European Journal of Orthodontics, 2018, 1–7
doi:10.1093/ejo/cjy029

OXFORD

Original article

Statistical testing against baseline in orthodontic research: a meta-epidemiologic study

Sophia Gratsia[1], Despina Koletsi[2,3], Padhraig S. Fleming[4] and Nikolaos Pandis[5,6]

[1]School of Dentistry, National and Kapodistrian University of Athens, Greece, [2]Clinic of Orthodontics and Paediatric Dentistry, Center of Dental Medicine, University of Zurich, Switzerland, [3]Private Practice in Athens, Greece, [4]Department of Orthodontics, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, UK, [5]Department of Orthodontics and Dentofacial Orthopedics, Dental School/Medical Faculty, University of Bern, Switzerland and [6]Private Practice in Corfu, Greece

Correspondence to: Despina Koletsi, 5 Kanari St, 15127 Melissia, Attica, Greece. E-mail: d.koletsi@gmail.com

Summary

Background/objectives: To assess the prevalence of within-group comparisons from baseline to follow-up in published orthodontic articles and to identify potential associations between this statistical problem and a number of study characteristics.

Materials/method: The most recent 24 issues of four leading orthodontic journals with highest impact factor (American Journal of Orthodontics and Dentofacial Orthopedics; AJODO, European Journal of Orthodontics; EJO, Angle Orthodontist; ANGLE, Orthodontics and Craniofacial Research; OCR) were electronically searched until December 31st 2017. The proportion of articles using comparisons against baseline and interpretation of findings according to within-group comparisons were recorded. The association of this practice with journal, year of publication, study design, continent of authorship, number of centres and researchers, statistical significance of results, and statistical analysis was tested. Univariable and multivariable modified Poisson regression were used to identify significant predictors.

Results: Overall, 339 articles were eligible for inclusion with the majority published in ANGLE ($n = 157$, 46%), followed by AJODO ($n = 75$, 22%), and EJO ($n = 75$, 22%). A total of 60 studies (18%) presented interpretation of their findings based on within-group comparisons against baseline in isolation. Statistical significance of the primary outcome was a very strong predictor of the prevalence of this flawed approach (RR: 2.33, 95% CIs: 1.22, 4.43; $P = 0.01$).

Limitations: The effect of time since publication was not addressed.

Conclusions/implications: Statistical testing and interpretation within groups is prevalent in orthodontic research. Endorsement of accurate conduct and reporting of statistical analyses and interpretation of research findings is important in order to promote optimal inferences to support clinical decision-making.

Introduction

Rationale

Methodological and reporting flaws are endemic in medical and dental research with orthodontic research also afflicted by both conduct and reporting limitations (1–3). Reporting guidelines have been endorsed in an attempt to promote clear and optimal reporting in order to raise correct inferences from research in support of clinical decision-making (4, 5).

Although the use of reporting guidelines has received increasing awareness over the years and their adoption has been actively implemented both by journal editors and the wider research community in medicine as well as dentistry, areas of obscure and substandard reporting persist (6–9). Statistical analysis is not immune to these shortcomings with areas of particular concern including over-reliance on *P*-values, while disregarding precision of the effect as represented by confidence intervals (1); erroneous selection of statistical methods to analyze the data (6, 10); and inappropriate handling of correlated data (9, 11).

Recently, an important problem with regard to statistical handling and interpretation of study findings has been identified in the field of oral medicine (12). Statistical testing within-groups and against baseline has been shown to generate inappropriate inferences and lead to erroneous interpretation of research findings (12, 13). Specific related problems include confounding of the outcome due to natural improvement over time or regression toward the mean (14) as both have been linked to potential changes over time irrespective of the intervention or exposure; other problems comprise multiple testing and increased likelihood of false positive errors (inflated type I errors). Intuitively, when conducting an experiment to examine the effectiveness or safety of one intervention over another, inferences should be based upon statistical testing on their difference. Examination of whether treatment effects within each intervention group in isolation is significant when compared with its own before treatment baseline value, risks incorrect inferences particularly in comparative research.

Moreover, to better illustrate reliance on statistical significance one may consider the following example. Imagine a study that investigates the effectiveness of either headgear or Class II elastics in the reduction of overjet. The authors do not examine the difference in treatment effectiveness between the two strategies but rather the reduction of overjet within each group; they come up with a reduction in overjet of 4 mm in the first group (*P*-value = 0.049) and 4.1 mm in the second group with a *P*-value of 0.051. The authors may erroneously conclude that treatment with headgear is more effective than Class II elastics based on the observed *P*-values. However, the absence of evidence for the latter together with evidence for the former does not imply evidence for difference and may lead to erroneous inferences. In addition, direction of the effect is a parameter that should also be considered. A similar magnitude of a non-significant effect in two treatment groups receiving different interventions is not indicative of absence of a between-group difference, since the effect might be of the opposite direction and particularly strong. Unfortunately, those erroneous practices (15) and interpretations can be transferred to meta-analyses.

Findings from published empirical data in dentistry indicate that nearly a quarter of studies involve interpretation of data based solely on within-group comparisons and changes from baseline to follow-up, while observational studies were found to be particularly prone to this error (12). Previous original reports from biomedical research have included analysis of the field of Neuroscience with approximately 15 per cent of related publications being affected (6). To date, this methodological issue has not been evaluated specifically within orthodontic research.

Objectives

Therefore, the aim of this meta-epidemiological report was to examine the presence of this statistical error in orthodontic journals and to identify possible associations of this practice with a range of study characteristics.

Materials and methods

Adapted PRISMA guidelines were followed for the present meta-epidemiologic study (16). No registered protocol exists.

The content of the most recent 24 consecutive issues from 4 major orthodontic journals with the highest impact factor were electronically searched by one author (SG) until December 2017 to identify publications including measurements over time that could potentially present within group comparisons against baseline. The journals searched were: American Journal of Orthodontics and Dentofacial Orthopedics (AJODO), European Journal of Orthodontics (EJO), Angle Orthodontist (ANGLE), Orthodontics, and Craniofacial Research (OCR).

Eligibility criteria

All original studies involving measurements over time [either comparisons with baseline (two time-points) or assessment of data at more than two time-points], were considered eligible for inclusion excluding editorials, case reports, opinion letters, and reviews. Single arm trials or cohort studies without a comparison group were also excluded. Included studies were categorized according to design as interventional or observational in human, while laboratory or animal studies were included separately and specifically recorded as such.

Study selection and data collection process

Data acquisition and recording was performed on pre-specified standardized piloted forms and calibration between two researchers (SG, DK) was undertaken prior to data extraction on 20 articles. Inter-examiner reliability was assessed on 15 additional papers. For each study, changes from baseline to follow-up (within-group comparison) or otherwise together with interpretation of study results were recorded. Judgment of interpretation of the findings from each article was based on specific parts of the discussion section pertaining to reporting of implications of the results and the conclusion section in both abstract and main manuscript. Only when there were clear indications that the authors had based the narration and the presentation of their findings primarily on within-group comparisons, were the manuscripts categorized as bearing this type of misconduct. Furthermore, study characteristics such as journal, year of publication, continent of authorship, number of centres and researchers involved, statistical significance of results (based on the primary outcome), statistical analysis used, and reporting of confidence intervals were recorded. Statistical analysis was recorded for the primary outcome and, if more than one analyses were reported, the most complex was selected, corresponding to the primary outcome.

Summary measures and synthesis of results

Descriptive statistics were performed for a range of study characteristics. To test the association of overall interpretation based on changes from baseline to follow-up with study characteristics, chi-square tests and Fisher's exact test were undertaken, as appropriate. Univariable and multivariable modified Poisson regression with robust standard errors (SE) for binary data was performed to assess the effect of study characteristics including journal of publication, study design, and statistical significance of outcomes on overall interpretation of the findings based on within-group comparisons with baseline. Predictors with $P > 0.10$ in their univariable analysis were excluded from the multivariable model. The Hosmer–Lemeshow test was used to check model fit. The unweighted kappa statistic was used to assess inter-rater agreement with regard to overall interpretation of the findings based on within group comparisons. All statistical analyses were conducted with Stata version 15.1 software (Stata Corporation, College Station, Texas, USA).

Results

Study selection and characteristics

A total of 1164 articles were initially identified, of which 339 were eligible for inclusion after consideration of inclusion and exclusion criteria (Figure 1). Reliability assessment yielded an unweighted kappa statistic of 0.88 for the outcome of interest (ie. overall interpretation of the findings based on within group comparisons), reflecting excellent agreement between the two reviewers. Overall, the highest percentage of the assessed articles were published in ANGLE (157/339, 46%), followed by AJODO (75/339, 22%) and EJO (74/339, 22%) and within the years 2016 (99/339, 29%) and 2017 (107/339, 32%). Most articles originated from Asia/Other (146/339, 43%), consisted of multi-centre efforts (210/339, 62%) and were authored by 4–6 researchers (195/339, 58%). The highest percentage of studies were either interventional (161/339, 47%) or observational (123/339, 37%) in design, while *in vitro* (28/339, 8%) and animal studies (27/339, 8%) were under-represented in the present sample. Statistically significant findings for the main outcome were found for the majority of the studies (230/339, 68%) (Table 1).

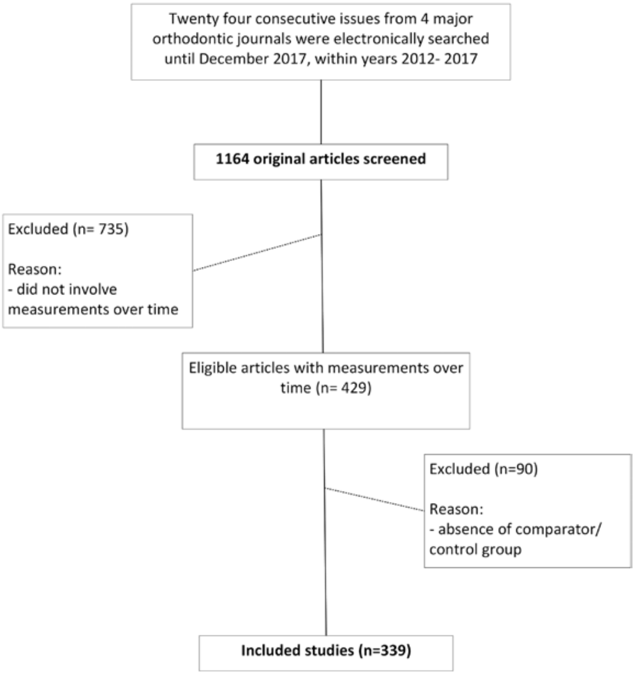
Twenty four consecutive issues from 4 major orthodontic journals were electronically searched until December 2017, within years 2012- 2017
1164 original articles screened
Excluded (n= 735)
Reason:
- did not involve measurements over time
Eligible articles with measurements over time (n= 429)
Excluded (n=90)
Reason:
- absence of comparator/ control group
Included studies (n=339)

Figure 1. Flow diagram of study selection.

Synthesis of results

Nearly one-fifth of the studies (n = 60, 18%) presented the interpretation of their results based on within-group comparisons for changes from baseline to follow-up (Appendix 1). Of those, one-third (n = 20, 33%) involved studies that conducted the statistical analysis solely within groups, while the rest (n = 40, 67%) presented analysis both within- and between-groups to evaluate treatment effects (Tables 1 and 2).

In vitro (8/28, 29%) and observational studies (27/123, 22%) revealed the highest percentage of this statistical problem; studies reporting statistically significant findings for the main outcome were also more likely to have this flaw (n = 50, 22%; P = 0.005; Table 1). Overall, univariable and multivariable Poisson regression showed strong evidence of association between reporting of statistically significant results and overall interpretation based on within-group comparisons (multivariable: RR: 2.33; 95% Confidence Intervals (CIs): 1.22, 4.43; P = 0.01; Table 3; Figure 2).

In the included studies where interpretation based on within-group comparisons against baseline was used, the most common statistical tests used were paired *t*-test (32/60, 53%). None of the studies reported Confidence Intervals for the estimated effect, although two-thirds of these articles (n = 40) involved analyses based on both within- and between-group comparisons (Table 4).

Discussion

Summary of evidence

The findings of this study are in keeping with previous research (6,12) with almost one-fifth of studies involving presentation of data based on within-group comparisons and changes from baseline to follow-up. To our knowledge, this was only the third meta-epidemiological study in biomedical fields concerning the prevalence of testing changes from baseline within groups and interpretation of data according to within-group comparisons. It was not surprising that empirical data from previous research in dentistry particularly related to orthodontic articles published 4–6 years ago revealed similar proportions of the presence of this statistical misconduct, although the previous study was based on just a single orthodontic journal (6). The relatively high prevalence of this statistical flaw within orthodontic research is indicative of the need for improvement in the statistical analysis of research data and interpretation of the results, in common with other research fields (6, 12).

Of those studies including testing changes from baseline within groups, it was somewhat encouraging that two-thirds did also incorporate between groups testing; as such, it is likely that the research question was addressed fully within the latter studies. No association of this statistical flaw with publication characteristics such as type of journal, continent of authorship, and number of authors, number of centres or type of study could be confirmed. Only studies with statistically significant results were more likely to base overall interpretation on comparisons from baseline to follow-up. Furthermore, testing and interpretation against baseline may reflect researchers' tendencies to consider statistically significant findings more important than non-significant ones (17, 18). This may represent a scenario, where between groups comparisons are non-significant while testing for changes from baseline to follow-up provides significant associations. Consequently, the former might be selectively withheld from publication or obscured somewhat, while the latter might be over-emphasized risking publication and selective reporting bias.

Table 1. Frequency distribution for the overall interpretation based on comparisons against baseline or otherwise by article characteristic (*n* = 339).

	Overall interpretation based on within-group comparison with baseline			*P*-value
	No *N* (%)	Yes *N* (%)	Total *N* (%)	
Year				0.47#
2012	4 (100)	0 (0)	4 (100)	
2013	6 (100)	0 (0)	6 (100)	
2014	52 (90)	6 (10)	58 (100)	
2015	51 (78)	14 (22)	65 (100)	
2016	80 (81)	19 (19)	99 (100)	
2017	86 (80)	21 (20)	107 (100)	
Journal				0.73*
AJODO	64 (85)	11 (15)	75 (100)	
ANGLE	130 (83)	27 (17)	157 (100)	
EJO	58 (78)	16 (22)	74 (100)	
OCR	27 (82)	6 (18)	33 (100)	
Continent				0.14*
America	66 (78)	19 (22)	85 (100)	
Europe	86 (80)	22 (20)	108 (100)	
Asia/other	127 (87)	19 (13)	146 (100)	
No. authors				0.74*
1–3	75 (83)	15 (17)	90 (100)	
4–6	158 (81)	37 (19)	195 (100)	
≥7	46 (85)	8 (15)	54 (100)	
No. centres				0.96*
Single centre	106 (82)	23 (18)	129 (100)	
Multi-centre	173 (82)	37 (18)	210 (100)	
Study category				0.09*
Observational	96 (78)	27 (22)	123 (100)	
Interventional	139 (86)	22 (14)	161 (100)	
In vitro	20 (71)	8 (29)	28 (100)	
Animal	24 (89)	3 (11)	27 (100)	
Significance				0.005*
No	99 (91)	10 (9)	109 (100)	
Yes	180 (78)	50 (22)	230 (100)	
Total	279 (82)	60 (18)	339 (100)	

*Pearson chi-square.

#Fisher's exact test.

Table 2. Frequency and percentage reporting of comparisons/interpretations within-group and against baseline or otherwise for the included studies (*n* = 339).

	N	%
Comparison		
Against baseline	20	6
Between groups	17	5
Both	302	89
Interpretation		
Based on group comparison against baseline	60	18
Based on between group comparison	81	24
Both	198	58
Total	339	100

A range of statistical tools are available for modelling panel data (longitudinal data with measurements over time) and this was recorded in the present study. Although simple statistics have always been rather straightforward for analysis in before-after studies, using paired *t*-tests, Wilcoxon signed rank tests or similar, more sophisticated analyses have also been endorsed to account for between-group comparisons and time-related or repeated measurements associations (19, 20). Analysis of covariance, repeated measures analysis of variance, linear and generalized linear mixed models, and generalized estimating equations (GEEs), may be used to adjust for baseline or treat baseline as another level of the time factor. Modelling differences in changes for between-group comparisons may also be easier to interpret alongside within-group comparisons, whether these within-group analyses are performed within the same model or separately. Furthermore, it is important to note that there are circumstances where intra-group comparisons may provide useful information, in isolation. An example is growth studies that assess certain population characteristics and no treatment effects are sought, or pilot studies with a comparison group where between-group comparisons are not necessarily powered for. In such cases, within-group differences for the intervention group may be desirable in an attempt to justify further investigation. It therefore appears that statistical testing and interpretation of comparisons from baseline to follow-up in isolation are prevalent in several medical domains within original research articles. The downstream use of these comparisons in future meta-analyses may lead to distorted impressions of treatment effects and incorrect inferences.

Table 3. Univariable and multivariable modified Poisson regression with Relative Risks (RR) and associated 95% confidence intervals (CIs) for the effect of a range of article characteristics on overall interpretation of the findings based on statistical testing against baseline (*n* = 339).

Category	Univariable			Multivariable		
	RR	95% CI	*P*-value	RR	95% CI	*P*-value
Journal			0.73*			
AJODO	Reference					
ANGLE	1.17	0.61, 2.24				
EJO	1.47	0.73, 2.96				
OCR	1.24	0.50, 3.07				
Continent			0.15*			
Asia/other	Reference					
America	1.72	0.96, 3.06				
Europe	1.57	0.89, 2.75				
No. authors			0.74*			
1–3	Reference					
4–6	1.14	0.66, 1.97				
≥7	0.89	0.40, 1.96				
No. centres			0.96			
Single centre	Reference					
Multi-centre	0.99	0.62, 1.59				
Study category			0.10			0.15*
Interventional	Reference					
Observational	1.61	0.96, 2.68		1.55	0.93, 2.58	
In vitro	2.09	1.03, 4.23		1.83	0.88, 3.82	
Animal	0.81	0.26, 2.53		0.70	0.23, 2.16	
Significance			0.008			0.01
No	Reference					
Yes	2.37	1.25, 4.50		2.33	1.22, 4.43	

*Wald test for the overall association.

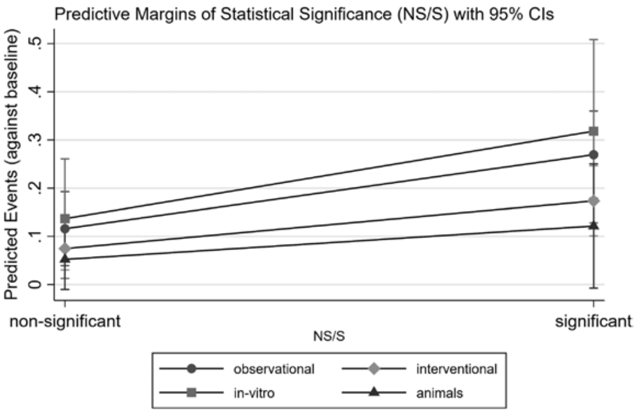

Figure 2. Predictive margins for overall interpretation of findings according to comparisons against baseline based on statistical significance and type of study.

Reporting guidelines have been regarded to mitigate against research reporting limitations having been widely endorsed to promote clear and accurate reporting in different types of research studies including clinical trials, observational research, and systematic reviews (4, 21–23). However, initiatives directed towards improving specific aspects of conduct and reporting within a research article, such as statistical analysis have been developed only relatively recently (24). Moreover, the latter are also less well-known among researchers, risking particularly poor levels of compliance. In addition, these initiatives have gained less traction among journal editors and peer reviewers. As such, the journals contributing the sample of the study are not known to follow specific guidelines with regard to statistical analyses or to use dedicated statistical reviewing as common practice for submitted articles. Notwithstanding this, compliance with reporting and conduct guidelines are known to be suboptimal even among those journals endorsing recognized guidelines necessitating the development of tailored approaches to enhance reporting of research both in orthodontics and other fields of research (25, 26).

Strengths and limitations

A potential limitation of the present meta-epidemiologic research was the inclusion of articles based on a subgroup of four orthodontic journals. Moreover, the effect of time or year since publication was not assessed by the final analysis model and the association between chronological year and the statistical problem of changes from baseline to follow-up or trend of this association could not be addressed. The selection of the four orthodontic journals included in the present work was based on the recent impact factor of these journals; this ranking is known to be dynamic. Notwithstanding this, it is likely that this cross-section is indicative of best practice within orthodontic research. In addition, the most recent issues of these were selected in order to reflect the current status of reporting quality in orthodontic literature; this approach has been common to most meta-epidemiologic research (7, 9).

Conclusions

Based on the present cross-section of four leading orthodontic specialty journals, statistical testing and interpretation within groups appears to be prevalent in orthodontic research, although the majority of studies (67%) incorporating within-group testing do also

Table 4. Type of statistical method used and reporting of confidence intervals (CIs) for the included articles, based on overall interpretation of comparisons against baseline or otherwise ($n = 339$).

	Overall interpretation based on comparison against baseline			
	No		Yes	
	N	%	*N*	%
Statistical analysis				
ANOVA	79	28	6	10
Repeated measures ANOVA/Friedman	15	5	4	7
Chi-square	7	2	1	2
Linear regression	7	2	2	3
Logistic regression	1	1	1	2
Mixed models	35	13	5	8
t-test	106	38	7	12
Paired *t*-test	27	10	32	53
Unclear	2	1	2	3
Reporting of CIs for between group comparison				
No	219	79	55	92
Yes	40	14	0	0
Only CIs against baseline	20	7	5	8
Total	279	100	60	100

Analysis of variance (ANOVA) category includes k-way ANOVA, multivariate analysis of variance, and nonparametric ANOVA; *chi-square* category includes χ^2, Fisher's exact test, Homogeneity test and McNemar's test; *mixed models* category includes mixed models (random effects, hierarchical models); *t*-test category includes independent and non-parametric equivalents (e.g. Mann–Whitney); paired *t*-test includes paired and non-parametric equivalents (Wilcoxon signed rank); *unclear*: use of statistical test not clearly stated.

include more relevant between-group analyses. The promotion of accurate conduct and reporting of statistical analyses is important in order to promote optimal inferences to support clinical decision-making; consequently, further work is required in order to improve the statistical rigor of orthodontic research outputs.

Conflict of Interest

None to declare.

References

1. Gardner, M.J. and Altman, D.G. (1986) Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal (Clinical Research Edition)*, 292, 746–750.
2. Fleming, P.S., Buckley, N., Seehra, J., Polychronopoulou, A. and Pandis, N. (2012) Reporting quality of abstracts of randomized controlled trials published in leading orthodontic journals from 2006 to 2011. *American Journal of Orthodontics and Dentofacial Orthopedics*, 142, 451–458.
3. Koletsi, D., Spineli, L.M., Lempesi, E. and Pandis, N. (2016) Risk of bias and magnitude of effect in orthodontic randomized controlled trials: a meta-epidemiological review. *European Journal of Orthodontics*, 38, 308–312.
4. Moher, D., Hopewell, S., Schulz, K.F., Montori, V., Gøtzsche, P.C., Devereaux, P.J., Elbourne, D., Egger, M. and Altman, D.G. (2010) CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ (Clinical research ed.)*, 340, c869.
5. Glasziou, P., Meats, E., Heneghan, C. and Shepperd, S. (2008) What is missing from descriptions of treatment in trials and reviews? *BMJ (Clinical research ed.)*, 336, 1472–1474.
6. Nieuwenhuis, S., Forstmann, B.U. and Wagenmakers, E.J. (2011) Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience*, 14, 1105–1107.
7. Neville, J.A., Lang, W. and Fleischer, A.B. Jr. (2006) Errors in the Archives of Dermatology and the Journal of the American Academy of Dermatology from January through December 2003. *Archives of Dermatology*, 142, 737–740.
8. Spanou, A., Koletsi, D., Fleming, P.S., Polychronopoulou, A. and Pandis, N. (2016) Statistical analysis in orthodontic journals: are we ignoring confounding? *European Journal of Orthodontics*, 38, 32–38.
9. Fleming, P.S., Koletsi, D., Polychronopoulou, A., Eliades, T. and Pandis, N. (2013) Are clustering effects accounted for in statistical analysis in leading dental specialty journals? *Journal of Dentistry*, 41, 265–270.
10. Kurichi, J.E. and Sonnad, S.S. (2006) Statistical methods in the surgical literature. *Journal of the American College of Surgeons*, 202, 476–484.
11. Koletsi, D., Pandis, N., Polychronopoulou, A. and Eliades, T. (2012) Does published orthodontic research account for clustering effects during statistical data analysis? *European Journal of Orthodontics*, 34, 287–292.
12. Koletsi, D., Madahar, A., Fleming, P.S. and Pandis, N. (2015) Statistical testing against baseline was common in dental research. *Journal of Clinical Epidemiology*, 68, 776–781.
13. Bland, J.M. and Altman, D.G. (2011) Comparisons against baseline within randomised groups are often used and can be highly misleading. *Trials*, 12, 264.
14. Bland, J.M. and Altman, D.G. (1994) Regression towards the mean. *BMJ (Clinical research ed.)*, 308, 1499.
15. Austin, P.C. and Hux, J.E. (2002) A brief note on overlapping confidence intervals. *Journal of Vascular Surgery*, 36, 194–195.
16. Murad, M.H. and Wang, Z. (2017) Guidelines for reporting meta-epidemiological methodology research. *Evidence Based Medicine*, 22, 139–142.
17. Koletsi, D., Karagianni, A., Pandis, N., Makou, M., Polychronopoulou, A. and Eliades, T. (2009) Are studies reporting significant results more likely to be published? *American Journal of Orthodontics and Dentofacial Orthopedics*, 136, 632.e1–5; discussion 632.
18. Dwan, K., *et al.* (2008) Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One*, 3, e3081.
19. West, B.T., Welch, K.B. and Galecki, A.T. (2007) *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman & Hall/CRC. Taylor and Francis Group, New York.
20. Fitzmaurice, G.M., Laird, L.M. and Ware, J.H. (2004) *Applied Longitudinal Analysis*. John Wiley & Sons Inc., Hoboken, New Jersey, USA. pp. 326–328.
21. von Elm, E., Altman, D.G., Egger, M., Pocock, S.J., Gøtzsche, P.C. and Vandenbroucke, J.P.; STROBE Initiative. (2007) Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement:

guidelines for reporting observational studies. *BMJ (Clinical research ed.)*, 335, 806–808.

22. Moher, D., Liberati, A., Tetzlaff, J. and Altman, D.G.; PRISMA Group. (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine*, 6, e1000097.
23. Stroup, D.F., Berlin, J.A., Morton, S.C., Olkin, I., Williamson, G.D., Rennie, D., Moher, D., Becker, B.J., Sipe, T.A. and Thacker, S.B. (2000) Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. *JAMA*, 283, 2008–2012.
24. Lang, T.A. and Altman, D.G. (2015) Basic statistical reporting for articles published in biomedical journals: the "Statistical Analyses and Methods in the Published Literature" or the SAMPL Guidelines. *International Journal of Nursing Studies*, 52, 5–9.
25. Koletsi, D., Fleming, P.S., Behrents, R.G., Lynch, C.D. and Pandis, N. (2017) The use of tailored subheadings was successful in enhancing compliance with CONSORT in a dental journal. *Journal of Dentistry*, 67, 66–71.
26. Turner, L., Shamseer, L., Altman, D.G., Weeks, L., Peters, J., Kober, T., Dias, S., Schulz, K.F., Plint, A.C. and Moher, D. (2012) Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *The Cochrane Database of Systematic Reviews*, 11, MR000030.

Appendix 1. Examples from included articles which used either interpretation based on within group comparisons or otherwise.

	Aim of study	Intervention/comparator	Outcome	Discussion/conclusion	Judgement
Example 1 (Study id: 4)	To investigate the efficiency of piezosurgery technique in accelerating miniscrew supported en-masse retraction and study the biological tissue response […]	Interventions: Piezosurgery-assisted versus conventional en-masse retraction anchored from miniscrews placed between second premolars and first molars […]	The main outcome was the en-masse retraction rate	- Our results showed that the difference between retraction rates was not significant, although piezosurgery group (G1) showed slightly higher rates. - No evidence was found to support the claim that piezosurgery technique is an efficient way of accelerating en-masse retraction. - Changes in the nature of incisor and molar movement, cephalometric, and dental cast variables were similar in two groups	Overall interpretation based on both within and between group comparisons
Example 2 (Study id: 63)	The aim was to find out if 1 year active treatment time with EGA was sufficient for achieving normal occlusal relationships and dental alignment in 7- to 8-year-old children […]	The participants were randomly assigned into a treatment group (N = 25) and a control group (N = 23). Children in the treatment group received treatment with the EGA for 1 year. The controls had no orthodontic treatment	Changes in overjet, overbite, Angle's Class, and crowding were used as primary outcome measures […]	- Our results showed distinct improvements in overjet, overbite, sagittal molar relationship, and crowding in the treated subjects - In conclusion, the present results suggest that the EGA may be an effective treatment option for improving incisal relationships, class II malocclusion, and crowding in young children	Overall interpretation based on within group comparisons from baseline to follow-up