*Original Article*

# Assessment of reliability in orthodontic literature: *A meta-epidemiological study*

**Richard E. Donatelli[a]; Ji-Ae Park[b]; Yasser Murdi Abdullah Alghamdi[c]; Nikolaos Pandis[d]; Shin-Jae Lee[e]**

**ABSTRACT**

**Objectives:** To map the statistical methods applied to assess reliability in orthodontic publications and to identify possible trends over time.

**Materials and Methods:** Original research articles published in 2009 and 2019 in a subset of orthodontic journals were downloaded. Publication characteristics, including publication year, number of authors, single vs multicenter study, geographic origin of the study, statistician involvement, study category, subject category, types of reliability assessment, and statistical methods applied to assess reliability, were recorded. Descriptive statistics, Chi-square tests, and logistic regression analyses were performed to investigate associations between reliability analysis and study characteristics.

**Results:** A total of 768 original research articles were analyzed. The most prevalent study category was observational (69%) with a statistician involved in 16% of studies. Overall, reliability was assessed in 47% of studies, and the most frequent methods applied to assess reliability were intraclass correlation coefficients or kappa statistics (60.4%). The odds of applying appropriate methods were greater in 2019 than in 2009 (odds ratio [OR]: 2.43; 95% confidence interval [CI]: 1.75, 3.37; $P < .001$). Involvement of a statistician resulted in greater odds of applying appropriate methods compared to no statistician involvement (OR: 1.88; 95% CI: 1.23, 2.87; $P < .01$).

**Conclusions:** Over the past decade (2009 vs 2019), reliability assessment became more common in the orthodontic literature, and studies applying correct statistical methods to assess reliability significantly increased. This trend was more apparent in studies that involved a statistician, which may highlight the role of the statistician. (*Angle Orthod.* 2022;92:409–414.)

**KEY WORDS:** Reliability statistics; Research trend

## INTRODUCTION

Reliability in clinical measurements pertains to reproducibility over time with minimal measurement errors. In a clinical study, it is important to assess the reliability of the applied measurements to exclude imprecision and biases due to use of inappropriate measures. Reliability should be assessed even for methods reported to be reliable in past studies because there is no guarantee that the same method implemented by a different investigator in a different setting will also be reliable.[1] Though some uncertainties are inevitable during an experiment or survey, interobserver variability also needs to be addressed.[2]

In 2000, when BeGole examined common statistical procedures in 203 articles published in three leading orthodontic journals, the proportion of studies that had included reliability assessment was relatively small, with only 33 using reliability statistics out of 407 statistical procedures (8%).[3] In the past, reliability

The first two authors contributed equally to this work.

[a] Assistant Professor and Program Director, Department of Orthodontics, College of Dentistry, University of Florida, Gainesville, Florida, USA.

[b] Postgraduate Student, Department of Orthodontics, Graduate School, Seoul National University, Seoul, Korea.

[c] Resident, Ministry of Health, Kingdom of Saudi Arabia.

[d] Professor, Department of Orthodontics and Dentofacial Orthopedics, School of Dental Medicine, University of Bern, Bern, Switzerland.

[e] Professor, Department of Orthodontics and Dental Research Institute, Seoul National University School of Dentistry, Seoul, Korea.

Corresponding author: Shin-Jae Lee, DDS, MSD, PhD (Stats), PhD, Professor, Department of Orthodontics and Dental Research Institute, Seoul National University School of Dentistry, 101 Daehakro, Jongro-Gu, Seoul 03080, Korea
(e-mail: nonext.shinjae@gmail.com)

Accepted: December 2021. Submitted: August 2021.
Published Online: January 31, 2022

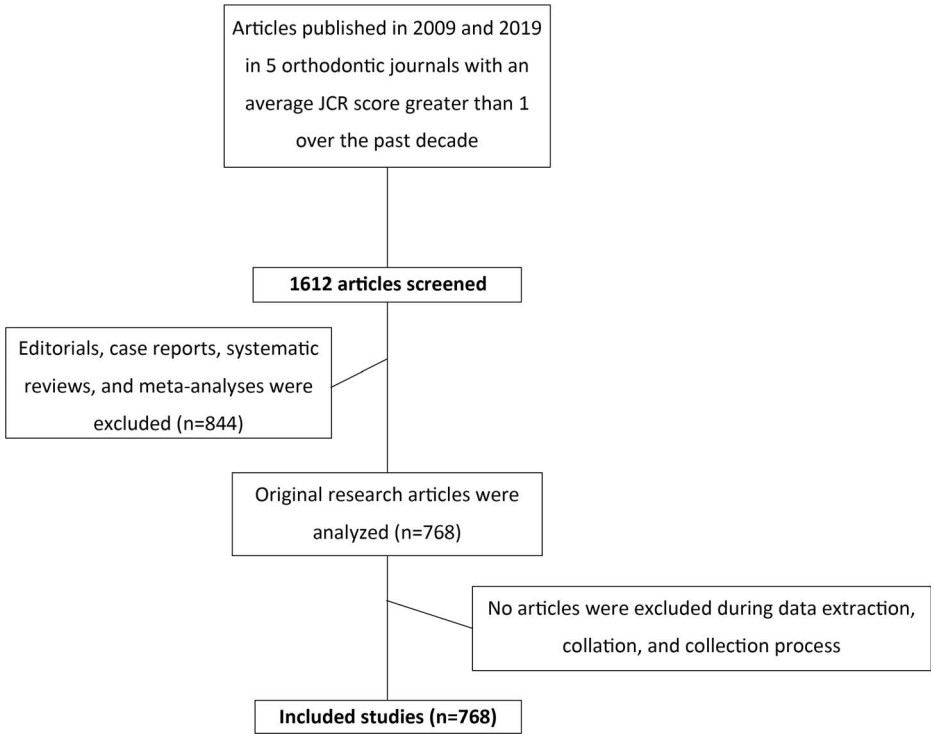


**Figure 1.** Flow diagram of study selection.

assessment in healthcare commonly relied on the root mean squared error (Dahlberg procedures),[4] the intraclass correlation coefficient (ICC),[5] and the kappa statistic[6] with no studies applying graphical methods such as the Bland-Altman plot.[7] More recently, reliability assessment has become routine in the orthodontic literature and rigorous and advanced statistical methods have been increasingly used. However, several studies still used suboptimal methods such as a *t*-test or the correlation coefficient, for reliability assessment.[8] The *t*-test was designed to compare the difference between two means and not the concordance between two measurements. In addition, the use of different statistical tests for reliability assessment can create significant interstudy variation and can prevent comparison of repeatability across studies. Use of more appropriate methods can provide a more optimal assessment of reliability and implementing such tests as standard practice would allow for comparison of interstudy reliability.

There has been no recent study assessing reliability analysis in the orthodontic literature. Therefore, the aim of the present study was to map the statistical methods that were applied to assess reliability in orthodontic publications. In addition, changes in reliability assessment over time were examined as well as possible associations between the use of optimal approaches and study characteristics.

## MATERIALS AND METHODS

### Data Sources, Search, and Selection

Original research articles published in 2009 and 2019 in five orthodontic journals with an average Journal Citation Reports (JCR) score $>$ 1 over the past decade were selected. JCR has often been used to select leading journals[9,10] with the highest impact factors in each specialty. The selected journals were *American Journal of Orthodontics and Dentofacial Orthopedics* (AJODO), *Angle Orthodontist* (AO), *European Journal of Orthodontics* (EJO), *Orthodontics and Craniofacial Research* (OCR), and *Korean Journal of Orthodontics* (KJO).

In the present study, all original research articles were downloaded and electronically searched by one investigator (Y.M.A.A.). Editorials, case reports, systematic reviews with or without meta-analyses were excluded. To extract reliable data, multiple calibration sessions were developed by two investigators (J.A.P. and S.J.L.) under detailed written instructions (Figure 1).

### Data Extraction, Collation, and Collection Process

From each publication, the following variables was extracted: publication year, number of authors, single- vs multicenter study, geographic origin of the study, statistician involvement, study category, subject cate-

gory, types of reliability assessment (intra- and interexaminer reliability), and statistical methods applied to assess reliability. Among the article characteristics, geographic origin (continent), study category, subject category, and types of statistical tests were consistent with previous categories reported by Koletsi et al.[10]

Statistician involvement (yes/no) was determined using the title page and acknowledgment statement. When statistician involvement was not clearly stated in these sections, statistician involvement was coded as "no." Data extraction followed an iterative process until all disagreements were eliminated.

### Statistical Analysis

Descriptive statistics were calculated and Chi-square tests and logistic regression analyses were performed to examine associations between study characteristics over time. Study characteristics that were associated with the primary outcome (appropriate/inappropriate reliability analysis) in the univariable logistic regression were added in the multivariable model. All statistical analyses were performed using Language R version 4.1.1 (R Foundation for Statistical Computing, Vienna, Austria).[11]

## RESULTS

A total of 768 original research articles published in five selected orthodontic specialty journals were analyzed (Figure 1).

First, a pilot calibration was conducted on 10 articles, which produced disagreement on seven items between the two investigators. After 4 days, the second calibration conducted on another 10 articles resulted in disagreement on four items. After 1 week, a third calibration conducted on another 60 articles resulted in disagreement on only two items. Not all the disagreements between the two investigators could be resolved by multiple calibration sessions. Overall, there were disagreements on 56 items out of a total of 6912 items, which were later corrected by unanimous consent.

AO included the highest proportion of original research articles, followed by AJODO. Fewer original research articles were published in 2019 (n = 335) compared to 2009 (n = 433) (Table 1).

Specifically, the number of original research articles published by authors in Europe and Asia had decreased, whereas the number of articles was similar between the two periods for America. Number of coauthors > 5 and number of multicenter studies significantly increased in 2019 relative to 2009. A statistician was involved in 14% and 18% of the studies in 2009 and 2019, respectively. Over time, the most prevalent study category was observational (n = 463, 60.3%) with the highest proportion concerning human subjects (n = 559, 72.8%). The numbers of in vitro and animal studies significantly decreased in 2019 relative to 2009 (Table 1).

**Table 1.** Assessment of 768 Original Research Articles According to Article Characteristics[a]

| Study Variables | 2009 n (%[b]) | 2019 n (%[b]) | Total n (%[b]) | *P* Value[c] |
|---|---|---|---|---|
| Total | 433 (56) | 335 (44) | 768 (100) | |
| Journal | | | | <.001 |
| AO | 154 (60) | 103 (40) | 257 (100) | |
| AJODO | 134 (55) | 111 (45) | 245 (100) | |
| EJO | 89 (77) | 27 (23) | 116 (100) | |
| OCR | 23 (28) | 60 (72) | 83 (100) | |
| KJO | 33 (49) | 34 (51) | 67 (100) | |
| Continent | | | | <.001 |
| Europe | 168 (64) | 96 (36) | 264 (100) | |
| America | 114 (49) | 120 (51) | 234 (100) | |
| Asia | 133 (59) | 92 (41) | 225 (100) | |
| Other countries | 18 (40) | 27 (60) | 45 (100) | |
| Number of authors | | | | <.001 |
| 1–3 | 141 (66) | 72 (34) | 213 (100) | |
| 4–5 | 195 (61) | 126 (39) | 321 (100) | |
| ≥6 | 97 (41) | 137 (59) | 234 (100) | |
| Number of centers | | | | <.001 |
| Single center | 252 (67) | 124 (33) | 376 (100) | |
| Multicenter | 181 (46) | 211 (54) | 392 (100) | |
| Statistician involvement | | | | .12 |
| No | 373 (58) | 274 (42) | 647 (100) | |
| Yes | 60 (50) | 61 (50) | 121 (100) | |
| Study category | | | | <.01 |
| Observational | 244 (53) | 219 (47) | 463 (100) | |
| In vitro | 100 (68) | 48 (32) | 148 (100) | |
| Interventional | 41 (49) | 42 (51) | 83 (100) | |
| Animal | 48 (65) | 26 (35) | 74 (100) | |
| Subject category | | | | <.001 |
| Human | 288 (52) | 271 (48) | 559 (100) | |
| Material | 83 (76) | 26 (24) | 109 (100) | |
| Animal | 46 (62) | 28 (38) | 74 (100) | |
| Human material | 16 (62) | 10 (38) | 26 (100) | |

[a] AJODO indicates *American Journal of Orthodontics and Dentofacial Orthopedics*; AO, *Angle Orthodontist*; EJO, *European Journal of Orthodontics*; OCR, *Orthodontics and Craniofacial Research*; KJO, *Korean Journal of Orthodontics*.
[b] Row percentage.
[c] Chi-square test.

Overall, 47% of studies assessed reliability. Among studies that assessed reliability, only 22% reported both intra- and interexaminer reliability. The proportion of studies that assessed both intra- and interexaminer reliability significantly increased in 2019 relative to 2009. The most frequent method applied to assess reliability was ICC or the kappa statistic (60.4%). In reporting reliability, the proportion of studies using a graphical method to assess reliability, for example, the Bland-Altman plot, increased in 2019 relative to 2009. The proportion of studies incorporating optimal reliability statistics increased over the past decade, and incorrect use of inferential tests as a reliability measure, such as *t*-tests, analysis of variance (AN-

**Table 2.** Type of Reliability Assessment and Statistical Analysis Method to Measure Reliability

| | 2009 n (%[a]) | 2019 n (%[a]) | *P* Value[b] |
|---|---|---|---|
| Type of reliability assessment | | | <.0001 |
| Not reported | 264 (61) | 142 (42) | |
| Reported | 169 (39) | 193 (58) | |
| Total | 433 (100) | 335 (100) | |
| Type of reliability assessment | | | .0025 |
| Intraexaminer reliability | 131 (78) | 117 (61) | |
| Interexaminer reliability | 13 (8) | 24 (12) | |
| Both inter- and intraexaminer reliability | 25 (15) | 52 (27) | |
| Total | 169 (100) | 193 (100) | |
| Statistical analysis to assess reliability | | | .0002[c] |
| Appropriate methods to assess reliability | | | .0009 |
| Intraclass correlation coefficient/kappa[d] | 74 (52) | 143 (68) | |
| Root mean squared error (Dahlberg) | 59 (42) | 48 (23) | |
| Bland-Altman plot | 9 (6) | 19 (9) | |
| Total | 142 (100) | 210 (100) | |
| Not appropriate methods to assess reliability | | | 1[e] |
| *t*-test | 47 (66) | 29 (63) | |
| ANOVA[f] | 2 (3) | 3 (7) | |
| Correlation analysis | 22 (31) | 14 (30) | |
| Total | 71 (100) | 46 (100) | |

[a] Column percentage.
[b] Chi-square test.
[c] Chi-square test only for the rows and columns named "Appropriate methods to assess reliability," and "Not appropriate methods to assess reliability."
[d] Intraclass correlation coefficient/kappa category includes concordance correlation coefficient reports three times, a Kendall coefficient of concordance report, a coefficient of reliability report, an iota coefficient report, a Crohnbach's alpha report, and a Krippendorff's alpha report.
[e] Chi-square test result after pooling the rows of *t*-test and ANOVA.
[f] ANOVA indicates analysis of variance.

OVA), and correlation statistics, decreased from 2009 to 2019 (Table 2).

Figure 2 depicts the counts of appropriate/inappropriate analyses per year and statistician involvement. In the adjusted multivariable analysis, the odds of applying appropriate methods to assess reliability were greater in 2019 than in 2009 (odds ratio [OR], 2.43; 95% confidence interval [CI]: 1.75, 3.37; $P < .001$). For in vitro, interventional, and animal studies, the odds ratios of applying correct methods were $< 1$ ($P < .001$). Involvement of a statistician resulted in greater odds of applying appropriate reliability assessments compared to no statistician involvement (OR, 1.88; 95% CI: 1.23, 2.87, $P < .01$) (Table 3).

## DISCUSSION

The present study was performed to record the statistical methods used to assess reliability in orthodontic publications, and to examine possible trends over time. Significant improvements were observed when the data extracted from 2019 publications were compared to data extracted from 2009 publications. For example, the percentage of studies that included reliability assessment increased from 39% in 2009 to 58% in 2019. The use of appropriate methods to assess reliability was greatly increased in 2019. It was also noticeable that involvement of a statistician increased the proportion of correct application of the reliability assessment. It could be conjectured that consultation with a statistician might be meaningful in conducting an appropriate method.

The quantity and quality of reliability statistics changed over time. According to a report published in 2000,[3] reliability measures were applied to <10% of the articles and were limited only to two kinds of procedures: the Dalberg procedure (56%) and ICC (43%).[3] More recently, the Dahlberg procedure was used less frequently (n = 107) than ICC statistics (n = 217). The proportion of incorrect methods to assess reliability decreased dramatically with increased application of the Bland-Altman plot in 2019 relative to 2009 being also noticeable. This trend likely indicates increased awareness and improvements in orthodontic research methodology.

The number of coauthors per study and the proportion of multicenter studies increased in 2019 relative to 2009 and was in agreement with the increased number of multidisciplinary research efforts in recent years.[9,10] However, despite the increased use of more advanced statistical tests, statistician involvement did not increase in 2019 relative to 2009 in the present study samples. This could have been due to the versatility and easy accessibility of commercial statistical software now available to investigators. Currently, much of the software is commercially available but often requires a trained statistician to

**Table 3.** Result of Univariable and Multivariable Logistic Regression Analyses to Examine Potential Associations Between Study Characteristics and Use of Appropriate Methods to Assess Reliability (n = 768)[a]

| Variable and its Category | Univariable | | | Multivariable | | |
|---|---|---|---|---|---|---|
| | Odds Ratio | 95% CI | *P* Value | Odds Ratio[b] | 95% CI | *P* Value |
| Publication year | | | <.001 | | | <.001 |
| 2009 | Reference | | | | | |
| 2019 | 2.63 | 1.93, 3.59 | | 2.43 | 1.75, 3.37 | |
| Journal | | | .10[b] | | | |
| AO | Reference | | | | | |
| AJODO | 1.67 | 1.14, 2.44 | | | | |
| EJO | 1.52 | 0.95, 2.43 | | | | |
| OCR | 1.27 | 0.71, 2.28 | | | | |
| KJO | 1.41 | 0.83, 2.41 | | | | |
| Continent | | | .06[c] | | | |
| Europe | Reference | | | | | |
| America | 0.78 | 0.54, 1.13 | | | | |
| Asia | 0.67 | 0.45, 0.98 | | | | |
| Other countries | 1.40 | 0.74, 2.65 | | | | |
| Number of authors | | | .65[c] | | | |
| 1–3 | Reference | | | | | |
| 4–5 | 1.18 | 0.81, 1.71 | | | | |
| ≥6 | 1.16 | 0.78, 1.74 | | | | |
| Number of centers | | | | | | |
| Single center | Reference | | .15 | | | |
| Multicenter | 1.25 | 0.92, 1.69 | | | | |
| Study category | | | <.001[c] | | | <.001[c] |
| Observational | Reference | | | | | |
| In vitro | 0.62 | 0.38, 1.02 | | 0.54 | 0.32, 0.91 | |
| Interventional | 0.14 | 0.08, 0.25 | | 0.09 | 0.03, 0.29 | |
| Animal | 0.32 | 0.17, 0.59 | | 0.14 | 0.01, 1.48 | |
| Subject category | | | <.001[c] | | | |
| Human | Reference | | | | | |
| Material | 0.36 | 0.20, 0.66 | | | | |
| Animal | 0.23 | 0.13, 0.41 | | | | |
| Human material | 0.13 | 0.03, 0.55 | | | | |
| Statistician involvement | | | | | | <.01 |
| No | Reference | | | | | |
| Yes | 2.18 | 1.47, 3.24 | <.001 | 1.88 | 1.23, 2.87 | |

[a] AJODO indicates *American Journal of Orthodontics and Dentofacial Orthopedics*; AO, *Angle Orthodontist*; CI, confidence interval; EJO, *European Journal of Orthodontics*; OCR, *Orthodontics and Craniofacial Research*; KJO, *Korean Journal of Orthodontics*.
[b] Adjusted odds ratio.
[c] Likelihood ratio test for the overall associations.

guide in the analytical process for it to be credible and reliable.

Use of inappropriate methods such as *t*-tests, ANOVAs, and correlation analyses to assess reliability decreased over time. However, those three methods for reporting reliability are still common and were used in 117/469 (25%) of the selected articles.

Mean comparison methods, such as *t*-tests or ANOVA, should be used to find differences between means of groups and not to measure reliability. No significant differences between groups after the use of mean comparing tests implies no difference in the means but provides no information regarding the range of deviations at the individual level. Large disagreements between pairs of individual reliability measurements can still result in small or even non-existent mean differences, completely masking the lack of agreement.[1,2]

The correlation coefficient was the second most prevalent among incorrect methods to assess reliability. However, the correlation coefficient should not be used as a reliability measure since it does not indicate the agreement, but the linear association between two variables. It is likely that two highly correlated sets of measurements never agree. For example, the pairs (1,2) (2,4) (3,6) (4,8) (5,10) have a correlation coefficient of 1 but, in reality, their individual pair values are evidently in great disagreement. In addition, the null hypothesis for a correlation test is testing whether the correlation coefficient is zero and, thus, the magnitude of the *P* values after the correlation test are not very meaningful. A common misconception was that small *P* values would indicate strong correlation;

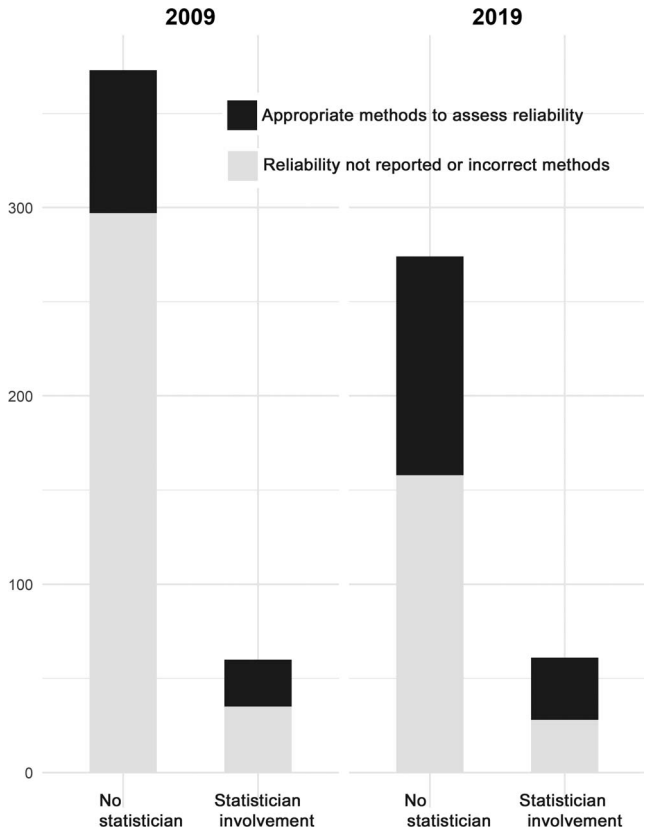


**Figure 2.** Stacked bar plot demonstrating that the proportion of studies conducting correct use of methods to assess reliability increased over time and were more likely in studies involved with a statistician.

however, low *P* values after a correlation test have nothing to do with the strength of the correlation.[1,2] In this regard, it was encouraging to see that studies applying incorrect methods to assess reliability decreased over time.

This study was not without limitations since it included articles in 2009 and 2019, omitting the years in between. Including all years would have resulted in an intractable number of publications. However, the sample might be adequate to map the area and provide evidence in the reliability analysis practices over time.

## CONCLUSIONS

- Based on this cross-sectional survey of original research articles published in five orthodontic journals in 2009 and in 2019, the results demonstrated that studies conducting correct methods to assess reliability significantly increased over time.
- Involvement of a statistician increased the odds of applying correct statistical methods to assess reliability, which may highlight the meaningful role of the statistician in orthodontic research.

## REFERENCES

1. Donatelli RE, Lee SJ. How to report reliability in orthodontic research: Part 1. *Am J Orthod Dentofacial Orthop*. 2013;144:156–161.
2. Donatelli RE, Lee SJ. How to report reliability in orthodontic research: Part 2. *Am J Orthod Dentofacial Orthop*. 2013;144:315–318.
3. BeGole EA. Statistics for the orthodontist. In: Graber TM, Vanarsdall RL, eds. *Orthodontics: Current Principles and Techniques*. St. Louis: Mosby; 2000:339–352.
4. Dahlberg G. *Statistical Methods for Medical and Biological Students*. London: George Allen & Unwin Ltd.; 1940.
5. Fleiss JL. *The Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons; 1985.
6. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–174.
7. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1:307–310.
8. Donatelli RE, Lee SJ. How to test validity in orthodontic research: a mixed dentition analysis example. *Am J Orthod Dentofacial Orthop*. 2015;147:272–279.
9. Aura-Tormos JI, Garcia-Sanz V, Estrela F, Bellot-Arcis C, Paredes-Gallardo V. Current trends in orthodontic journals listed in Journal Citation Reports. A bibliometric study. *Am J Orthod Dentofacial Orthop*. 2019;156:663–674.e661.
10. Koletsi D, Madahar A, Fleming PS, Pandis N. Statistical testing against baseline was common in dental research. *J Clin Epidemiol*. 2015;68:776–781.
11. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2021.