Clin Oral Invest (2005) 9: 70–75
DOI 10.1007/s00784-005-0001-8

ORIGINAL ARTICLE

H. P. Müller · K. M. Barrieshi-Nusair

Gingival bleeding on repeat probing after different time intervals in plaque-induced gingivitis

Received: 5 April 2005 / Accepted: 7 June 2005 / Published online: 9 July 2005

Abstract The aim of this study was to assess agreement and association of gingival bleeding after repeated probing at different time intervals in subjects with gingivitis. Twenty adults participated. Periodontal probing ($P \approx 1.27$ MPa) was conducted at six sites of every tooth present. Probing was repeated in different quadrants immediately after the first probing (T0), after 1 h (T1), 4 h (T4), and 24 h (T24). A total of 3,459 sites were probed twice. The mean proportion of sites bleeding on probing (BOP) was 0.23 (standard deviation 0.08, range 0.10–0.41). Probing itself had a significant effect on the results of repeated probing. For T0 through T24, respective mean differences of proportions were 0.04, 0.04, 0.01, and −0.03. Ninety-five percent repeatability coefficients of proportions were 0.17–0.18. Estimates of overall kappa were 0.390, 0.234, 0.233, and 0.046 for T0 through T4, respectively. Adjusted two-level binary response models revealed odds ratios (95% confidence interval) for BOP at T0 through T4 of 6.52 (4.34, 9.80), 3.23 (2.19, 4.76), 3.80 (2.63, 5.50), and 2.68 (1.85, 3.89). It was concluded that a certain degree of agreement of site-specific bleeding scores in subjects with plaque-induced gingivitis could be observed only if probing was repeated at once. Adjusted associations between repeat BOP were weak in general, but strongest immediately after first probing. There appears to be a significant effect of probing itself, which may last for more than 1 h, whereas 24-h results are obtained under different conditions.

Keywords Gingival inflammation · Bleeding on probing · Reliability · Agreement · Association

H. P. Müller · K. M. Barrieshi-Nusair
Faculty of Dentistry, Kuwait University,
P.O. Box 24923 Safat, 13110, Kuwait

H. P. Müller (✉)
Faculty of Dentistry, Kuwait University,
P.O. Box 24924 Safat, 13110, Kuwait
e-mail: hp.muller@hsc.edu.kw
Tel.: +965-252-2664502
Fax: +965-252-2634247

Introduction

Since the early 1960s, bleeding of gingival tissue after some sort of mechanical manipulation is a widely used clinical parameter to characterize its inflamed condition [19]. Increased bleeding tendency of gingiva relates to an inflammatory infiltrate in the tissue [3, 4, 11, 21], pointing to a certain degree of validity of the observation. In several attempts, reliability of bleeding on probing (BOP) and related indices had been determined in replication studies, after varying or ill-defined time intervals, often employing questionable statistical methods, and frequently with rather poor results [7, 9, 23, 25, 26].

There is limited and contradictory information regarding the effect of time on gingival bleeding after repeat probing [1, 12, 25]. In preparation of a clinical trial in young adults with plaque-induced gingivitis, we conducted a study to analyze results of repeat gingival probing after different time intervals. First, we did a thorough analysis of agreement of the *proportion of sites* with BOP. Furthermore, *site-specific agreement* of BOP scores was assessed, and finally, multilevel modeling for analyzing the *association* between gingival bleeding after varying time intervals between probing was employed whereas simultaneously considering various covariates at subject and site levels.

Material and methods

Volunteers

Twenty subjects (six men) participated. Volunteers were between 21 and 39 years old (median age 32 years) and systemically healthy. Inclusion criteria were mild or moderate plaque-induced gingivitis and a minimum of 25 teeth. Three subjects were mild smokers (7.5 to 12 pack-years). The following exclusion criteria applied: (1) any indication for antibiotic prophylaxis, (2) pregnancy or lactation, (3) any long-term medication with a possible effect on the gingiva, (4) any non-plaque-induced gingival dis-

ease, (5) destructive periodontal disease with a possible exception of localized attachment loss, (6) extensive tooth restoration or tooth replacement, and (7) large amounts of dental calculus. After briefing on the study's aim and procedures, volunteers gave written informed consent for participation.

Periodontal examinations

Systematic periodontal probing to the bottom of the "sulcus" was conducted at six sites of every tooth present with a pressure-controlled probe (ClickProbe 1395, KerrHawe, Bioggio, Switzerland). The probe has a tip diameter of 0.5 mm, and probing force is, according to the manufacturer, 0.25 N when the probe bends with a palpable click. Thus, pressure was about 1.27 MPa. No time limit for assessing presence or absence of BOP was preset because bleeding may be delayed in certain individuals [20]. Probing was repeated in different quadrants immediately (following thorough mouth rinsing with tap water) after the first probing (T0), after 1 h (T1), 4 h (T4), and 24 h (T24). Four sampling schemes were randomly assigned to volunteers. For a given quadrant probed immediately after the first probing, the contralateral quadrant in the opposite jaw was probed after 1 h. Probing was repeated after 4 h in the quadrant of the same side in the opposite jaw, and in the remaining quadrant, after 24 h. One examiner was employed who had no access to previous recordings.

Data analysis

A statistical program was used for calculations and graphical display (SYSTAT 8.0 for WINDOWS, SYSTAT, Evanston, IL, USA). Mean and standard deviations (SD) of subjects' BOP proportions at different time intervals were calculated and repeated measures ANOVA applied to test the hypothesis of no time-dependent change of differences between first and second probing. We performed a thorough analysis of agreement/repeatability for BOP proportions [2] including analysis of data pairs, calculation of repeatability coefficient, and graphical display of differences against means of two measurements. It was expected that mean differences were zero and 95% of differences less than 1.96 SD, the repeatability coefficient. Differences were also regressed on means of BOP proportions, as were residuals from the models.

Site-specific analyses, stratified at the subject level, followed [8]. First, fourfold tables were constructed for each volunteer and time interval, and Cohen's kappa calculated. We estimated overall kappa coefficients by considering individual weights defined as reciprocals of the squared standard errors (SEs) of individual kappa estimates. We also tested homogeneity of individual kappa values. Homogeneity of marginal proportions (pointing to no direct effect of the probing procedure on bleeding results) was tested in the same way. Reciprocals of squared SE of differences of proportions of discrepant bleeding results were used as weights and respective overall estimates constructed.

For assessing multilevel associations between bleeding scores, multilevel modeling [10] was applied using special software (*MLwiN* 2.0, Centre for Multilevel Modelling, Institute of Education, University of London, London, UK). We built two-level (site, subject) binary response models with bleeding after second probing (BOP2) as dependent variable. Given a sample of gingival units from individuals, we can write the model for the probability of BOP2 as

$$\operatorname{logit}(\pi_{ij}) = (X\beta)_{ij} + u_j$$
$$y_{ij} \sim \operatorname{Bin}(\pi_{ij}, 1)$$

using a standard logit link function assuming binomial error distribution for the (0,1) response y_{ij}. The level 2 (subject) random variation is described by the term u_j within the linear predictor. Although initially, no separate estimate for the level 1 (site) variance is assumed, extra-binomial variation was also fitted. First, two-level binary response models were built with bleeding after first probing (BOP1) as independent variable. Next, stepwise adjustment for tooth type and site, and thereafter, age, gender, smoking status, percent sites with instant BOP, and sampling scheme, allowed for a detailed analysis of changes of the observed association between BOP scores.

Results

Nine sites (0.26%) in five patients were excluded from the analysis because of recent tooth extraction adjacent to the sites or attachment loss (≥2 mm) in combination with increased probing depth (≥5 mm). A total of 3,459 sites were probed twice.

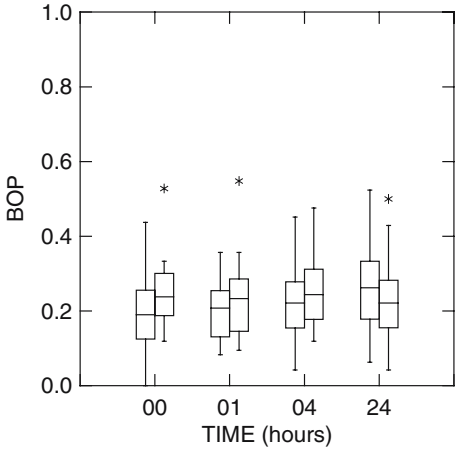

Fig. 1 Box plots showing proportions of sites with BOP at first (*left*) and second (*right*) probing with different time intervals (0, 1, 4, and 24 h)

Table 1 Analysis of repeatability of the proportion of sites bleeding on probing (BOP) when probing was repeated immediately after the first probing (T0), after 1 h (T1), 4 h (T4), and 24 h (T24)

	Mean difference±SD*	Repeatability coefficient
T0	0.044±0.091	0.178
T1	0.035±0.094	0.184
T4	0.012±0.089	0.174
T24	−0.031±0.085	0.167

*p=0.04 (repeated measures ANOVA)

Proportion of sites BOP

At first probing, 22.8% (SD 8.1) sites bled, on average. Box plots in Fig. 1 display the distribution of the proportion of sites BOP before and after probing at different time intervals. Mean differences of proportions of sites with BOP tended to be higher when probing was either repeated immediately or after 1 h and lower after 24 h (p=0.04). Ninety-five percent repeatability coefficients were between 0.167 and 0.184 (Table 1). No systematic trend of change in differences with higher BOP percentage could be ascertained at any time interval.

Site-specific agreement

Overall kappa estimates for bleeding results at different time intervals varied considerably (Table 2). A certain level of agreement could only be established when probing was repeated immediately after the first probing (overall kappa 0.390, SE 0.036). After 24 h, hardly any agreement could be ascertained. Moreover, although individual kappa values were rather homogeneous for repeat probing at T0 and T1, results after 4 and 24 h differed significantly among subjects. Weighted overall differences of proportions with discrepant bleeding results after first and second probing indicate a deleterious effect for immediate repetition of probing and repetition after 1 h (0.042 and 0.032), whereas the opposite was noticed after 24 h (−0.036).

Site-specific association

We employed multilevel modeling to adjust time interval-related associations of repeated BOP for subject and tooth-related covariates. Three sets of models are presented in Table 3. Crude odds ratios for BOP were 7.43 at T0, 3.67 at T1, 4.56 at T4, and 3.42 at T24. When the models were adjusted for tooth type and site, odds ratios were decreased by about 10–15% (model 2). Model 3 was adjusted for age, gender, smoking status, proportion of sites with BOP at the outset, and sampling scheme. Association of BOP scores at different time intervals remained largely unaffected. Subject variation, as expressed by u_{1j}, was between 0.022 and 0.218 in unadjusted models and largely reduced in adjusted models (between 0 and 0.081).

Discussion

Seven criteria have been listed for good performance of a gingival or periodontal examination [24]: it should be simple, accurate, reproducible, objective, quantitative, quickly done, and amenable to statistical analysis. For the sake of simplicity, gingival inflammation was confined here to one factor, namely, bleeding after probing to the bottom of the sulcus. Because probing is done in most gingivitis experiments by one examiner, we tested intraexaminer repeatability only. The chosen population presented with mild or moderate plaque-induced gingivitis, similar to the intended main study population. Whether gingiva bleeds after probing does not only depend on the presence or absence of an inflammatory infiltrate in the tissue but is also largely affected by probing pressure [15]. Probing with a constant force has led to greater reproducibility of gingival bleeding [25], making standardization of probing force mandatory.

The present study design allowed detailed analysis of the time factor on bleeding tendency after repeat probing. When probing was repeated up to 1 h after the first examination, higher mean proportions (about +0.04) of bleeding sites were discerned. The difference became small after 4 h (+0.01) whereas lower mean bleeding

Table 2 Agreement of BOP as expressed by overall kappa and overall weighted differences of marginal proportions p_2-p_1[a] with respective 95% confidence intervals (CI) when probing was repeated immediately after the first probing (T0), after 1 h (T1), 4 h (T4), and 24 h (T24)

	Overall kappa		Homogeneity		Marginal proportions		Homogeneity	
	$\hat{\hat{\kappa}}$	95% CI	$\chi^2_{(19)}$	p	Overall weighted p_2-p_1	95% CI	$\chi^2_{(19)}$	p
T0	0.390	0.319, 0.461	12.032	n.s.	0.041	0.012, 0.070	27.131	n.s.
T1	0.234	0.164, 0.304	18.623	n.s.	0.032	0.000, 0.064	27.964	n.s.
T4	0.233	0.172, 0.294	57.072	<0.001	0.019	−0.012, 0.051	24.750	n.s.
T24	0.046	0.004, 0.088	63.335	<0.001	−0.036	−0.067, −0.005	19.254	n.s.

Homogeneity tests for equal individual kappas and differences of proportions

[a]If results of probing at n sites in a certain subject are tabulated in a fourfold table as frequencies of pairs with agreement (a, d) and disagreement (b, c), then $p_1=(a+c)/n$, $p_2=(a+b)/n$, and $p_2-p_1=(b-c)/n$

Table 3 Odds ratios and 95% CI derived from three sets of two-level binary response models describing the influence of different time intervals (T1–T24) on association between first and second BOP at second probing

	Model 1[a]		Model 2[b]		Model 3[c]	
	Odds ratio	95% CI	Odds ratio	95% CI	Odds ratio	95% CI
T0	7.43	5.18, 10.68	6.66	4.52, 9.82	6.52	4.34, 9.80
T1	3.67	2.56, 5.26	3.23	2.22, 4.72	3.23	2.19, 4.76
T4	4.56	3.23, 6.42	4.04	2.82, 5.77	3.80	2.63, 5.50
T24	3.42	2.41, 4.87	2.82	1.96, 4.07	2.68	1.85, 3.89

[a]Unadjusted
[b]Adjusted for tooth type and site
[c]Adjusted for tooth type, site, age, gender, smoking, proportion of sites instantly bleeding after probing, and sampling scheme

proportion (−0.03) was observed after 24 h, probably due to the Hawthorne effect [6]. Any analysis of repeatability assumes zero mean difference between measurements. Systematic error (bias) indicates that values are not true replicates [2]. It may be concluded, therefore, that repeatability of BOP proportion is best to be assessed after 4 h. Repeatability coefficients, indicating the range in which 95% of differences between measurements can be expected, were about 0.17. Whether coefficients of that magnitude are sufficiently small for correct interpretation of data acquired in a gingivitis trial may be a matter of debate.

By summarizing observations at the patient level, "reliability" may be spuriously increased and can be high [22], even if the index is unreliable at all at the site level. Reliability of *site-specific* observations is to be assessed at the unit of examination. The 2×2 intraclass kappa directly estimates reliability as defined in the classical sense and is the ideal reliability coefficient for a binary measure [14]. Kappa coefficients are sensitive to the form of marginal distributions, which makes comparison of kappa values from different tables questionable. It is well-known that high reliability of any measure in a homogeneous population is difficult to achieve. But that is not a flaw in kappa. It merely reflects difficulties in making clear distinctions between units of observations in a population in which those distinctions are very rare or fine. In such populations, noise quickly overwhelms the signals [14]. Therefore, kappa does not describe reliability of a diagnostic as such but rather in a specific situation. It is a common misconception, e.g., in periodontal research, that inclusion of too many nondiseased sites might give a false picture of reproducibility because kappa may be low [26]. Instead, the sample population of a replication study should be similar in its parameter in question to that of the main study.

When numerous observations are made in a given individual, SEs of any measure of association will be spuriously minimized if sites in subjects are (erroneously) regarded independent. Here, we considered individual values of kappa to calculate a weighted estimate of common kappa. The highest overall kappa (0.390) was determined when probing was repeated immediately after first probing. Very low, if any, chance-corrected agreement was noted after 24 h (0.046). Tests for equality of individual kappa values revealed significant heterogeneity when probing was repeated after 4 and 24 h. Individual fourfold tables were also used for estimating weighted overall differences of marginal proportions [8] to analyze, at the site level, discordance of bleeding results. Somewhat smaller in magnitude as figures in Table 1, they provide a weighted site-specific average of differences, and again, point to a relevant traumatizing effect of the examination procedure itself, which may be traced at least 1 h after probing without significant heterogeneity. Comparison of present results with findings in other studies is difficult. Chosen time intervals between repeat probing vary from a few minutes [25] to even 2 days [23] and are frequently ill-defined. Our observations may corroborate the findings of investigators who reported increased bleeding frequencies for up to 2 h after initial probing in subjects with gingivitis [1, 26], but may be in contrast to others who examined nonsurgically treated periodontitis patients with increased probing depths, and did not observe an increased percentage of bleeding sites when probing was repeated after 15 [25] or 100 min [12]. However, investigators employed different probes, probing forces, probing techniques, and frequently used inappropriate statistics. In general, reported kappa coefficients for bleeding after repeat probing (inter- and/or intraexaminer reliability) were slightly higher in several studies [5, 23, 26]. But most important, kappa coefficients that have been obtained in different populations and/or situations cannot simply be compared. Moreover, as a diagnostic measure, BOP has definitively different meaning/interpretation in individuals with gingivitis, or patients with either untreated or treated periodontitis. All the more, its diagnostic accuracy should be assessed before any main study is conducted.

The present results further substantiate the claim that *reliability* of an invasive diagnostic measure cannot be determined. Instead, measures of *association* between findings after first and second probing, such as the odds ratio, may be more appropriate. Interpretation of the odds ratio as measure of association has recently been criticized [13]. When marginal proportions P or Q in a fourfold table are equal and 0.5, cut points for kappa (defining different levels of agreement) of 0.2, 0.4, 0.6, or 0.8 correspond to odds ratios of 2.3, 5.4, 16, and 81, respectively. Odds ratios are, however, much higher (4.5, 13.3, 42.7, and 223) if marginals are equal, but 0.1 or 0.9 (similar to those in the

present study), suggesting much stronger associations. An unassailable argument for using the odds ratio is the likelihood ratio test statistic for the null hypothesis of equality between two binomial distributions, in particular, in multiple logistic regression. In its strict sense, assumptions of regression models are not appropriate in the study of *reliability* [2, 13, 16, 27], where both variables are treated as random and neither is regarded independent or dependent. However, as already outlined, in replication of invasive probing, the second response indeed depends on the first.

As compared to any other time interval, the crude odds ratio for bleeding was twice as high when probing was immediately repeated. Adjustment for tooth type and site attenuated the association between BOP scores. In all models, the association was significantly lower at buccal sites, and in the model considering immediate repeat probing, association between probing results was stronger at first and second molars (not shown). The relationship between different BOP results at different tooth types, and sites may largely be influenced by various thickness of soft tissues [17]. Adjusting for other factors such as gender, age, smoking status, proportion of sites with BOP, and sampling scheme slightly affected overall associations between BOP at different time intervals. The present results have to be considered in view of a recent longitudinal study in which standardized probing was repeated in smokers and nonsmokers with gingivitis [18]. Examinations were performed every 8 weeks for 6 months in a steady-state plaque environment. Associations between BOP scores were rather low in smokers. However, in nonsmokers, the common odds ratio for BOP at baseline and at 8-week reexamination was 6.01 (95% confidence interval 4.33–8.36), comparable to that in the present study when probing was immediately repeated after the first probing.

Recently, a more complicated procedure for provoking gingival bleeding has been advocated for clinical studies on gingivitis [26]. The protocol of the *bleeding on marginal probing index* includes insertion into the gingival crevice of a 0.5-mm-diameter probe to a depth of approximately 2 mm, which is then run at an angle of approximately 60° to the long axis of the tooth. Bleeding is assessed within 30 s after probing. How are these technical details controlled? According to the authors, bleeding upon probing to the bottom of the sulcus is a poor indicator of early gingivitis. Instead, they recommend that gingivitis should be assessed by probing the marginal gingiva. In that study, however, measurement agreement of any procedure (irrespective within and between examiners) could hardly be rated as "fair" on a site basis [26].

It is self-evident that a particular diagnostic should be interpreted with caution if observations cannot be replicated. This may have serious consequences for the conduction of clinical studies in which *site-specific* gingival bleeding is the main outcome variable (or an integral part of a more "precise" index). As regards to the present results in subjects with gingivitis, the following might be concluded. The probing procedure itself has a small detrimental effect as indicated by an increase in the proportion of bleeding sites after second probing. It was consistent among individuals and might last for more than 1 h. If repeatability of full-mouth bleeding scores is to be assessed 4 h after the first probing when this effect seems to be small, the 95% repeatability coefficient may be as large as 0.17. After 24 h, a small trial effect may be discernible. A certain degree of agreement of site-specific bleeding scores can only be observed when probing is repeated at once, whereas after 24 h, a different situation may be present. The study confirms that *reliability* of gingival bleeding after probing in subjects with mild or moderate plaque-induced gingivitis can hardly be determined in a replication study. Adjusted *associations* between repeat BOP were weak in general, but strongest immediately after first probing. However, an odds ratio of 6.5 (with marginal proportions of about 0.8) hardly corresponds to a desired degree of repeatability of a diagnostic measure.

Acknowledgements ClickProbes were provided by KerrHawe, Switzerland. The study was supported in part by Kuwait University Research Administration, Grant # DS02/02.

References

1. Birkeland JM, Jorkjend L (1975) The influence of examination on the assessment of the intra-examiner error by using the plaque and gingival index systems. Community Dent Oral Epidemiol 3:214–216
2. Bland JM, Altman DG (1999) Measuring agreement in method comparison studies. Stat Methods Med Res 8:135–160
3. Cooper PG, Caton JG, Polson A (1983) Cell population associated with gingival bleeding. J Periodontol 54:497–502
4. De Souza PH, De Toledo BE, Rapp GE, Zuza EP, Neto CB, Mendes AJ (2003) Reliability of bleeding and non-bleeding on probing to gingival histological features. J Int Acad Periodontol 5:71–76
5. Eaton KA, Rimini FM, Zak E, Brookman DJ, Newman HN (1997) The achievement and maintenance of inter-examiner consistency in the assessment of plaque and gingivitis during a multicentre study based in general dental practices. J Clin Periodontol 24:183–188
6. Feil PH, Grauer JS, Gadbury-Amyot CC, Kula K, McCunniff MD (2002) Intentional use of the Hawthorne effect to improve oral hygiene in orthodontic patients. J Dent Educ 66:1129–1135
7. Feldman R, Douglass C, Loftus E, Kapur K, Chauncey H (1982) Interexaminer agreement in the measurement of periodontal disease. J Periodontal Res 17:80–90
8. Fleiss JL (1981) Statistical methods for rates and proportions, 2nd edn. Wiley, New York, pp 112–119, 160–187, 212–236
9. Fleiss JL, Chilton NW (1983) The measurement of interexaminer agreement on periodontal disease. J Periodontal Res 18:601–606
10. Goldstein H, Browne W, Rasbash J (2002) Multilevel modelling of medical data. Stat Med 21:3291–3315
11. Greenstein G, Caton J, Polson AM (1981) Histological characteristics associated with bleeding after probing and visual signs of inflammation. J Periodontol 52:420–425
12. Janssen PTM, Faber JAJ, van Palenstein Helderman WH (1986) Reproducibility of bleeding tendency measurements and the reproducibility of mouth bleeding scores for the individual patient. J Periodontal Res 21:653–659
13. Kraemer HC (2004) Reconsidering the odds ratio as a measure of 2×2 association in a population. Stat Med 23:257–270
14. Kraemer HC, Periyakoil VS, Noda A (2002) Kappa coefficients in medical research. Stat Med 21:2109–2129

15. Lang NP, Nyman S, Senn C, Joss A (1991) Bleeding on probing as it relates to probing pressure and gingival health. J Clin Periodontol 18:257–261
16. Müller HP, Behbehani E (2005) Methods of measuring agreement: glucose levels in gingival crevice blood. Clin Oral Investig 9:65–69
17. Müller HP, Könönen E (2005) Variance components of gingival thickness. J Periodontal Res 40:239–244
18. Müller HP, Stadermann S, Heinecke A (2001) Bleeding on probing in smokers and non-smokers in a steady state plaque environment. Clin Oral Investig 5:177–184
19. Newbrun E (1996) Indices to measure gingival bleeding. J Periodontol 67:555–561
20. Nowicki D, Vogel R, Melcer S, Deasy M (1981) The gingival bleeding time index. J Periodontol 52:260–262
21. Oliver RC, Holm-Pedersen P, Löe H (1969) The correlation between clinical scoring, exudate measurements and microscopic evaluation of inflammation in the gingival. J Periodontol 40:201–209
22. Shaw L, Murray JJ (1977) Diagnostic reproducibility of periodontal indices. J Periodontal Res 12:141–147
23. Spolsky VW, Gornbein JA (1996) Comparing measures of reliability for indices of gingivitis and plaque. J Periodontol 67:853–859
24. Stratford JM (1975) Methods of assessing gingival and periodontal disease: a review. Int J Epidemiol 4:79–86
25. Van der Velden U (1980) Influence of probing force on the reproducibility of bleeding tendency measurements. J Clin Periodontol 7:421–427
26. Van der Weijden GA, Timmerman MF, Saxton CA, Russell JI, Huntington E, Van der Velden U (1994) Intra-/inter-examiner reproducibility study of gingival bleeding. J Periodontal Res 29:236–241
27. White SA, van den Broek NR (2004) Methods for assessing reliability and validity for a measurement tool: a case study and critique using the WHO haemoglobulin colour scale. Stat Med 23:1603–1619