoston Hroon, Nea Ionia 14231, Greece. E-mail: teliades@athforthnet.gr

SUMMARY The purpose of this study was to search the orthodontic literature and determine the frequency of reporting of confidence intervals (CIs) in orthodontic journals with an impact factor. The six latest issues of the *American Journal of Orthodontics and Dentofacial Orthopedics*, the *European Journal of Orthodontics*, and the *Angle Orthodontist* were hand searched and the reporting of CIs, *P* values, and implementation of univariate or multivariate statistical analyses were recorded. Additionally, studies were classified according to the type/design as cross-sectional, case–control, cohort, and clinical trials, and according to the subject of the study as growth/genetics, behaviour/psychology, diagnosis/treatment, and biomaterials/biomechanics. The data were analyzed using descriptive statistics followed by univariate examination of statistical associations, logistic regression, and multivariate modelling.

CI reporting was very limited and was recorded in only 6 per cent of the included published studies. CI reporting was independent of journal, study area, and design. Studies that used multivariate statistical analyses had a higher probability of reporting CIs compared with those using univariate statistical analyses. Misunderstanding of the use of *P* values and CIs may have important implications in implementation of research findings in clinical practice.

Introduction

In most research studies, where comparisons are made between groups, some form of statistical analysis is performed and a test or a number of tests of significance are reported with corresponding *P* values. The *P* value shows the probability of observing the recorded treatment effect/difference or a more extreme one between the study groups when in reality no difference exists between these groups, i.e., when the null hypothesis is true. If the observed *P* value is small enough, the null hypothesis may be rejected and it may be said that there is evidence that there is a true difference between the study groups. To use a practical example to clarify the meaning of the *P* value, let us assume that a study has been conducted on a population sample and a difference has been found in treatment duration between two bracket systems, A and B, of 20 per cent with a calculated *P* value of 0.01. The observed difference of 20 per cent in this particular trial would occur, just by chance, in only one out of 100 identical studies when $P = 0.01$; a rather unlikely outcome resulting in the conclusion that there is a true difference in treatment duration between the two bracket systems.

In order to find the real difference in effect between study groups, the whole target population must be studied. However, since this is impossible, statistical inference are made using a small, representative sample of the target population in order to draw conclusions about the whole. *P* values, although they may be indicative of a statistically significant result, provide limited insight into the clinical relevance of the findings. The more clinically relevant and important data obtained from the study results would be the actual difference/effect size and its range (Gardner and Altman, 1986; Goodman, 1999). Small *P* values depend heavily on large sample sizes and low variances, whereas their correlation with the observed effect size and its clinical importance is limited. For example, if a two-arm parallel trial with 1000 patients in each arm is run in order to evaluate differences in treatment duration between two bracket systems, A and B, and additionally (it is assumed that), the mean treatment duration for the first group A is 500 days, whereas for the second group B it is 510 days and the standard deviation (SD) of the mean treatment duration is 50 days for both groups, a *t*-test between the two groups would give a highly significant result ($P < 0.001$). However, a difference of 10 days [95 per cent confidence interval (CI): 5.6–14.4 days] in treatment duration is clinically not important.

The problem in publications is related to the fact that frequently only *P* values are reported and used in order to draw conclusions concerning treatment effectiveness disregarding the size of the effect, its range, and the clinical importance of the observed results. A more appropriate presentation of the trial results would focus on the size of the difference between the treatment groups and its range, i.e. the CI. As mentioned previously in the example trial, the mean difference in treatment duration between the study groups is 10 days and the 95 per cent CI of the difference is 5.6–14.4 days. Relying only on the *P* value of the trial would lead to the conclusion that bracket system A is superior to bracket system B; however, reporting the actual difference in days (10 days) and the 95 per cent CI (5.6–14.4 days) would result in a different interpretation of the results.

It has long been recognized that over-reliance on *P* values when presenting and interpreting results is inappropriate and often misleading (Rothman, 1978; Mainland, 1984). There is a tendency for the result to be interpreted in terms of significance or non-significance based solely on *P* values. This way, any significant result, regardless of its clinical importance or plausibility, is considered important, whereas any non-significant result regardless of its clinical importance it is considered as indicating 'no difference of effect" (Simon, 1986; Savitz, 1993; Barnett and Mathisen, 1997; Chia, 1997). CIs show the range of the plausible difference of effect/association between study groups that helps to determine whether the observed differences are suggestive of true benefits and superiority of one treatment over the other, and offer valuable information that may be adopted by the clinician and used to make clinical decisions. The meaning of the 95 per cent CI level is as follows: if 100 samples are drawn from the target population, 95 per cent of them would contain the true population value. CIs always contain the effect point estimate and depending on the confidence level, usually set at 95 per cent, we are confident, at a defined level (95 or 99 per cent, etc.), that they contain the true population value. Increased sample size only narrows the width of the CIs around the same size of effect, thus increasing precision, unlike in the case of *P* values where increasing the sample size lowers the *P* value. Reporting CIs moves the interpretation of the results from the dichotomy of significant/non-significant to the size of the effect/association and its range of plausible values given by the data under study.

There is a lack of studies evaluating the quality of statistical reporting in the orthodontic literature and whether only *P* values and/or CIs are included in the results. The objective of this study was to search the most recent orthodontic journals with an impact factor and evaluate the frequency of CI reporting and its possible associations with publication characteristics such as journal, study type, study subject, and use of univariate or multivariate statistical analyses.

Materials and methods

The following three orthodontic journals were included in the study:

1. American Journal of Orthodontics and Dentofacial Orthopedics (AJODO).
2. Angle Orthodontist (AO).
3. European Journal of Orthodontics (EJO).

The content of the six most recent issues of the journals published up to July 2009 was hand searched by one author (NP).

During the first screening, all editorials and letters were excluded and the following guidelines were used in order to classify article-reporting characteristics:

1. Case reports, reviews, and descriptive articles, where there were no statistical comparisons for the main research question, were excluded from the statistical analysis.
2. The four broad subject categories were (1) craniofacial growth, morphology, and genetics (growth/genetics); (2) behaviour and psychology (behaviour/psychology); (3) diagnostic procedures, aesthetics, treatment, and oral hygiene (diagnosis/treatment); and (4) biomaterials and biomechanics (biomaterials/biomechanics).
3. No distinction was made for either animal or human clinical studies.
4. Multivariate analyses were considered as the studies' analyses where two or more variables were used as predictors (model driven or through stratification).
5. Reporting of *P* values and/or notation of significance or non-significance was treated as the same.
6. CIs recording was carried out for treatment group differences, and not for within groups summary values.

In total, 378 articles were examined; 101 were eventually excluded for not adhering to the pre-determined criteria, leaving 277 to be included in the data analysis. The data were processed and analysed by means of Stata® 10.0 version software (Stata Corporation, College Station, Texas, USA). The level of statistical significance for all tests was set at 0.05. Initial data analysis relied on descriptive statistics; subsequent univariate examination of statistical associations was conducted using Pearson chi-square or Fisher's exact test. Logistic regression analysis was performed for effect estimation, whereas simultaneous investigation of a number of predictors was accomplished through multivariate modelling.

Results

All examined studies provided statistical analysis with *P* values or notation of significance/non-significance; however, only 17 (6 per cent) presented CIs. All studies providing CIs presented point effect estimates; five studies reported odds ratios, one study relative risk, and 11 absolute risk (difference). Stratification and multivariate analysis accounting for possible confounders were observed in 32 studies (11.5 per cent).

Table 1 shows the distribution of 277 orthodontic articles by journal, subject area, study design, and type of statistical analysis. Table 2 presents the results of univariate and multivariate logistic regression modelling. Univariate analysis revealed that the AO and the EJO showed an increased probability of publishing an article reporting CIs compared with the AJODO; however, this finding did not achieve statistical significance. Also, non-growth/genetic subjects showed a non-significant decreased probability of reporting CIs compared with investigations in the growth/

Table 1 Distribution of 277 orthodontic articles by journal, subject area, study type, statistical analysis, and confidence interval (CI) reporting.

		CI reporting			
		Total	No	Yes	
Variable	Category	*n*	*n* (%*)	*n* (%*)	*P* value
Journal	*American Journal of Orthodontics and Dentofacial Orthopedics*	70	67 (95.71)	3 (4.29)	NS**
	Angle Orthodontist	134	126 (94.03)	8 (5.97)	
	European Journal of Orthodontics	73	67 (91.78)	6 (8.22)	
Subject area	Growth/genetics	32	29 (90.63)	3 (9.38)	
	Behaviour/psychology	11	10 (90.91)	1 (9.09)	NS***
	Diagnosis/treatment	159	147 (92.45)	12 (7.55)	
	Biomaterials/biomechanics	75	74 (98.67)	1 (1.33)	
Study type	Cross-sectional	134	127 (94.78)	7 (5.22)	NS**
	Case–control	9	8 (88.89)	1 (11.11)	
	Cohort	62	57 (91.94)	5 (8.06)	
	Clinical trial	72	68 (94.44)	4 (5.56)	
Statistical analysis	Univariate	245	235 (95.92)	10 (4.08)	$<10^{-3}$
	Stratified/multivariate	32	25 (73.13)	7 (21.88)	
	Total	277	260 (93.86)	17 (6.14)	

NS: non-significant; *Row percentage; **Pearson chi-square test; ***Fisher's exact test.

Table 2 Logistic regression modelling-derived odds ratios (OR) and confidence intervals (CIs) for a CI-reporting article finding over a CI-non-reporting article, by a series of publication characteristics (*n* = 277).

		Univariate model			Adjusted model*		
Variable	Category or increment	OR	95% CI	*P* value	OR	95% CI	*P* value
Journal	*American Journal of Orthodontics and Dentofacial Orthopedics*	Baseline					
	Angle Orthodontist	1.41	0.36–5.52	NS			
	European Journal of Orthodontics	2.00	0.48–8.32	NS			
Subject area	*Growth/genetics*	Baseline					
	Behaviour/psychology	0.13	0.01–1.30	NS			
	Diagnosis/treatment	0.78	0.20–2.97	NS			
	Biomaterials/biomechanics	0.96	0.09–10.39	NS			
Study type	Cross-sectional	Baseline					
	Case–control	2.26	0.24–20.75	NS			
	Cohort	1.59	0.48–5.22	NS			
	Clinical trial	1.06	0.30–3.77	NS			
Statistical analysis	Univariate	Baseline			Baseline		
	Stratified/multivariate	6.58	2.30–18.80	$<10^{-3}$	6.60	2.25–21.46	0.001

NS: non-significant; *Model included journal, subject area, study type, and statistical analysis type.

genetics category, and a etiologic study type (case–control, cohort, intervention) showed an increased, but still non-significant, probability of reporting CIs compared with cross-sectional investigations. Studies employing multivariate or stratified analysis had an increased, statistically significant probability, of publishing CIs compared with those presenting only univariate analysis. Multivariate logistic regression modelling revealed an increased probability of reporting CIs after multivariate analysis was used compared with when univariate statistical analysis was employed even after accounting for the possible confounding effect of journal, study type, and subject.

Discussion

The CIs take a range of values, which is believed to include the 'true' population value, with a defined level of certainty, and they represent the precision of the outcome and are a function of the sample size, the variability of the characteristic being studied, and the selected level of confidence. Smaller sample sizes are associated with greater standard errors and wider CIs, leading to lower precision of the results. The *P* value indicates the strength of the evidence against the null hypothesis, when the null hypothesis is true, but *P* values give no indication of the direction, size, or precision of the effect (Rothman *et al.*, 2008). On the contrary, CIs shift the interpretation from a qualitative judgment to a quantitative estimation of the effect.

The standardization and quality of reporting of biomedical research is an important obligation of the health care community. Efforts have been made and guidelines have been published referring to quality reporting (Bailar and Mosteller, 1988; International Committee of Medical Journal Editors, 1997; Altman, 2000). One important set of guidelines is included in the Consolidated Standards of Reporting Trials (CONSORT) statement (Altman *et al.*, 2001).

The CONSORT guidelines require that for each outcome, study results should be reported as a summary of the outcome in each group (for example, mean, proportion) together with the effect size (risk ratio or relative risk, odds ratio, risk difference, hazard ratio or difference in median survival time, and difference in means). CIs should be presented for the difference rather than separately for the outcome in each group. Presentation of CIs is especially valuable when non-significant differences are found and therefore a judgment based on clinical relevance could be made on the importance of even the non-significant difference of effect.

The interpretation of *P* values derived from statistical testing most often becomes a qualitative one where the study results are presented as either being significant or not significant, whereas the CIs provide a range of values within which the true difference of the study groups is believed to exist, thus giving the reader the opportunity to interpret the results in relation to clinical practice. Furthermore, *P* values have no units whereas CIs are in the units of the dependent variable, a fact that makes interpretation of the results easier. Freiman *et al.* (1978) re-analysed 71 negative studies, based on significance testing, using CIs for study result interpretation. The re-analysis using CIs indicated that probably many of the treatments were beneficial and a focus on CI interpretation rather than *P* values would have indicated this effect. On the contrary, Vavken *et al.* (2009), in a recent publication, found the CI reporting in orthopaedic research is around 20 per cent. Additionally, they found that the probability of statistically significant results predicting at least a 10 per cent between-group difference was only 69 per cent (95 per cent CI: 55–83 per cent), indicating that a high proportion of statistically significant results do not reflect large treatment effects. The use of CIs could help avoid such erroneous results.

The findings of the present investigation show that there is very limited adoption of CI reporting in the orthodontic literature and that there is no evidence that journal type, and consequently impact factor, type of study, and study subject are significant predictors of CI reporting. On the contrary, studies where more advanced statistical analyses were used show a higher probability of CI reporting. This seems logical since complicated analyses are most likely to be performed by more experienced investigators in statistics compared with simple analyses.

The limited reporting of CIs in the major orthodontic journals possibly indicates that there is misunderstanding within the orthodontic community regarding the misinterpretations associated with *P* values and CIs. These findings may have important implications on the interpretation of orthodontic research and the extrapolation of the results into clinical practice.

Conclusions

The results of this investigation of orthodontic research articles published in three orthodontic journals suggest that:

1. Reporting of CIs in orthodontic journals with an impact factor is limited (6 per cent).
2. CI reporting was independent of specific journal, subject area, or study design.
3. Studies that use multivariate statistical analyses have a higher probability of including CIs compared with those using univariate statistical analyses.

References

Altman D G 2000 Confidence intervals in practice. In: Altman D G, Machin D, Bryant T N, Gardner M J (eds). Statistics with confidence: confidence intervals and statistical guidelines. 2nd edn. British Medical Journal Books, London, pp. 6–14.

Altman D G *et al.* 2001 CONSORT Group (Consolidated Standards of Reporting Trials). The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Annals of Internal Medicine 134: 663–694

Bailar J C III, Mosteller F 1988 Guidelines for statistical reporting in articles for medical journals. Amplifications and explanations. Annals of Internal Medicine 108: 266–273

Barnett M L, Mathisen A 1997 Tyranny of the *p*-value: the conflict between statistical significance and common sense. Journal of Dental Research 76: 534–536

Chia K S 1997 'Significant-itis'—an obsession with the *P*-value. Scandinavian Journal of Work Environmental Health 23: 152–154

Freiman J A, Chalmers T C, Smith H Jr, Kuebler R R 1978 The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 'negative' trials. New England Journal of Medicine 299: 690–694

Gardner M J, Altman D G 1986 Confidence intervals rather than *p* values: estimation rather than hypothesis testing. British Medical Journal (Clinical Research Education) 292: 746–750

Goodman S N 1999 Toward evidence-based medical statistics I. The *P* value fallacy. Annals of Internal Medicine 130: 995–1004

International Committee of Medical Journal Editors 1997 Uniform requirements for manuscripts submitted to biomedical journals. Annals of Internal Medicine 126: 36–47

Mainland D 1984 Statistical ritual in clinical journals: is there a cure? British Medical Journal 288: 841–843

Rothman K J 1978 A show of confidence. New England Journal of Medicine 299: 1362–1363

Rothman K J, Greenland S, Lash T L 2008 Modern epidemiology. Lippincot Williams and Wilkins, Philadelphia, pp. 156–162.

Savitz D 1993 Is statistical significance testing useful in interpreting data? Reproductive Toxicology 7: 95–100

Simon R 1986 Confidence intervals for reporting results of clinical trials. Annals of Internal Medicine 105: 429–435

Vavken P, Heinrich K M, Koppelhuber C, Rois S, Dorotka R 2009 The use of confidence intervals in reporting orthopaedic research findings. Clinical Orthopedics and Related Research 467: 3334–3339