Are colored periodontal probes reliable to classify the gingival phenotype in terms of gingival thickness?

Kristina Bertl¹,² | Mehdi Al-Hotheiry¹ | David Sun¹ | John Olofsson¹ | Stefan Lettner³,⁴ | Klaus Gotfredsen⁵ | Andreas Stavropoulos¹,⁶,⁷

¹ Department of Periodontology, Faculty of Odontology, University of Malmö, Malmö, Sweden
² Division of Oral Surgery, University Clinic of Dentistry, Medical University of Vienna, Vienna, Austria
³ Division of Oral Surgery, Karl Donath Laboratory for Hard Tissue and Biomaterial Research, University Clinic of Dentistry, Medical University of Vienna, Vienna, Austria
⁴ Austrian Cluster for Tissue Regeneration, Vienna, Austria
⁵ Department of Oral Rehabilitation, School of Dentistry, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
⁶ Division of Regenerative Dentistry and Periodontology, University Clinics of Dental Medicine (CUMD), University of Geneva, Geneva, Switzerland
⁷ Division of Conservative Dentistry and Periodontology, University Clinic of Dentistry, Medical University of Vienna, Vienna, Austria

Abstract

Background: This cross-sectional study assessed the potential of colored periodontal probes (CPP) to classify gingival phenotype in terms of gingival thickness (GT).

Methods: Buccal GT in three anterior teeth in each of 50 patients was measured by transgingival sounding and classified by three different methods by eight examiners. Specifically, the diagnostic potential of visual judgment and transparency of a standard periodontal probe (SPP) to discriminate thin and thick gingiva, and of CPP to discriminate thin, medium, thick, or very thick gingiva was assessed.

Results: GT ranged from 0.57 to 2.37 mm. Using CPP resulted in a medium judgment in 87% of the cases, on average, and only between 1-10 cases/examiner were judged as thick or very thick. Considering 1 mm GT as relevant cut-off value, all methods showed a high positive predictive value (≥ 0.82) to identify thick cases, but also a high false omission rate (≥ 0.73) indicating that many cases classified as thin were actually thick. Further, 88% of the cases being ≤ 1 mm, were not classified as thin with CPP; this was inferior to SPP, for which, however, still 64% of the cases being ≤ 1 mm thick were wrongly classified. The highest, yet moderate agreement among examiners was achieved by SPP (κ = 0.427), whereas visual judgment and CPP showed only fair (κ = 0.211) and slight agreement (κ = 0.112), respectively.

Conclusion: Using CPP resulted in most of the cases in a medium judgment. It seems that CPP cannot distinctly discriminate between “thick” and “very thick” cases and fails to capture the thin high-risk cases.

Keywords
diagnosis, gingiva, phenotype, reproducibility of results, sensitivity and specificity
INTRODUCTION

Most patients demand for prosthetic restorations not only a functional, but also a highly aesthetic outcome for which periodontal and peri-implant phenotype are important parameters.\(^1,2\) Periodontal phenotype comprises bone morphotype and gingival phenotype, the latter being a composite of gingival thickness (GT) and keratinized tissue width (KTW),\(^2\) whereas peri-implant phenotype is composed of keratinized mucosa width, mucosal thickness, supracrestal tissue height, and peri-implant bone thickness.\(^1\) GT and mucosal thickness have an impact on the outcome of various clinical procedures, including surgical, procedures around natural teeth\(^2\) and dental implants,\(^1\) respectively. For example, a thin compared to thick gingiva is associated with a reduced rate of complete root coverage after mucogingival surgery.\(^1,5\) Therefore, assessment of GT and mucosal thickness around natural teeth and dental implants is relevant in the clinic for treatment planning and in research for reliable assessment of new clinical procedures and/or dental materials.\(^1,2,6\)

Various methods for GT assessment/classification are described in the literature. The most common method is the placement of a standard periodontal probe (SPP) into the gingival sulcus at the midfacial aspect of the tooth.\(^7–17\) If the SPP is visible/shining through the tissue at the gingival margin, the gingiva is considered “thin,” whereas it is judged “thick” if the probe is invisible.\(^13\) However, the probe transparency method has a certain subjective component; in a previous study complete agreement in classifying GT as thin or thick with SPP was achieved in 80% of the cases among three evaluators.\(^12\) Other methods, such as visual judgment or transgingival sounding are either even more challenging and unreliable\(^6,18\) or invasive\(^12,19\). Hence, there is a need for non-invasive methods to measure/classify GT and amongst ultrasonography,\(^20\) caliper,\(^9,16\) and radiographic assessment\(^17\) also a new set of colored periodontal probes (CPP) for the probe transparency method has been introduced.\(^21\) This CPP set consists of three probes, with a white, green, or blue colored tip (see Supplementary Figure 1 in online Journal of Periodontology); depending on which type of CPP is visible (i.e., the white, green, blue, or none), GT can supposedly be classified into four categories, i.e., “thin,” “medium,” “thick,” and “very thick.”

A recent methodological study comparing different methods for GT assessment included the CPP set;\(^22\) at 400 lower central incisors the judgment of GT with CPP was in about 70% of the cases “medium” and only 11 out of 20 repeated judgments resulted in the same category. Although these data question somewhat the diagnostic potential of CPP, data on the upper anterior teeth, which are aesthetically more relevant, are still lacking. Therefore, the aim of the present study is to assess the potential of CPP to classify correctly, repeatably, and reproducibly gingival phenotype in terms of GT in upper anterior teeth.

MATERIALS AND METHODS

Study population

The present cross-sectional study was approved by the regional ethical review board of Lund (Sweden; Dnr. 2017/679) and performed between October and December 2017 at the Department of Periodontology (Faculty of Odontology, University of Malmö, Sweden); reporting complies with the STROBE guidelines\(^23\) (see Supplementary Table 1 in online Journal of Periodontology). A convenience sample of 50 participants, which was based on previous publications with a similar study design,\(^10,14\) was recruited according to the following eligibility criteria: (1) presence of a central and lateral incisor and canine; (2) no buccal filling, veneer or crown restoration; (3) no signs of gingival inflammation and/or probing pocket depths (PD) > 4 mm; (4) ≥ 2 mm KTW; (5) no highly pigmented marginal gingiva; (6) no intake of any medication affecting the gingiva (e.g., amiodipine, Cyclosporine A, hydantoin); and (7) not pregnant. Oral and written informed consent was obtained from all participants prior to any intervention and gender and age were recorded.

Methods of GT assessment/classification

GT was assessed by four different methods at one maxillary central and lateral incisor and canine in each participant: (1) visual judgment; (2) probe transparency with SPP;\(^2\) (3) probe transparency with CPP;\(^1\); and (4) transgingival sounding with an endodontic file. Judgment of the first three methods was made on standardized intraoral photographs, which were taken at room light (but without direct light of the dental unit), with a digital camera mounted with a ring flash.\(^3\) Photographs were taken after drying the tooth, with a black contrastor placed palatal to the tooth, and with a magnification ratio of 1/1 and a fixed distance of 12 cm, perpendicular to the buccal aspect of the tooth. The photographer ensured that the digital camera and the patient’s Frankfurt horizontal plane were both parallel to the ground. The following five shots of the

\(^1\) CP-12, Hu-Friedy Mfg. Co., LLC, Chicago, IL, USA
\(^2\) Colorvue® biotype probe, Hu-Friedy Mfg. Co., LLC, Chicago, IL, USA
\(^3\) Canon EOS 1000D, Canon SA, Tokyo, Japan
Tooth of interest including approximately 4 to 5 mm of the marginal gingiva have been captured (Figure 1): (1) without any probe; (2) with SPP; and 3-5) with each CPP (i.e., white, green, and blue). Each probe was placed by a single calibrated examiner with a force of approximately 0.2 N into the sulcus at the midfacial aspect of the tooth of interest. After taking all intraoral photographs GT was measured at each included tooth by transgingival sounding (see Supplementary Figure 2 in online Journal of Periodontology). After application of local anesthesia gel§ an endodontic file (ISO 20) mounted with a silicon stop was inserted perpendicularly into the buccal aspect of the gingiva (1 mm below the gingival margin) until touching the tooth surface. The insertion depth was secured by placing the silicone stop in contact with the gingiva and fixing its position with flowable composite**. After removal of the endodontic file, a picture including a ruler was taken and the insertion depth was measured in an image analysis program†† after calibrating the program using the included ruler.

2.3 Judgment of the intraoral photographs

All photographs were imported into an image editing program‡‡ and by means of a grey card, which was included in each series of photographs, the white balance was adjusted. The photographs without a probe were cropped to the marginal aspect but included both adjacent papillae, whereas the photographs with a probe were cropped to display only the marginal gingiva. The probes and 0.5 mm of the marginal gingiva were masked artificially to prevent any distraction/bias by seeing the type/color of the probe not covered by the gingiva (Figure 1 and 2). All photographs were uploaded in a random order to an online questionnaire service§§. For verification of the “probe masking” method, 50 randomly chosen photographs without a probe were included twice in this catalogue; these duplicate images incorporated the above-described artificial probe masking, to “pretend/simulate” the presence of a probe (Figure 1 and 2). The final catalogue consisted of 800 photographs (i.e., five photographs per tooth, three teeth per participant, 50 participants, and 50 photographs twice). In the photographs without a probe, a judgment as either “thin gingiva” or “thick gingiva” was required, whereas in all other photographs the options were “probe is visible” or “probe is invisible.” The examiners were instructed to judge whether the color of the probe is shining through the marginal gingiva, but not to judge the “visibility” of the presence of a probe, e.g., by identifying any wrinkles at the gingival surface because of inserting the probe into the sulcus (Figure 2).

The questionnaire was sent out to eight examiners without known colorblindness together with a tablet computer***. The screen settings of the tablet computer were predefined (i.e., set at highest brightness and “true tone” deactivated) and the function of the tablet computer was restricted to the survey page; suitability of a tablet computer for such analysis has been previously reported‡‡. All examiners were informed that the judgment will take in total about 3 to 4 hours, but they should avoid continuous judgment beyond 60 minutes. The eight examiners had different backgrounds (i.e., two last-year dental students, two

---

§ Xylocain 2%, AstraZeneca AB, Södertälje, Sweden
** Tetric EvoFlow, Ivoclar Vivadent AG, Schaan, Lichtenstein
†† Photoshop CC, Version 20.0.2, Adobe Systems, San Jose, CA, USA
‡‡ Adobe® Lightroom Classic 8.1, Adobe Systems, San Jose, CA, USA
*** SUNET - Swedish University Computer Network, Stockholm, Sweden
**** iPad Pro 10.5”, Apple Inc., Cupertino, CA, USA
All probes were artificially masked including 0.5 mm of the marginal gingiva (i.e., the red part of the masking) to prevent any distraction by the color of the probe not covered by the gingiva (A, B). For verification of the method, 50 randomly chosen photographs without a probe (C) were included twice in this catalogue; however, in the second version a masking in the same manner as in the photographs with the probes was included to “pretend” the presence of a probe (D). Additionally, in some cases, inserting the probe into the sulcus caused wrinkles of the gingiva (E versus F); the examiners were instructed not to judge the “visibility” of a probe based on the presence of wrinkles of the gingiva, but to judge whether the color of the probe shines through the marginal gingiva.

Prothodontists, two periodontists, and two general practitioners; one randomly chosen examiner of each of these four groups repeated after 2 weeks half of the questionnaire (i.e., 400 randomly chosen photographs).

### 2.4 Statistical analysis

For statistical analysis, a visible SPP was interpreted as “thin gingiva” and an invisible SPP as “thick gingiva.” The classification based on the CPP allowed four categories, i.e., “thin” gingiva (the probe with the white tip is visible through the gingival margin), “medium” gingiva (the probe with the white tip is not visible, but the probe with the green tip is visible), “thick” gingiva (the probes with the white and green tip are not visible, but the probe with the blue tip is visible), and “very thick” gingiva (none of the probes are visible). Further, for specific parts of the analysis the cases were categorized into four categories according to their GT measured with the endodontic file: <1 mm, ≥1 to <1.25 mm, ≥1.25 to <1.5 mm, and ≥1.5 mm thick.

Based on a confusion matrix including 150 judgments of eight examiners, positive predictive value (PPV), false omission rate (FOR), true positive rate (TPR), and false positive rate (FPR) were calculated for all assessment methods. For these parameters, 95% confidence intervals (CI) are provided, estimated by resampling 10,000 times with replacement on the patient level (i.e., a clustered bootstrap) as described by Rutter 2000 and applying the percentile method. Further, plots of PPV versus FOR are provided, in which GT measurements with the endodontic file are considered as the “true” values. The ideal case would be to achieve PPV of 1 and FOR of 0, whereas the worst case...
would be a PPV of 0 and FOR of 1. These plots correspond to the well-known receiver operating characteristic (ROC) plots of sensitivity (or TPR) versus 1 – specificity (or FPR), in which the results based on the visual assessment, SPP, or CPP are considered as the “true” results. In other words, the plots presented herein differ from the ones shown by others only in the axis labeling, which was chosen herein to stress that GT measurements should actually be considered as “true” values. Inter-examiner reproducibility and intra-examiner repeatability of GT evaluation was tested with Fleiss’ kappa and the level of agreement was assessed according to a 6-level ranking: (1) poor agreement: < 0.0; (2) slight agreement: 0.0 to 0.2; (3) fair agreement: > 0.2 to 0.4; (4) moderate agreement: > 0.4 to 0.6; (5) substantial agreement: > 0.6 to 0.8; and (6) almost perfect agreement: > 0.8 to 1.0.

Finally, approximately normal distribution of GT measurements was confirmed graphically by Q-Q plots and descriptive analysis was performed for the assessed parameters (i.e., mean, standard deviation, minimum, maximum, and frequency distribution). Statistical analyses were performed with two programs (SPSS Version 25.0, Chicago, IL, USA and R version 4.0.5, The R Foundation, Vienna, Austria); a p-value ≤ 0.05 was considered statistically significant.

3 | RESULTS

Twenty-six males and 24 females (mean age: 25.44 ± 4.97 years, range: 18-39 years) contributed with three teeth each (i.e., a central and lateral incisor and canine). GT assessed by transgingival sounding ranged from 0.57 to 2.37 mm with the central incisor presenting the highest mean value (i.e., 1.44 mm). Further, out of the 150 teeth, 27 teeth presented GT of < 1 mm, 44 GT ≥ 1 to < 1.25 mm, 43 GT ≥ 1.25 to < 1.5 mm, and 36 GT ≥ 1.5 mm, respectively.

3.1 | Distribution of the GT categories

Based on the visual judgment between 30.7 and 79.3% of the cases have been judged as “thin” by the various examiners; this rate was lower when judging the probe transparency with SPP (i.e., between 8.7 and 46% were judged as “thin”). Using the CPP grading most of the cases accumulated in the “medium” category (i.e., 65.3-96%) with six out of eight examiners judging > 85% of the cases as “medium”; further, only 1 to 10 cases per examiner were judged as either “thick” or “very thick” (Table 1). Additionally, among all eight examiners, the green CPP was

---

### TABLE 1

<table>
<thead>
<tr>
<th>Judgment Category</th>
<th>Category</th>
<th>Visual judgment</th>
<th>CPP grading</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Thin/Thick</td>
<td>Thin (visible)/Thick (invisible)</td>
<td>Thin/Medium/Thick/Very Thick</td>
</tr>
<tr>
<td></td>
<td>#1</td>
<td>#2</td>
<td>#1</td>
</tr>
<tr>
<td></td>
<td>Dental student</td>
<td>General practitioner</td>
<td>Periodontist</td>
</tr>
<tr>
<td></td>
<td>#1</td>
<td>#2</td>
<td>#1</td>
</tr>
<tr>
<td>Visual</td>
<td>98</td>
<td>92</td>
<td>66</td>
</tr>
<tr>
<td>judgment</td>
<td>104</td>
<td>104</td>
<td>94</td>
</tr>
<tr>
<td>SPP</td>
<td>42</td>
<td>32</td>
<td>27</td>
</tr>
<tr>
<td>Thin/sparetissue</td>
<td>137</td>
<td>144</td>
<td>137</td>
</tr>
<tr>
<td>CPP grading</td>
<td>42</td>
<td>42</td>
<td>0</td>
</tr>
</tbody>
</table>

CPP, colored periodontal probes; SPP, standard periodontal probe.
judged as “visible” 20-times, although the blue CPP was judged as “invisible” for the same case. This represents an “incorrect” sequence, i.e., the blue CPP should not have been judged as invisible, once the green CPP was already visible. Considering that the green and blue CPP were judged as “invisible” only 39-times in total, these 20 cases with an “incorrect” sequence represented a high percentage. An overview on the frequency distribution of each method divided into four GT categories (i.e., < 1 mm, ≥ 1 to < 1.25 mm, ≥ 1.25 to < 1.5 mm, and ≥ 1.5 mm) is presented in Supplementary Table 2 in the online Journal of Periodontology, and an overview of the mean GT values per category and per method is provided in Supplementary Table 3 in the online Journal of Periodontology.

3.2 Parameters derived from the confusion matrix

For each assessment method PPV, FOR, TPR, and FPR were calculated by combining the data of all examiners (Table 2, Figure 3). Considering GT of 1 mm as relevant cut-off value, the following values were derived:

- Visual judgement: PPV 0.84 (CI: 0.78; 0.91), FOR 0.78 (CI: 0.71; 0.86), TPR 0.49 (CI: 0.44; 0.54), and FPR 0.39 (CI: 0.30; 0.48)

This is indicating that (1) 84% of the cases classified as thick were correctly identified being > 1 mm (PPV), but (2) 78% of the cases classified as thin, were actually > 1 mm (FOR); further, (3) 49% of the cases being > 1 mm, were correctly classified as thick (TPR), and (4) 39% of the cases being ≤ 1 mm, were wrongly classified as thick (FPR).

- SPP: PPV 0.84 (CI: 0.78; 0.90), FOR 0.73 (CI: 0.62; 0.84), TPR 0.77 (CI: 0.72; 0.82), and FPR 0.64 (CI: 0.52; 0.76)

This is indicating that (1) 84% of the cases classified as thick were correctly identified being > 1 mm (PPV), but (2) 73% of the cases classified as thin, were actually > 1 mm (FOR); further, (3) 77% of the cases being > 1 mm, were correctly classified as thick (TPR), and (4) 64% of the cases being ≤ 1 mm, were wrongly classified as thick (FPR).

- CPP grading thin vs. medium/thick/very thick: PPV 0.82 (CI: 0.75; 0.88), FOR 0.74 (CI: 0.70; 0.90), TPR 0.89 (CI: 0.88; 0.92), and FPR 0.88 (CI: 0.85; 0.94)

This is indicating that (1) 82% of the cases classified as medium/thick/very thick were correctly identified being > 1 mm (PPV), but (2) 74% of the cases classified as thin, were actually > 1 mm (FOR); further, (3) 89% of the cases being > 1 mm, were correctly classified as medium/thick/very thick (TPR), and (4) 88% of the cases being ≤ 1 mm, were wrongly classified as medium/thick/very thick (FPR).

For the comparison “CPP grading thin/medium vs. thick/very thick” and “CPP grading thin/medium/thick vs. very thick” a higher cut-off value was considered as relevant to display the aim to identify correctly thicker cases, i.e., based on 1.5 mm GT as cut-off, the following values were derived:

- CPP grading thin/medium vs. thick/very thick: PPV 0.39 (CI: 0.11; 0.62), FOR 0.24 (CI: 0.17; 0.31), TPR 0.05 (CI: 0.01; 0.09), and FPR 0.03 (CI: 0.01; 0.05)

This is indicating that (1) 39% of the cases classified as thick/very thick were correctly identified being > 1.5 mm (PPV), but (2) 24% of the cases classified as thin/medium, were actually > 1.5 mm (FOR); further, (3) 5% of the cases being > 1.5 mm, were correctly classified as thick/very thick (TPR), and (4) 3% of the cases being ≤ 1.5 mm, were wrongly classified as thick/very thick (FPR).

Figure 3: Plots of positive predictive value (PPV) versus false omission rate (FOR), in which gingival thickness measurements with the endodontic file are considered as the “true” values. The ideal case would have been to achieve PPV of 1 and FOR of 0, whereas the worst case would have been the opposite with PPV of 0 and FOR of 1.
<table>
<thead>
<tr>
<th>Gingival thickness (mm)</th>
<th>Visual judgment</th>
<th>SPP</th>
<th>CPP grading (thin vs. medium/thick/very thick)</th>
<th>CPP grading (thin/medium/thick vs. very thick)</th>
<th>CPP grading (thin/medium/thick vs. very thick)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PPV</td>
<td>FOR</td>
<td>TPR</td>
<td>FPR</td>
<td>PPV</td>
</tr>
<tr>
<td>0.80</td>
<td>0.99</td>
<td>0.97</td>
<td>0.47</td>
<td>0.25</td>
<td>1.00</td>
</tr>
<tr>
<td>0.90</td>
<td>0.96</td>
<td>0.90</td>
<td>0.49</td>
<td>0.26</td>
<td>0.95</td>
</tr>
<tr>
<td>1.00</td>
<td><strong>0.84</strong></td>
<td>0.78</td>
<td><strong>0.49</strong></td>
<td>0.39</td>
<td><strong>0.84</strong></td>
</tr>
<tr>
<td>1.10</td>
<td>0.76</td>
<td>0.66</td>
<td>0.50</td>
<td>0.39</td>
<td>0.74</td>
</tr>
<tr>
<td>1.20</td>
<td>0.64</td>
<td>0.54</td>
<td>0.51</td>
<td>0.42</td>
<td>0.62</td>
</tr>
<tr>
<td>1.30</td>
<td>0.52</td>
<td>0.42</td>
<td>0.52</td>
<td>0.42</td>
<td>0.50</td>
</tr>
<tr>
<td>1.40</td>
<td>0.38</td>
<td>0.31</td>
<td>0.52</td>
<td>0.44</td>
<td>0.38</td>
</tr>
<tr>
<td>1.50</td>
<td>0.23</td>
<td>0.24</td>
<td>0.47</td>
<td>0.27</td>
<td>0.26</td>
</tr>
<tr>
<td>1.60</td>
<td>0.15</td>
<td>0.16</td>
<td>0.45</td>
<td>0.47</td>
<td>0.18</td>
</tr>
<tr>
<td>1.70</td>
<td>0.09</td>
<td>0.08</td>
<td>0.49</td>
<td>0.47</td>
<td>0.10</td>
</tr>
<tr>
<td>1.80</td>
<td>0.06</td>
<td>0.06</td>
<td>0.49</td>
<td>0.47</td>
<td>0.07</td>
</tr>
<tr>
<td>1.90</td>
<td>0.05</td>
<td>0.05</td>
<td>0.48</td>
<td>0.47</td>
<td>0.06</td>
</tr>
<tr>
<td>2.00</td>
<td>0.03</td>
<td>0.01</td>
<td>0.67</td>
<td>0.47</td>
<td>0.03</td>
</tr>
</tbody>
</table>

CPP, colored periodontal probes; FPR, false positive rate; FOR, false omission rate; PPV, positive predictive value; SPP, standard periodontal probe; TPR, true positive rate. The values at a gingival thickness of 1 respectively 1.5 mm are indicated in bold.
Examples of two patient cases (A-C and D-F) where there was no agreement among the examiners for the colored periodontal probe (CPP) grading. Specifically, for the first case (A-C), two examiners judged it as medium [i.e., only the white CPP in (A) was judged as invisible, whereas the green and blue CPP were judged as invisible in (B-C)], but the remaining 6 examiners judged this case as very thick [i.e., none of CPP was judged as visible in (A-C)]; the actual gingival thickness of this canine measured by transgingival sounding was 1.42 mm. For the second case (D-F), all examiners agreed that the green and blue CPP are visible in (E-F), but four examiners judged also the white CPP in (D) as visible, whereas four examiners did not, resulting in an either thin or medium CPP grading; the actual gingival thickness of this central incisor measured by transgingival sounding was 1.49 mm.

- CPP grading thin/medium/thick vs. very thick: PPV 0.29 (CI: 0.10; 1.00), FOR 0.24 (CI: 0.17; 0.31), TPR 0.02 (CI: 0.00; 0.03), and FPR 0.01 (CI: 0.00; 0.03)

This is indicating that (1) 29% of the cases classified as very thick were correctly identified being > 1.5 mm (PPV), but (2) 24% of the cases classified as thin/medium/thick, were actually > 1.5 mm (FOR); further, (3) 2% of the cases being > 1.5 mm, were correctly classified as very thick (TPR), and (4) 1% of the cases being ≤ 1.5 mm, were wrongly classified as very thick (FPR).

### 3.3 Inter-examiner reproducibility and intra-examiner repeatability

The inter-examiner reproducibility and intra-examiner repeatability is presented in Supplementary Table 4 in the online Journal of Periodontology. The highest, although only moderate ($\kappa = 0.427$) agreement among all eight examiners was achieved by SPP; visual judgment has shown fair agreement ($\kappa = 0.211$) and CPP grading only a slight agreement ($\kappa = 0.112$). Figure 4 illustrates two patient cases for which the examiners did not agree well on CPP grading.

The intra-examiner repeatability showed high variability among the examiners and among the various judgment methods (see Supplementary Table 4 in online Journal of Periodontology). Specifically, a fair to moderate agreement was shown for the visual judgment, a slight to substantial agreement for the SPP, a poor to moderate for the white CPP, a fair to substantial for the green CPP, and a poor to almost perfect agreement for the blue CPP.

### 3.4 Verification of the method

Masking of the probe appeared not to influence the judgment of the examiners. Specifically, only two out of eight examiners judged a single picture out of the 50 pictures, in which the presence of a probe was “pretended/simulated,” as “probe is visible”; all other pictures (i.e., 99.5%) were correctly judged as “probe is invisible.”

### 4 DISCUSSION

Within the last decade, new devices promising a more reliable and/or detailed GT assessment/classification have been introduced into the market. In this context, it is important that the accuracy, inter-examiner reproducibility, and intra-examiner repeatability of any new device is properly assessed. The present study tested one of the newer devices—the CPP—consisting of a set of three probes allowing a classification into four GT groups (see Supplementary Figure 1 in online Journal of Periodontology). The rationale of having such a specific/detailed classification is that it may allow an improved detection rate of cases with very thick and very thin gingiva. However, in the present study most cases were categorized as “medium” with CPP, i.e., six out of eight examiners judged > 85% of the cases in this
category, whereas only few cases were classified to one of the other categories (i.e., “thin,” “thick,” and “very thick”). This clustering in the “medium” category corresponds well with what was reported in a recent study using CPP to judge GT in lower anterior teeth; therein, about 70% of the cases were classified by a single examiner as “medium.” Herein, this clustering was accompanied by a high false omission rate (FOR, i.e., 74% of the cases classified as thin with CPP, were actually > 1 mm) and a high false positive rate (FPR, i.e., 88% of the cases being ≤ 1 mm, were wrongly classified as thick). Although still high, the FPR was better when using SPP (FPR: 0.64) with similar values for positive predictive value (PPV), true positive rate (TPR), and FPR. Visual judgment provided an even lower/better FPR (0.39) but on costs of a lower TPR (i.e., only 49% of the cases being > 1 mm, were really classified as thick). Considering a lower cut-off value of 0.8 mm, as recently suggested, did not improve FOR (0.97) and FPR (0.81) for CPP, but decreased/improved FPR to 0.04 for SPP. Additionally, the green and blue CPP did not discriminate well between “thick” and “very thick” cases; in 20 out 39 times where a case was judged as “thick” or “very thick,” the proper probe sequence did not work (i.e., although the green CPP was judged visible, the blue CPP was not). This was also accompanied by a low PPV (i.e., only 39 and 29% of the cases classified as thick or very thick, respectively, were really > 1.5 mm) and a low TPR (i.e., only 5 and 2% of the cases being > 1.5 mm, were really classified as thick or very thick, respectively). This lack of discrimination between “thick” and “very thick” was also indicated by Kloukos et al., reporting an “overlapping” among “thick” and “very thick” cases.

Besides the accuracy of a method, intra-examiner repeatability and inter-examiner reproducibility are relevant parameters to judge its clinical applicability. Herein a high variability regarding the intra-examiner repeatability was observed for all methods. Specifically, for CPP the repeatability was poor to moderate for the white CPP, fair to substantial for the green CPP, and poor to almost perfect agreement for the blue CPP. The only previous study testing the CPP indicated also low repeatability; specifically, based on 20 re-judgments, only 11 cases were classified in the same category, whereas nine cases changed category by one step. Similarly, regarding inter-examiner reproducibility, only slight agreement among the eight examiners was observed for CPP method (κ = 0.112), whereas visual judgment and SPP showed fair and moderate agreement, respectively. These findings are somewhat in accordance with what was previously reported about the relatively low potential of the visual judgement and at least partly also of the SPP transparency method.

However, the studies on SPP transparency method may not be easily comparable, because various probe designs were used, i.e., probes differed in thickness or color with the first 3 mm might being either continuously silver or black or with black 1 mm markings.

The process to assess probe transparency in the present study is different compared to how the probes are used in the clinic. Specifically, the probe and the first 0.5 mm of the gingival margin were artificially masked on the photographs so that the examiners could not see which type of probe was used. This, in combination with the randomization process (i.e., not all five pictures of a given tooth were sequentially judged), prevented that the judgment was influenced by knowing how the previously used probes have been judged on the same tooth. Further, the adjacent papillae and teeth were cropped, and only the central aspect of the tooth was visible. This prevented that the judgment was influenced by any unintentional “pre”-judgment of the gingival phenotype because of seeing a larger aspect of the dentition. Altogether, these measures supported an unbiased evaluation of the CPP. Nevertheless, a few limitations should be considered when interpreting the results: (1) only upper anterior teeth were assessed herein, based on the rationale that they are aesthetically most relevant; nevertheless, similar results were presented in the previous study assessing CPP at lower anterior teeth; (2) a minimum amount of KTW (≥ 2 mm) was defined as inclusion criterion, i.e., the present results cannot be transferred to cases with a lack of keratinized tissue; (3) the sample size was based on previous studies with similar number of patients and not on any sample size calculation, however, the CI show in general a narrow range, and (4) in the lack of a real “gold standard” the transgingival sounding was accepted herein as reference method, as it is easy-to-use, minimal invasive, without radiation exposure, and, it has been previously referred to as a reliable method.

In perspective, there is no consensus on which GT measured at which vertical level should be considered as relevant/critical for clinical decision making. Specifically, the literature provides a range of cut-off values for a “thin” gingiva from 0.5 to 1.5 mm with 1 mm—although still rather arbitrary—being most commonly used. Further, GT is measured at various vertical levels, ranging from about 1 to 3 mm from to the gingival margin to a level related to the probing pocket depth. One of these previous studies reported an increasing GT from the gingival margin to the alveolar crest, whereas thereafter the gingiva covering the alveolar process presented with a reduced thickness again; the highest correlation between the SPP transparency and GT was described at a vertical level of 0.5 to 1 mm from the gingival margin, which corresponds to the vertical level chosen herein and is reachable in each patient by the tip of a probe. Nevertheless, knowing that GT is changing...
according to the vertical level, this makes any cut-off level for a thin or thick gingiva even more arbitrary. In this context, a digital technique allowing GT assessment at any chosen height could be advantageous. Very recently, superimposition of STL and DICOM files was shown to achieve excellent agreement with transgingival sounding with an endodontic spreader.20

5 | CONCLUSION

The results of the present study do not support the use of the CPP transparency method for classifying the gingival phenotype in terms of GT, neither for clinical decision making nor research purposes. CPP was not able to accurately discriminate among the various GT categories and showed low intra-examiner repeatability and inter-examiner reproducibility.

ACKNOWLEDGMENTS

The authors thank Dr. Lars-Åke Johansson, DDS, dr. odont., and Dr. Philip Arneryd (both, Department of Periodontology, Faculty of Odontology, University of Malmö, Sweden) for being judges of the photographs.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest. No external funding was obtained for performing the present study.

AUTHOR CONTRIBUTIONS

Kristina Bertl, Klaus Gotfredsen, and Andreas Stavropoulos made substantial contributions in conception and design of the work and in drafting and/or revising the work critically for important intellectual content; Mehdi Al-Hotheiry, David Sun, John Olofsson, and Stefan Lettner made substantial contributions in the acquisition and analysis of data for the work and in drafting the work; all authors approved the version to be published and agreed to be accountable for all aspects of the work.

ORCID

Kristina Bertl https://orcid.org/0000-0002-8279-7943
Andreas Stavropoulos https://orcid.org/0000-0001-8161-3754

REFERENCES


**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of the article.