Journal of Clinical Periodontology

Official Journal of The European Federation of Periodontology
Founded by the British, Dutch, French, German, Scandinavian, and Swiss Societies of Periodontology

A comparison of methods of aesthetic assessment in root coverage procedures

Stéphane Kerner[1], Sandrine Katsahian[2], Alexandre Sarfati[1], Stéphane Korngold[1], Sébastien Jakmakjian[1], Bruno Tavernier[3], Fabien Valet[4] and Philippe Bouchard[1]

[1]Department of Periodontology, Service of Odontology, Hôtel-Dieu Hospital, AP-HP, Paris 7 – Denis Diderot University, U.F.R. of Odontology, Paris, France; [2]Department of Biostatistics, Saint Louis Hospital, AP-HP, Paris 7- Denis Diderot University, Paris, France; [3]Department of Prosthetic Dentistry, Service of Odontology, Hôtel-Dieu Hospital, AP-HP, Paris 7 – Denis Diderot University, U.F.R. of Odontology, Paris, France [4]Department of Biostatistics, Institut Curie, Paris, France.

Abstract

Aim: To evaluate the reliability of professional qualitative scoring methods used in evaluating aesthetic results after root coverage therapy and to evaluate the relationship between subjective and objective measurements.

Material and Methods: A review panel of seven professional and non-professional, trained and untrained observers used photographic records to assess the overall cosmetic results of 162 root coverage surgical procedures in 133 patients (mean follow-up 17.51 ± 17.37 months). Two different methods were used. In the before–after panel scoring system, observers evaluated the difference between preoperative and postoperative views, whereas in the random panel scoring system, observers rated each photograph independently.

Results: For both methods, intrarater agreement ranged from substantial to almost perfect for the periodontists. The best interrater agreement was found for trained periodontists using the five-point ordinal scale of the before–after panel scoring system ($\kappa = 0.68$). Neither root coverage percentage nor gingival augmentation was correlated to cosmetic assessment.

Conclusions: The before–after scoring system is an acceptable and reliable method for professional cosmetic assessment of root coverage therapy. The overall cosmetic evaluation does not appear to be related to the percentage of root coverage.

With the widespread adoption of periodontal plastic surgery, trials comparing different types of root coverage therapy are increasingly being conducted. Aesthetic results are recognized as an important outcome; yet these results are not widely incorporated as an outcome measurement in clinical trials (Cairo et al. 2008). The relative contribution of different assessment modalities to the ultimate overall cosmetic outcome will be of great importance in

future trials that compare different types of root coverage therapies (Palmer & Cortellini 2008). Thus, as aesthetics gains importance in trials, it will be necessary to define reproducible and meaningful methods of professional assessment.

In previous studies, cosmetic assessment methods have rarely been used; most authors summarize the aesthetic outcome of root coverage surgeries as ''good'' or ''excellent.'' To our knowledge, only five comparative studies have published consistent aesthetic evaluations by professionals. The methods are highly variable, but all of them use photographic assessment and a scale that divides outcomes into categories. The first comparative trial used a three-

point photographic scale and impressions assessment (poor, moderate and good) by two blind independent observers ($\kappa = 0.70$) (Bouchard et al. 1994). In the study of (Rosetti et al. 2000) five calibrated observers used a three-point scale in the aesthetic assessment. Unfortunately, no information was provided on the calibration process or on observers' agreement . In a comparative clinical trial, (Aichelmann-Reidy et al. 2001) used a four-point scale (poor, fair, good and excellent), and scored the overall aesthetic evaluation according to independent clinicians. The number of clinicians evaluating the results was not indicated; consequently, no information on examiner reliability was available. In the study of Wang et al. (2001)

aesthetic results were scored by one independent periodontist. As a result, the outcomes were dependent on the subjective assessment of a single professional observer. (Cheung & Griffin 2004) asked three independent examiners to rate three variables (colour, texture and contour) using a four-point scale. However, no overall aesthetic evaluation was performed, and no information was provided in the statistical analysis section on the reliability test used.

The paucity of the literature on this specific question suggests that the influence of root coverage therapy on aesthetic outcomes is yet to be completely defined, and that there is no consensus as how best to assess the cosmetic result. Yet, agreement on such a method is a pre-requisite to comparing clinical studies and to unambiguously analysing the impact of various parameters on the aesthetic outcome. The ideal assessment method should rely on simple quantitative measures, and involve a permanent record that can be reviewed, because it is not possible, especially in multi-centre trials, for one panel to evaluate all patients in a live setting.

Photographic assessment appears to lend itself well to the development of image databases, which could use large sample sizes in future trials. The use of photography has been validated previously for quantitative root coverage evaluation, using ImageJ, an image processing program (Kerner et al. 2007). Thus, the subjective information, such as aesthetic outcome, that is available to an observer on a photographic picture can be correlated to objective quantitative measures, such as the percentage of root coverage. Moreover, it may be assumed that the professional assessment of aesthetic results based on photography is independent of patient satisfaction, as compared with live assessment, which may potentially influence observer opinion.

To address this issue, we conducted an analysis in which patients who were previously treated with a variety of root coverage techniques were subjected to two different methods of cosmetic evaluation by a panel of professional and non-professional, trained and untrained observers. With the above purpose in mind, we used the database that was previously analysed to quantitatively evaluate root coverage procedures with an image analysis system (Kerner et al. 2008). The current analysis focuses on qualitative outcomes so as to assess the reliability of a qualitative panel.

The goal of this methodological study was (1) to validate a professional cosmetic assessment method through a qualitative panel scoring system and (2) to describe the relationship between subjective global aesthetic scores and objective quantitative measurements as measured by the ImageJ analysis system.

Material and Methods

In June 2006, a retrospective study on the effect of root coverage therapy was initiated. In December 2006, the collection of the cases was closed, and the database including 363 eligible surgeries, 232 coincident patients and 691 coincident recessions was locked. Eligible patients had at least one buccal Classes 1, 2, 3, or 4 Miller's gingival recession defect to be treated (Miller 1985). More detailed information on the protocol can be found in the publication of the primary analysis (Kerner et al. 2008). To summarize, the sample consisted of consecutive outpatients who underwent a root coverage surgical procedure between 1981 and 2005 in seven practices that were limited to periodontology. Various root coverage surgical techniques were used by seven operators. To keep the description of the sample easily interpretable, the surgical techniques were pooled into the following four categories: (1) pedicle soft tissue graft; (2) non-submerged graft; (3) submerged grafts; and (4) envelope techniques. Categories 1, 2, and 3 correspond to those described by Bouchard et al. (2001). Envelope techniques were defined as all types of submerged grafts without releasing incision. Each patient was documented under standard conditions with pre- and post-operative photographs that were measured using ImageJ for windows. Patient, defect and surgical characteristics as well as surgical indications were recorded for each patient. Percentage of root coverage and gingival augmentation measurements were calculated using ImageJ, a public domain Java image processing program.

Study population

In the present study, a subset of 287 surgeries, 215 coincident patients and 495 coincident recessions were subjected to analysis. Surgeries were suitable for aesthetic assessment if they were documented with two high-quality photographs – one at the date of the surgery and one at least 6 months after the surgical procedure. Exclusion photographic criteria were the following (Fig. 1):

- Lack of visibility of the cementoenamel junction and/or of the muco-gingival line on at least one tooth mesially and distally located to the treated area.
- Difference in colour contrasts and/or framing between the pre-operative and the post-operative views.
- Presence of an edentulous area on any photograph.

Aesthetic assessment

Preoperative and corresponding post-operative slides were digitized under 300 dpi with a scanner, and displayed using Adobe® Photoshop® software (version 7.0 Adobe Systems Europe Ltd., Uxbridge, UK). To be included in the database, the image deformation was calculated using the image analysis system ImageJ for windows, and must be $\leqslant 5\%$. Each pre-operative and post-operative photograph was reframed with Arcsoft® Photostudio5.5®, and imported in presentation software (PowerPoint®, Microsoft, Redmond, WA, USA).

Photographic assessment was performed by a panel of seven observers that included five professionals and two non-professionals. The observers were members of the same institution (Hôtel-Dieu Hospital), and were not involved in the surgical procedures. Photographs were magnified on a screen, and the views were rated in one session by the review panel, which was blind to the patient, the operator and the given treatment. Before starting the evaluation, the panelists attended a briefing to ensure their understanding of the rating form. No time limitation was given to the panelists to evaluate the results, but the photographic assessment of each surgery by each observer took approximately 3 min. to complete. The review panel was asked to score the global cosmetic evaluation for each view. Two different methods of assessment were successively used. Table 1 summarizes the characteristics of the observers and their training according to the evaluation design.
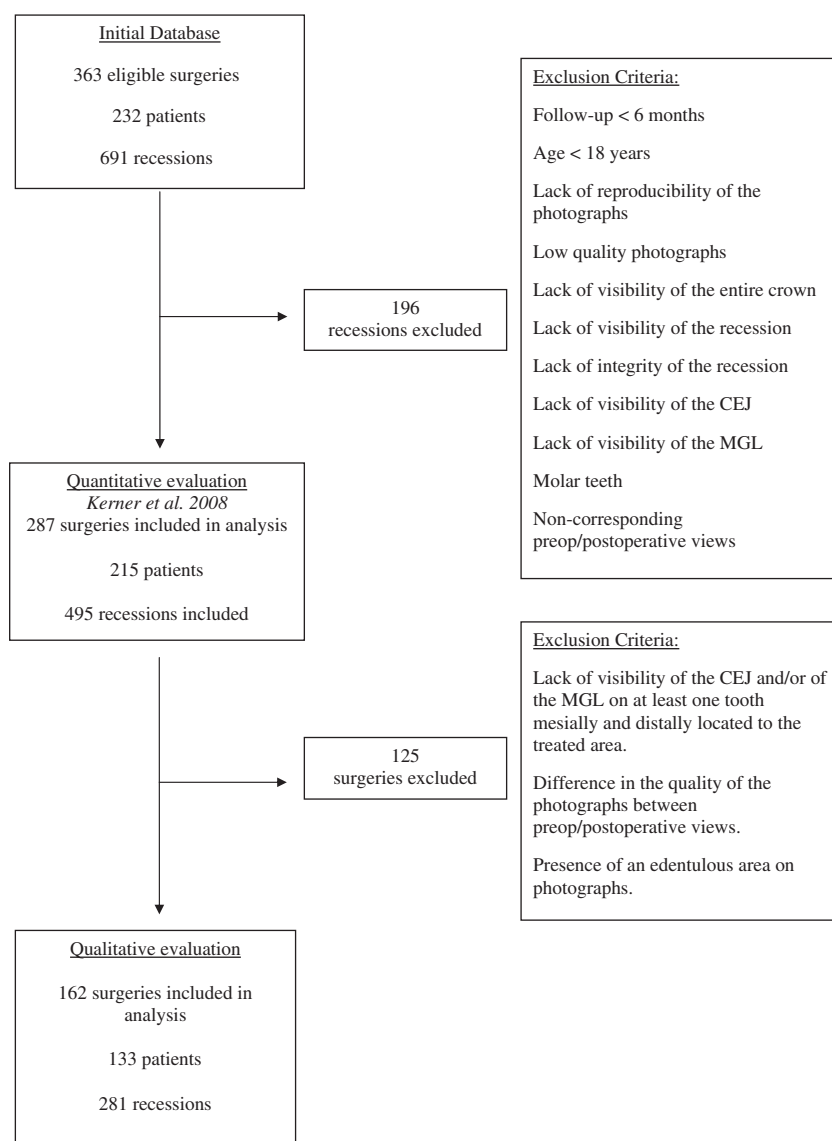
Fig. 1. Diagram of the inclusion cases.

Table 1. Observers' characteristics indicating their training and participation to the methods of cosmetic assessment

Observers			Training	Before–after scoring	Random scoring
Professional	1	Periodontist		•	
	2	Periodontist		•	
	3	Periodontist	•	•	•
	4	Periodontist	•	•	•
	5	Prosthodontist			•
Non-professional	6	Nurse		•	
	7	Nurse			•

Before–after panel scoring system

In this method, observers had to evaluate the overall aesthetic improvement based on the difference between the pre-operative and the post-operative photographs. The pre-operative and the corresponding post-operative photographs were matched in a single view. The views, each one corresponding to one surgical procedure, were randomized using a proprietary randomization program under PowerPoint®. The overall cosmetic result was scored independently on an evaluation form, using the following five-point ordinal improvement scale: poor (1), fair (2), good (3), very good (4) and excellent (5). To evaluate the intraobserver agreement, a set of 20 selected pictures, representative of the five categories, were duplicated and randomly included in the photographic assessment for a double rating. The review panel consisted of four independent professional observers (periodontists) and one non-professional control (nurse). Before starting the evaluation, two out of four periodontists were trained on a standard set of 10 couples of matched photographs.

Random panel scoring system

In this method, the observers had to rate each photograph independently. Photographs, with the same magnification as in the before–after design, were randomized, using the same program. The cosmetic value of each photograph was scored independently, using the following four-point ordinal scale: poor (1), fair (2), good (3) and excellent (4). The overall aesthetic outcome of the root coverage procedure was calculated using the difference between the pre-operative and the post-operative scores. Similar to the before–after panel method, the intraobserver agreement was ensured by a double scoring of 40 photographs representative of the four categories. The review panel consisted of three independent professional observers (two periodontists and one prosthodontist) and one non-professional control (nurse). The two periodontists were also involved in the before–after panel. They were chosen among the four available periodontists due to their training. (Table 1).

Statistical analysis

Collected data were organized into a spreadsheet using a computer program (Excel, Microsoft, Redmond, WA). After proofing for entry errors, the database was locked and loaded in statistical software. The statistician was blind to the given treatments. All statistical tests were performed with R 2.4.1 software (R Foundation for Statistical Computing, Vienna, Austria) on PC architecture. The surgery was used as the unit of analysis. Descriptive statistics were reported as means and standard deviations, or as numbers and percentages.

Table 2. Sample characteristics according to categories of root coverage procedure

Parameter	Unit	Categories of root coverage procedure				
		pedicle soft tissue graft (%)	non-submerged graft (%)	submerged grafts (%)	envelope (%)	total*
Patient						
	Number	26 (17.9)	28 (19.3)	64 (44.2)	27 (18.6)	133
	Mean age (SD)	37.9 (11.46)	37.4 (11.31)	37.6 (12.13)	37.7 (12.05)	37.71 (12.15)
	Male/female	6/20	6/22	15/49	6/21	31/102
Treated defects	Number	42 (14.95)	45 (16.02)	141 (50.17)	53 (18.86)	281
Surgical procedure	Number	30 (18.52)	28 (17.28)	74 (45.68)	30 (18.52)	162
Mean follow-up†	Months (SD)	16.70 (12.20)	18.29 (16.70)	17.21 (14.35)	18.32 (18.53)	17.51 (17.37)

*Each patient being submitted to one or more surgical procedures, the total corresponds to the sample characteristics that may differ from the addition of the subgroup's characteristics.
†Calculated per surgery.

Table 3. Before–after scoring: intrarater and interrater agreement for overall cosmetic results, Cohen's κ statistic (SD)

	Untrained periodontist 1	Untrained periodontist 2	Trained periodontist 3	Trained periodontist 4	Nurse 1
Untrained periodontist 1	0.79 (0.25)	0.51 (0.08)	0.56 (0.08)	0.58 (0.08)	0.3 (0.08)
Untrained periodontist 2		0.66 (0.31)	0.54 (0.10)	0.21 (0.08)	0.21 (0.08)
Trained periodontist 3			0.83 (0.24)	0.68 (0.08)	0.36 (0.08)
Trained periodontist 4				0.83 (0.27)	0.29 (0.08)
Nurse 1					0.42 (0.24)

Agreement within and between the different observers was measured by means of Cohen's weighted κ statistic (Cohen 1968) with Fleiss–Cohen quadratic weights (Fleiss & Cohen 1973). To interpret the level of agreement, a six-level nomenclature was used (Landis & Koch 1977):

- Poor agreement < 0.00.
- Slight agreement $= 0.00$–0.20.
- Fair agreement $= 0.21$–0.40.
- Moderate agreement $= 0.41$–0.60.
- Substantial agreement $= 0.61$–0.80.
- Almost perfect agreement $= 0.81$–0.92.

The relationship between the observers' scores and the quantitative variables (percentage of root coverage and percentage of gingival augmentation) was measured by means of Pearson's correlation coefficients. We used the following crude rule of thumb for interpreting the size of the correlation (Colton 1974):

- 0–0.25: little or no relationship.
- 0.25–0.50: fair degree of relationship.
- 0.50–0.75: moderate to good relationship.
- > 0.75: very good to excellent relationship.

Results

The final sample, after control for inclusion/exclusion criteria, included 162 surgeries, which corresponded to 133 patients and 281 coincident recession defects. Sample characteristics are shown in Table 2. Twenty-four "current" smoking patients were included in the analysis ($\geqslant 5$ cigarettes per day). The mean number of teeth treated per surgery was 1.73 ± 0.89 (median $= 1$). The range of follow-up time per surgery was 6–130 months (median $= 11.72$). The mean percentage of root coverage was $70 \pm 29\%$. Complete root coverage was observed in 32.38% of the defects. The mean percentage of gingival augmentation was $103 \pm 176\%$.

Panel scoring agreement

Before–after panel scoring system

Table 3 indicates the results of the agreement within and between the observers.

Intraobserver variability. Intraobserver agreement of the four periodontists for the global score as measured by the weighted κ was substantial (0.66 and 0.79 for the two untrained periodontists) to almost perfect (0.83 for the trained periodontists). The weighted κ value

was 0.42 for the non-professional control (nurse), showing moderate intraobserver reliability.

Interobserver variability. A specific agreement pattern was observed between trained and untrained periodontists. It was fair to moderate, ranging from 0.21 to 0.58, whereas substantial agreement was found between the trained periodontists as shown by a 0.68 κ value. The reliability between the professional observers and the nurse was fair, with κ values ranging from 0.21 to 0.36.

Random panel scoring system

Preoperative and postoperative scores were analysed independently before being pooled for a global analysis.

Preoperative observer's agreement. Intraobserver agreement was almost perfect for the trained periodontist and the prosthodontist, both showing weighted κ values > 0.81 (Table 4). Intraobserver reliability for the untrained periodontist and the nurse was substantial ($\kappa = 0.67$). The best interobserver agreement was found between the two periodontists. However, this agreement was moderate ($\kappa = 0.55$). The lowest agreement was

Table 4. Random scoring: preoperative intrarater and interrater agreement for overall cosmetic results, Cohen's κ statistic (SD)

	Untrained periodontist 2	Trained periodontist 3	Prosthodontist	Nurse 2
Untrained periodontist 2	0.67 (0.28)	0.55 (0.07)	0.38 (0.08)	0.43 (0.08)
Trained periodontist 3		0.89 (0.25)	0.3 (0.08)	0.37 (0.08)
Prosthodontist			0.85 (0.26)	0.46 (0.08)
Nurse 2				0.67 (0.26)

Table 5. Random scoring: postoperative intrarater and interrater agreement for overall cosmetic results, Cohen's κ statistic (SD)

	Untrained periodontist 2	Trained periodontist 3	Prosthodontist	Nurse 2
Untrained periodontist 2	0.83 (0.23)	0.49 (0.08)	0.53 (0.08)	0.35 (0.08)
Trained periodontist 3		0.88 (0.23)	0.39 (0.08)	0.43 (0.08)
Prosthodontist			0.49 (0.23)	0.14 (0.08)
Nurse 2				− 0.53 (0.20)

Table 6. Random scoring: global intrarater and interrater agreement for overall cosmetic results, Cohen's κ statistic (SD)

	Untrained periodontist 2	Trained periodontist 3	Prosthodontist	Nurse 2
Untrained periodontist 2	0.86 (0.20)	0.66 (0.06)	0.57 (0.06)	0.51 (0.06)
Trained periodontist 3		0.92 (0.18)	0.44 (0.06)	0.48 (0.06)
Prosthodontist			0.73 (0.18)	0.47 (0.06)
Nurse 2				0.32 (0.17)

found between the trained periodontist and the prosthodontist ($\kappa = 0.30$).

Postoperative observer's agreement. Table 5 shows that again, the intrarater agreement between the two periodontists was almost perfect, with κ values > 0.81, whereas that of the prosthodontist was moderate ($\kappa = 0.49$). No intraobserver agreement could be found with the control; the weighted κ for the nurse was a negative value. The best interrater agreement was found between the untrained periodontist and the prosthodontist. Nevertheless, this agreement was moderate ($\kappa = 0.53$). Interestingly, the worst interrater agreement was found between the prosthodontist and the nurse; the weighted κ value was 0.14, the lowest value shown in this report.

Global observer's agreement. Almost perfect intrarater agreement was found between the periodontists (Table 6). The intraobserver agreement was substantial for the prosthodontist, and fair for the nurse. The interobserver agreement was substantial among periodontists ($\kappa = 0.66$). Fair to moderate interobserver agreements were found for the prosthodontist and the nurse, regardless of the observed score.

Relationship between panel scoring and qualitative measurements

Root coverage

The before–after scoring system showed that the percentage of root coverage was significantly correlated with the overall aesthetic evaluation of one trained periodontist (Table 7). Nevertheless, the correlation was weak ($r < 0.50$). With the random scoring system, no significant relationship was found for the preoperative assessment, whatever the observer qualification. However, a significant correlation was found for all the observers for the postoperative assessment, showing little or no relationship for the prosthodontist and the nurse and a fair degree of relationship for the two periodontists. The periodontists only showed a significant but weak relationship for the global assessment ($r = 0.23$, and $r = 0.21$).

Gingival augmentation

A weak to fair significant negative relationship between the percentage of gingival augmentation and the qualitative rating was found with the before–after scoring system for three out of four periodontists (Table 8). No specific relationship pattern was found for the preoperative and postoperative assessment with the random scoring system, except that when the relationship was significant, the r value was negative and low. Little or no significant positive relationship was found for the nurse's evaluation with the global assessment.

Discussion

In attempting to define the optimal methods for the overall cosmetic assessment, we compared two methods of photographic assessment, as well as assessment with scales of different designs and observers with varied levels of training. Our analysis demonstrates acceptable intrarater reliability in evaluating the aesthetic results of root coverage with a method that uses trained professional observers who directly score the difference between pre-operative and post-operative views with a five-point ordinal scale. Indeed, Table 3 shows a specific pattern of agreement among the observers with the before–after panel scoring system, indicating substantial agreement between the two trained periodontists ($\kappa = 0.68$). Table 6 shows that the interrater agreement of the two periodontists with the global random scoring system was somewhat lower ($\kappa = 0.66$). Taking into account that the first method is twice as less time-consuming as the second one, there is no doubt that the before–after scoring system is the preferable method and should be used for aesthetic evaluation.

Photographic assessment was performed by a panel of seven observers representing a good mix of professional background and training. It should be noted that intraobserver and interobserver agreement is generally stronger for the periodontists both in the before–after method and in the random scoring method. This suggests that photographic

Table 7. Pearson's correlation coefficients between the observers' scores and the percentage of root coverage

Observer	Before–after panel scoring system		Random panel scoring system					
			preoperative assessment		postoperative assessment		global assessment	
	r	*p*	*r*	*p*	*r*	*p*	*r*	*p*
Untrained periodontist 1	0.13	NS	NA	NA	NA	NA	NA	NA
Untrained periodontist 2	0.06	NS	0.12	NS	0.37	$<10^{-3}$	0.23	0.004
Trained periodontist 3	0.21	0.007	0.10	NS	0.30	$<10^{-3}$	0.21	0.008
Trained periodontist 4	0.10	NS	NA	NA	NA	NA	NA	NA
Nurse 1	0.14	NS	NA	NA	NA	NA	NA	NA
Prosthodontist	NA	NA	0.14	NS	0.21	0.006	0.06	NS
Nurse 2	NA	NA	0.04	NS	0.22	0.004	0.14	NS

NS, non significant; NA, not applicable.

Table 8. Pearson's correlation coefficients between the observers' scores and the percentage of keratinized tissue

Observer	Before–after panel scoring system		Random panel scoring system					
			preoperative assessment		postoperative assessment		global assessment	
	r	*p*	*r*	*p*	*r*	*p*	*r*	*p*
Untrained periodontist 1	− 0.10	NS	NA	NA	NA	NA	NA	NA
Untrained periodontist 2	− 0.24	0.002	− 0.22	0.004	− 0.04	NS	0.15	NS
Trained periodontist 3	− 0.33	$<10^{-3}$	− 0.14	NS	− 0.21	0.005	− 0.11	NS
Trained periodontist 4	− 0.29	$<10^{-3}$	NA	NA	NA	NA	NA	NA
Nurse 1	− 0.11	NS	NA	NA	NA	NA	NA	NA
Prosthodontist	NA	NA	− 0.11	NS	− 0.09	NS	0.02	NS
Nurse 2	NA	NA	− 0.20	0.010	− 0.03	NS	0.16	0.040

NS, non significant; NA, not applicable.

assessment is a well-adapted method for periodontology specialists. However, the amount of information available to an observer in a photograph is limited to aesthetic appearance, and cannot replace the live clinical assessment. For example, neither the clinical attachment level nor the probing depth is measurable. This limitation has been further discussed in a previous report dealing with qualitative measurements (Kerner et al. 2007).

A fair to moderate intrarater and interrater agreement was consistently found with the controls, regardless of the method used. Interestingly, Table 4 shows that the preoperative intrarater agreement of the nurse was substantial ($\kappa = 0.67$), whereas Table 5 indicates that the corresponding postoperative value was negative ($\kappa = -0.53$). This means that it was relatively easy for the nurse to score the initial situation, whereas she was not able to correctly score the result of the surgical procedure.

In any evaluation process, examiner training is key to reliability. The present study design offers an insight into whether a learning curve for the process of photographic assessment may exist. Tables 3–6 indicate that trained periodontists have the best intrarater and interrater agreements as compared with other observers without formal training. Furthermore, the before–after evaluation was performed before the random evaluation. Two periodontists participated in both assessment protocols. A substantial improvement in the intrarater and interrater agreement was found among these observers during the second protocol. This suggests a learning effect that should be taken into account in further studies. This observation is in accordance with other research showing the importance of training in improving the reliability of observers' judgements of the quality of medical care (Koran 1975).

The system chosen for scoring the results is also critical to evaluate observer agreement. In the present study, we do not use a visual analogue scale (VAS) because the use of VAS implies that continuous data must be converted to categories to be analysed using the κ statistic. Furthermore, the literature aiming to evaluate aesthetic results shows that VAS is a less reliable tool than an ordinal scale (Lowery et al. 1996). The literature dealing with aesthetic assessment shows that optimal aesthetic evaluation using ordinal scales can be obtained by a four-point ordinal scale (Harris et al. 1979, Sneeuw et al. 1992). However, these evaluations cannot be taken as gold standards because they do not specifically deal with dental aesthetics in general and periodontal plastic surgeries in particular. A five-point ordinal scale was chosen in the before–after protocol to improve the accuracy of cosmetic assessment. Because of the relatively modest κ values of the before–after method, the number of scale categories was reduced in order to limit the dispersion of the rating and to potentially improve the level of agreement. Thus, the decision was made to change this five-point ordinal score in the random protocol after the

before–after evaluation was completed. The results indicate that reduction of the magnitude of the scale does not compensate the lack of direct comparative evaluation of the random method; this definitely weakens the interobserver agreement compared with the before–after method (Tables 3 and 6). Thus, a five-point ordinal scale may be used for cosmetic assessment in root coverage studies.

Our results show that neither the percentage of root coverage nor the percentage of gingival augmentation was correlated to subjective scores (Tables 7 and 8). It may be concluded that the amount of root coverage is not the most critical variable in the overall aesthetic judgement of the observers. This conclusion is in contrast with an interesting survey using photographic simulation, which shows that complete root coverage was perceived as the most successful outcome by periodontists and patients (Rotundo et al. 2008). In this study, three variables were modified and combined in a set of clinically simulated images; these variables included: recession depth, colour of the root and amount of root coverage. The difference in the conclusion between the study of Rotundo et al. and ours may be related to (1) the study design, and (2) the fact that the variables in the simulation study analysis did not take into account the soft tissue appearance that accounts for the overall clinical assessment. Consequently, root coverage percentage cannot be the main goal of root coverage surgeries that aim to improve the global aesthetic appearance of patients' smiles.

Although a reasonably substantial agreement was found between trained periodontists, concordance was far from perfect. Furthermore, the fact that the examiners belonged to the same institution may have influenced their judgement, and improved the rate of agreement. This suggests the identification and quantification of certain treatment outcome variables, such as those described previously (Bouchard et al. 2001) and that may be part of the overall aesthetic judgement. Objective quantitative measurements, such as the percentage of root coverage and the percentage of gingival augmentation, offer the advantage of reliability. However, they do not evaluate parameters that account for the global aesthetic evaluation (i.e. scarring, texture, volume, colour, gingival contour, location of the mucogingival line and root

colour). The subjective evaluation of root coverage is per se imperfect, and more explicit criteria are required to improve the reliability of scales for aesthetic assessment.

It may be assumed that patient judgement is a key factor in evaluating aesthetic outcomes. From a methodological point of view, patient judgement is one of the most difficult approaches because aesthetics is considered as part of overall patient satisfaction, which includes overlapping subjective variables (Sacchini et al. 1991). Reliable self-assessment methods are needed in the field of periodontology. It is the patient who judges surgery results, while it is the surgeon who selects the technique used; as such, professionals may develop different opinions concerning aesthetic appearance than patients. Consequently, future research should explore statistical models that combine professional, objective and subjective measurements as well as patient satisfaction. The present report is a first step towards a better understanding of aesthetic assessment in periodontal plastic surgery, and may help professionals interpret the results of studies that use them, especially for comparative purposes.

It can be concluded from this study that photographic assessment of quantifiable outcome variables is a useful method to compare treatment outcomes in root coverage trials when cosmetic outcome is important. Results that are reasonably reproducible can be obtained by periodontists with limited training. The assessment should be performed on the direct evaluation of the difference between preoperative and postoperative views, and not on the evaluation of each photograph *a posteriori* compared with the scores. A five-point ordinal scale is a valuable and recommended tool for subjective assessment of root coverage therapy.

References

Aichelmann-Reidy, M. E., Yukna, R. A., Evans, G. H., Nasr, H. F. & Mayer, E. T. (2001) Clinical evaluation of acellular allograft dermis for the treatment of human gingival recession. *Journal of Periodontology* **72**, 998–1005.

Bouchard, P., Etienne, D., Ouhayoun, J. P. & Nilvéus, R. (1994) Subepithelial connective tissue grafts in the treatment of gingival recessions. A comparative study of 2 procedures. *Journal of Periodontology* **65**, 929–936.

Bouchard, P., Malet, J. & Borghetti, A. (2001) Decision-making in aesthetics: root coverage revisited. *Periodontology 2000* **27**, 97–120.

Cairo, F., Pagliaro, U. & Nieri, M. (2008) Treatment of gingival recession with coronally advanced flap procedures: a systematic review. *Journal of Clinical Periodontology* **35** (Suppl. 8), 136–162.

Cheung, W. S. & Griffin, T. J. (2004) A comparative study of root coverage with connective tissue and platelet concentrate grafts: 8-month results. *Journal of Periodontology* **75**, 1678–1687.

Cohen, J. (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* **70**, 213–220.

Colton, T. (1974) *Statistics in Medicine.* Boston: Little, Brown and Company.

Fleiss, J. L. & Cohen, J. (1973) The equivalence of weighted kappa and the intra-class correlation coefficient as measures of reliability. *Educational and Psychological Measurement* **33**, 613–619.

Harris, J. R., Levene, M. B., Svensson, G. & Hellman, S. (1979) Analysis of cosmetic results following primary radiation therapy for stage I and II carcinoma of the breast. *International Journal of Radiation, Oncology, Biology Physics* **5**, 257–261.

Kerner, S., Borghetti, A., Katsahian, S., Etienne, D., Malet, J., Mora, F., Monnet-Corti, V., Glise, J. M. & Bouchard, P. (2008) A retrospective study of root coverage procedures using an image analysis system. *Journal of Clinical Periodontology* **35**, 346–355.

Kerner, S., Etienne, D., Malet, J., Mora, F., Monnet-Corti, V. & Bouchard, P. (2007) Root coverage assessment: validity and reproducibility of an image analysis system. *Journal of Clinical Periodontology* **34**, 969–976.

Koran, L. M. (1975) The reliability of clinical methods, data and judgments (second of two parts). *The New England Journal of Medicine* **293**, 695–701.

Landis, J. R. & Koch, G. G. (1977) An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* **33**, 363–374.

Lowery, J. C., Wilkins, E. G., Kuzon, W. M. & Davis, J. A. (1996) Evaluations of aesthetic results in breast reconstruction: an analysis of reliability. *Annals of Plastic Surgery* **36**, 601–606.

Miller, P. D. Jr. (1985) A classification of marginal tissue recession. *International Journal of Periodontics and Restorative Dentistry* **5**, 8–13.

Palmer, R. M. & Cortellini, P. (2008) Periodontal tissue engineering and regeneration: consensus report of the sixth European workshop on periodontology. *Journal of Clinical Periodontology* **35** (Suppl. 8), 83–86.

Rosetti, E. P., Marcantonio, R. A., Rossa, C. Jr., Chaves, E. S., Goissis, G. & Marcantonio, E. Jr. (2000) Treatment of gingival recession: comparative study between subepithelial connective tissue graft and guided tissue regeneration. *Journal of Periodontology* **71**, 1441–1447.

Rotundo, R., Nieri, M., Mori, M., Clauser, C. & Pini Prato, G. (2008) Aesthetic perception after root coverage procedure. *Journal of Clinical Periodontology* **38**, 705–712.

Sacchini, V., Luini, A., Tana, S., Lozza, L., Galimberti, V., Merson, M., Agresti, R., Veronesi, P. & Greco, M. (1991) Quantitative and qualitative cosmetic evaluation after conservative treatment for breast cancer. *European Journal of Cancer* **27**, 1395–1400.

Sneeuw, K. C., Aaronson, N. K., Yarnold, J. R., Broderick, M., Regan, J., Ross, G. & Goddard, A. (1992) Cosmetic and functional outcomes of breast conserving treatment for early stage breast cancer. 1. Comparison of patients' ratings, observers' ratings, and objective assessments. *Radiotherapy and Oncology* **25**, 153–159.

Wang, H. L., Bunyaratavej, P., Labadie, M., Shyr, Y. & MacNeil, R. L. (2001) Comparison of 2 clinical techniques for treatment of gingival recession. *Journal of Periodontology* **72**, 1301–1311.

Address:
Philippe Bouchard
U.F.R. d'Odontologie
5, rue Garancière
75006 Paris
France
E-mail: phbouch@noos.fr

Clinical Relevance

Scientific rationale for the study: Aesthetic outcome is an important outcome of root coverage therapies. A unique and universal system to evaluate cosmetic results should be defined and regularly included in clinical trials comparing different types of root coverage surgeries.

Principal findings: The before–after scoring panel system based on photographic assessment is a reasonably reproducible method of aesthetic assessment that may be used by periodontists with limited training.

Practical implications: This method can be advised for clinicians and researchers in root coverage aesthetic evaluation. It may help evaluate and select the surgical procedure that best corresponds to the individual patient's aesthetic request.