



The staging and grading system in defining periodontitis cases: consistency and accuracy amongst periodontal experts, general dentists and undergraduate students

Lorenzo Marini¹ | Maurizio S. Tonetti^{2,3} | Luigi Nibali⁴ | Mariana A. Rojas¹ |
Mario Aimetti⁵ | Francesco Cairo⁶ | Raffaele Cavalcanti⁷ | Alessandro Crea⁸ |
Francesco Ferrarotti⁵ | Filippo Graziani⁹ | Luca Landi¹⁰ | Nicola M. Sforza¹¹ |
Cristiano Tomasi¹² | Andrea Pilloni¹

¹Section of Periodontics, Department of Oral and Maxillofacial Sciences, Sapienza University of Rome, Rome, Italy

²Division of Periodontology and Implant Dentistry, Faculty of Dentistry, the University of Hong Kong, Hong Kong, China

³Department of Oral and Maxillo-facial Implantology, Shanghai Key Laboratory of Stomatology, National Clinical Research Centre for Stomatology, Shanghai Ninth People Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

⁴Periodontology Unit, Centre for Host-Microbiome Interactions, Faculty of Dental and Craniofacial Sciences, King's College London, London, UK

⁵Department of Surgical Sciences, C.I.R. Dental School, Section of Periodontology, University of Turin, Turin, Italy

⁶Research Unit in Periodontology and Periodontal Medicine, Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy

⁷Section of Periodontology, Department of General Surgery and Medical-Surgical Specialties, School of Dentistry, University of Catania, Catania, Italy

⁸Private Practice, Viterbo, Italy

⁹Sub-Unit of Periodontology, Halitosis and Periodontal Medicine, Department of Surgical, Medical and Molecular Pathology and Critical Care Medicine, University of Pisa, Pisa, Italy

¹⁰Private Practice, Rome and Verona, Italy

¹¹Private Practice, Bologna, Italy

¹²Department of Periodontology, Institute of Odontology, The Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden

Correspondence

Lorenzo Marini, Section of Periodontics,
Department of Oral and Maxillofacial
Sciences, Sapienza University of Rome, 6
Caserta Street, Rome 00161, Italy.
Email: lorenzo.marini@uniroma1.it

Abstract

Aim: The objective of this study was to evaluate consistency and accuracy of the periodontitis staging and grading classification system.

Methods: Thirty participants (10 periodontal experts, 10 general dentists and 10 undergraduate students) and a gold-standard examiner were asked to classify 25 fully documented periodontitis cases twice. Fleiss kappa was used to estimate consistency across examiners. Intraclass correlation coefficient (ICC) was used to calculate consistency across time. Quadratic weighted kappa and percentage of complete agreement versus gold standard were computed to assess accuracy.

Results: Fleiss kappa for stage, extent and grade were 0.48, 0.37 and 0.45 respectively. The highest ICC was provided by students for stage (0.91), whereas the lowest ICC by general dentists for extent (0.79). Pairwise comparisons against gold standard showed mean value of kappa >0.81 for stage and >0.41 for grade and extent. Agreement with the gold standard for all three components of the case definition was

achieved in 47.2% of cases. The study identified specific factors associated with lower consistency and accuracy.

Conclusions: Diagnosis was highly consistent across time and moderately between examiners. Accuracy was almost perfect for stage and moderate for grade and extent. Additional efforts are required to improve training of general dentists.

KEYWORDS

classification, data accuracy, diagnosis, periodontitis, reproducibility of results

1 | INTRODUCTION

The 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions introduced a new periodontitis case definition system (Tonetti et al., 2018). It is based on three components: (a) diagnosis of an individual as a periodontitis case; (b) identification of the specific form of periodontitis (Albandar et al., 2018; Herrera et al., 2018); (c) case assignment through the novel process of staging and grading (Tonetti et al., 2018).

The case definition provides a uniform description of a periodontitis patient, overcoming the difficulties of the previous classification in differentiating between aggressive and chronic periodontitis (Armitage, 1999; Lang et al., 1999). Periodontitis case definition can be easily communicated to patients or other clinicians/researchers. Furthermore, it could be relevant in assessing prognosis and may enhance individual patient management (Sanz, Herrera, et al., 2020).

As for all new re-classification of disease modalities, introducing a new periodontitis case definition system in clinical practice and education requires a learning curve to understand and become acquainted with its novel nature. In order to facilitate this process, empiric decision-making algorithms to guide clinicians and trainees in the assignment of cases to the proper periodontal diagnosis were suggested (Tonetti & Sanz, 2019). Furthermore, additional guidelines in the identification of potential grey zones, practical tips to help clinicians and, more recently, clarifications on how to apply the extent criterion and how to calculate tooth loss due to periodontitis were provided (Kornaman & Papapanou, 2020; Sanz et al., 2020).

Since its introduction, the periodontitis case definition system progressively started to be applied in research and clinical practice. However, to the best of our knowledge, no studies have been published yet to evaluate the reliability and accuracy when defining periodontitis cases.

The objective of this study was to describe the consistency across time and across examiners in the definition of stage, extent and grade of periodontitis cases amongst periodontal experts, general dentists and undergraduate dental students. The study also compared the case definitions of examiners to a gold standard to verify their accuracy in the assignment of stage, extent and grade of periodontitis.

Clinical Relevance

Scientific rationale for study: To date no study evaluated the consistency and accuracy when staging and grading periodontitis cases.

Principal findings: Consistency across time was almost perfect, whilst across examiners was moderate. Accuracy for stage was high whereas it was moderate for extent and grade. In nearly half of the cases, a complete agreement was reached with the gold standard for all the three components of case definition.

Practical implications: Education and training are needed to improve consistency and accuracy. Empiric decision-making algorithms or dedicated software might help the professionals and the trainee in this purpose.

2 | MATERIALS AND METHODS

2.1 | Study design

The study was based on the examination of the baseline digital documentation and subsequent stage, extent and grade definition of 25 untreated periodontitis cases. All cases were evaluated by 30 examiners, equally subdivided in three groups according with their level of education and experience in periodontology. Each case was assessed twice by all the participants to calculate the consistency across time and across examiners. The assessments of each examiner were compared to those of a gold standard (MST) directly involved with the development of the staging and grading system in order to assess accuracy.

The study was conducted according to the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) (Kottner et al., 2011).

2.2 | Ethical Considerations

The baseline clinical and radiographic documentation of periodontitis cases were collected in the context of routine care in

the Periodontology clinic of the University of Rome from June to December 2019. Anonymized data were used in the study. All subjects had provided informed consent to the use of the collected data in the context of training and research. According with the U.S. Department of Health and Human Services (HHS) definition, this investigation is not considered human subjects research. The study protocol was approved by the Department of Oral and Maxillofacial Sciences of Sapienza, University of Rome (Prot. N. 0000598/2020). Prior to starting the study, all the examiners signed an informed consent.

2.3 | Examiners

The following 30 participants, equally divided in three groups according to their educational level and expertise in periodontology, were selected to contribute to this study:

- (i) Ten final year undergraduate dental students of Sapienza, University of Rome, School of Dentistry were randomly selected using a computer-generated sequence;
- (ii) Ten general dentists with >10 years of clinical experience, who did not attend advanced graduate education programs in periodontology and do not exclusively focus on any specific field of dentistry in their own practice.
- (iii) Ten periodontal experts selected amongst certified periodontists by the Italian Society of Periodontology.

Furthermore, one examiner (MST) – not included in the previously described groups of participants – was selected amongst the authors of the case definitions for periodontitis developed in the context of the 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions (Caton et al., 2018).

2.4 | Procedures

2.4.1 | Selection and preparation of the documentation of the periodontitis cases

From 50 available fully documented periodontitis cases collected in the context of routine care, 25 were selected to ensure high quality and diagnostic precision of clinical, photographical and radiographical records by two investigators (LM and MAR) not involved in the assessments. All cases selected for this study received the diagnosis of periodontitis according to the 2017 World Workshop definition (Tonetti et al., 2018). Necrotizing forms or systemic manifestations of periodontitis were excluded from the study.

For staging the periodontitis case, full-mouth radiographs, a periodontal chart and a periodontal history of tooth loss are needed. For grading the periodontitis case previous periodontal records or, when not available, the bone/age ratio of the most affected tooth

calculated on the full-mouth radiographs and information related to the presence of recognized risk factors such as smoking and diabetes are necessary (Tonetti & Sanz, 2019). Therefore, the baseline documentation of each case provided the following information:

- a. age and gender;
- b. anamnestic data presented in a standardized format and subdivided in two sections. Section one comprised the general medical history and included any relevant systemic diseases and pharmacological treatment, as well as cigarette consumption (0, ≤10/day or >10/day). In patients with diabetes, values of glycated haemoglobin (<7% or ≥7%) acquired from the patient's medical record were provided. Section two comprised the dental history and included dichotomously recorded information (yes or no) about: (1) gingival bleeding, (2) tooth mobility, (3) dentin hypersensitivity, (4) halitosis, (5) family history of periodontitis, (6) use of interdental oral hygiene devices, (7) use of mouthwashes, (8) para-functional habits, (9) chewing difficulties, (10) tooth migration, (11) previous orthodontic treatment, (12) previous periodontal treatment and (13) previous prosthetic treatment. Moreover, the last dental examination and professional oral hygiene procedure (≤1 year, >1 year or >3 year) and the number of tooth loss attributable to periodontitis (0, ≤4 or ≥5) were reported;
- c. nine intra-oral photographs displaying the buccal and palatal/lingual view of all sextants;
- d. full-mouth long-cone, parallel technique, periapical radiographs;
- e. a periodontal chart displaying: (1) probing depth (PD) recorded at six sites per tooth of the entire dentition; (2) clinical attachment level (CAL) recorded at six sites per tooth of the entire dentition; (3) bleeding on probing (BOP) recorded dichotomously at six sites per tooth of the entire dentition, (4) furcation involvement (FI) according to the Hamp classification (Hamp et al., 1975), (5) tooth mobility (M) according to the Miller index (Miller, 1950), (6) full-mouth plaque score (FMPS) (O'Leary et al., 1972) and (7) full-mouth bleeding score (FMBS). CAL was estimated as the sum of PD and gingival margin (GM) at each site. GM measurements were performed simultaneously with the PD measurements. GM was measured by recording the distance from the cemento-enamel junction (CEJ) to the margin of the gingiva at 6 sites on each tooth. In periodontal sites with the gingival margin located on the root and a visible CEJ, the GM was given a positive sign. In periodontal sites with no visible CEJ, the periodontal probe (PCP-UNC 15, Hu-Friedy, Chicago, IL, USA) was inserted into the periodontal pocket and angulated approximately 45° in order to manually detect the cervical line. The depth of insertion into the periodontal pocket was recorded as GM and the measurement received a negative sign.

Two slideshow presentation files containing the complete documentation of the periodontitis cases were assembled. In the two presentations, there were the same twenty-five cases, but they were randomly ordered. Furthermore, a data collection file was prepared. The first presentation is provided as Appendix S1.

2.4.2 | Training of participants

Before beginning the study, all participants received a copy of the study procedures and detailed instructions. Subsequently, the examiners were provided with three clinical cases, not included in the study, for explaining the case presentation and assessment modalities. When necessary, the examiners' doubts were clarified and the procedure was re-explained.

Each participant previously attended at least one course/seminar on how to apply the periodontitis case definition system. No additional training on the new classification was performed prior to the start of the study.

2.4.3 | Staging and grading of periodontitis cases

The three groups of participants blindly to each other and independently examined the first presentation containing the twenty-five periodontitis cases and defined stage, extent and grade of each case, according to the new classification scheme. Examiners did not have the support of any implementation tool except for the staging and grading tables for their convenience (Tonetti et al., 2018). After an interval of one week, the second presentation was examined by the three groups and all cases were again diagnosed. The examiners carried out the assessments from their own workstations and no time limits were given to the examiners to define cases. However, participants had to record the exact time necessary for staging and grading of each case.

The reference examiner examined all the periodontitis cases as well. Stage, extent and grade that he provided were chosen *a priori* and considered as the gold standard. After scoring all cases in each presentation, raters returned the data collection forms for statistical analysis.

2.5 | Outcomes

The primary outcome was the consistency of stage, extent and grade definitions across examiners. The secondary outcomes were: (a) the consistency of stage, extent and grade definitions across time; (b) the accuracy of the stage, extent and grade definitions; (c) the scoring time.

2.6 | Statistical analysis

The consistency of stage, extent and grade definitions across examiners, selected as primary outcome, was evaluated as an inter-examiner agreement between overall evaluators and between evaluators within each group. It was calculated based on the results of the examination of the periodontitis cases included in the first presentation using the Fleiss kappa statistics (Fleiss, 1981).

The consistency of stage, extent and grade definitions across time was estimated as intra-examiner agreement by evaluators of

each group between two separate evaluations 1 week apart. It was assessed using intraclass correlation coefficient (ICC).

The accuracy of the assessments was evaluated by comparing the stage, extent and grade definitions of the cases collected in the first presentation file provided by each evaluator with those of the gold standard. Quadratic weighted kappa was calculated for each pairwise comparisons. Percentage and frequencies of complete agreement for stage, extent and grade with gold standard were also calculated. A sub-analysis was performed based on the group of the examiners, the stage, the grade and the presence of modifying factors to study the variables that could affect accuracy. In the respect of the test assumptions (Bewick et al., 2004), chi-squared test was used to determine whether there was a statistically significant difference between the expected and the observed frequencies. The significance level of statistical tests was set at 0.05.

A six-level nomenclature was used to interpret the kappa and the ICC values: poor agreement = <0.00; slight agreement = 0.00 to 0.20; fair agreement = 0.21 to 0.40; moderate agreement = 0.41 to 0.60; substantial agreement = 0.61 to 0.80 and almost perfect agreement = 0.81 to 1.00 (Landis & Koch, 1977).

In the absence of previous data in the field, the expected values of kappa are inevitably chosen arbitrarily (Sim & Wright, 2005). The more common range of kappa values in medical reliability studies is between 0.4 and 0.6 (Koran, 1975). As noted by McHugh (2012), the lowest kappa value of 0.41 may be considered adequate, even though any kappa equal or greater than 0.61 should be preferred. For this study, it was considered reasonable to expect at least kappa values of 0.41 for the consistency of stage, extent and grade definitions across examiners and of 0.61 for at least 50% of the pairwise comparisons with the gold standard.

Mean and SD of time taken for overall case definitions (stage, extent and grade) according with the different groups of examiners, the stage and the grade assigned by the gold standard and the accuracy of the diagnosis were presented. Scoring time recorded during the examination of periodontitis cases collected in the first presentation file was considered for analysis. The normality of distribution of the considered variables was evaluated with Shapiro-Wilk test or Kolmogorov-Smirnov test. In absence of normally distributed variables, differences were compared with Kruskal-Wallis test. The significance level of statistical tests was set at 0.05.

The statistical analysis was carried out by two investigators (LN and LM) using a statistical software package (IBM Corp. Released 2017. IBM SPSS Statistics for Macintosh, Version 25.0. Armonk, NY: IBM Corp.)

2.7 | Sample size

In reliability studies, the number of subjects has a much greater impact on the precision than the number of raters does (Streiner & Norman, 2003). Therefore, it is recommended determining the number of raters based on generalizability and feasibility, then estimating the number of subjects required to achieve the desired precision (Karanicolas et al.,

TABLE 1 Intraclass correlation coefficient for different groups of examiners for stage, extent and grade

Examiners	Stage	Extent	Grade
Periodontal experts (<i>n</i> = 10)	0.818	0.882	0.871
General dentists (<i>n</i> = 10)	0.916	0.792	0.860
Undergraduate Students (<i>n</i> = 10)	0.949	0.985	0.879

2009). For this investigation, the convenience number of the examiners for each of the 3 groups was established to be 10, based on previous comparable studies (Cairo et al., 2010; Isaia et al., 2018; Rotundo et al., 2015). Then, using pairwise comparisons with a required kappa of 0.61, lower end of the 95% confidence interval (CI) for kappa as 0.28 and expected agreement 50% of the time, the required sample size was estimated to be 25 cases (Donner & Rotondi, 2010).

3 | RESULTS

3.1 | Descriptive characteristics of periodontitis cases

Twenty-five periodontitis cases were examined in the present study. The sample consisted of 14 (56%) females and 11 (44%) males, aged 29 to 74 years with mean age 47.6 ± 13.3 years. No smoking habit, cigarette consumption of <10/day and cigarette consumption of ≥ 10 /day were observed in 17 (68%), 4 (8%) and 4 (8%) of cases respectively. The periodontitis cases were normoglycemic/no diabetes diagnosis, diabetes diagnosis with HbA1c <7% and diabetes diagnosis with HbA1c $\geq 7\%$ in 22 (88%), 2 (8%) and 1 (4%) of cases respectively.

According to the diagnoses made by the gold-standard examiner, the distribution of periodontitis cases by stage, extent and grade

was: 2 cases were defined as stage I (8%), 4 as II (16%), 12 as III (48%) and 7 as IV (28%); 20 were assessed as generalized (80%) and 5 as localized (20%); and 10 were assigned to grade B (40%) and 15 to grade C (60%).

3.2 | Consistency of stage, extent and grade definitions across time

The intraclass correlation coefficients (ICC) for stage, extent and grade definitions of examiners of each group are presented in Table 1. Generally, consistency across time was almost perfect (ICC = 0.81 – 1.00) and higher amongst undergraduate students.

3.3 | Consistency of stage, extent and grade definitions across examiners

Table 2 shows results of Fleiss kappa between periodontal experts, general dentists, undergraduate students and overall 30 examiners. Mostly, consistency across examiners was moderate (Fleiss Kappa = 0.41–0.60).

When testing in pairs, periodontal experts and students had the highest consistency for staging (Fleiss kappa = 0.60), whilst values for grading and extent appeared similar between groups (Table 2).

3.4 | Accuracy of stage, extent and grade definitions compared to the gold standard

Individual stage, extent and grade of the 25 periodontitis cases defined by the gold-standard examiner and the 30 raters are summarized in Figure 1.

TABLE 2 Fleiss kappa statistics (95% confidence interval) for different groups of examiners, for pairs of comparisons and for overall examiners for stage, extent and grade

Examiners	Stage	Extent	Grade
Groups			
Periodontal Experts (<i>n</i> = 10)	0.58 (0.53–0.61)	0.36 (0.30–0.42)	0.42 (0.38–0.46)
General dentists (<i>n</i> = 10)	0.36 (0.32–0.40)	0.31 (0.25–0.36)	0.44 (0.39–0.48)
Undergraduate students (<i>n</i> = 10)	0.65 (0.61–0.68)	0.64 (0.58–0.69)	0.52 (0.47–0.57)
Pairs of comparisons			
Periodontal experts – General dentist (<i>n</i> = 20)	0.44 (0.41–0.45)	0.35 (0.31–0.37)	0.43 (0.41–0.45)
Periodontal experts – undergraduate students (<i>n</i> = 20)	0.60 (0.57–0.61)	0.42 (0.39–0.45)	0.46 (0.35–0.48)
General dentists – undergraduate students (<i>n</i> = 20)	0.45 (0.43–0.47)	0.38 (0.35–0.41)	0.46 (0.43–0.48)
Overall (<i>n</i> = 30)	0.48 (0.47–0.49)	0.37 (0.35–0.39)	0.45 (0.43–0.46)

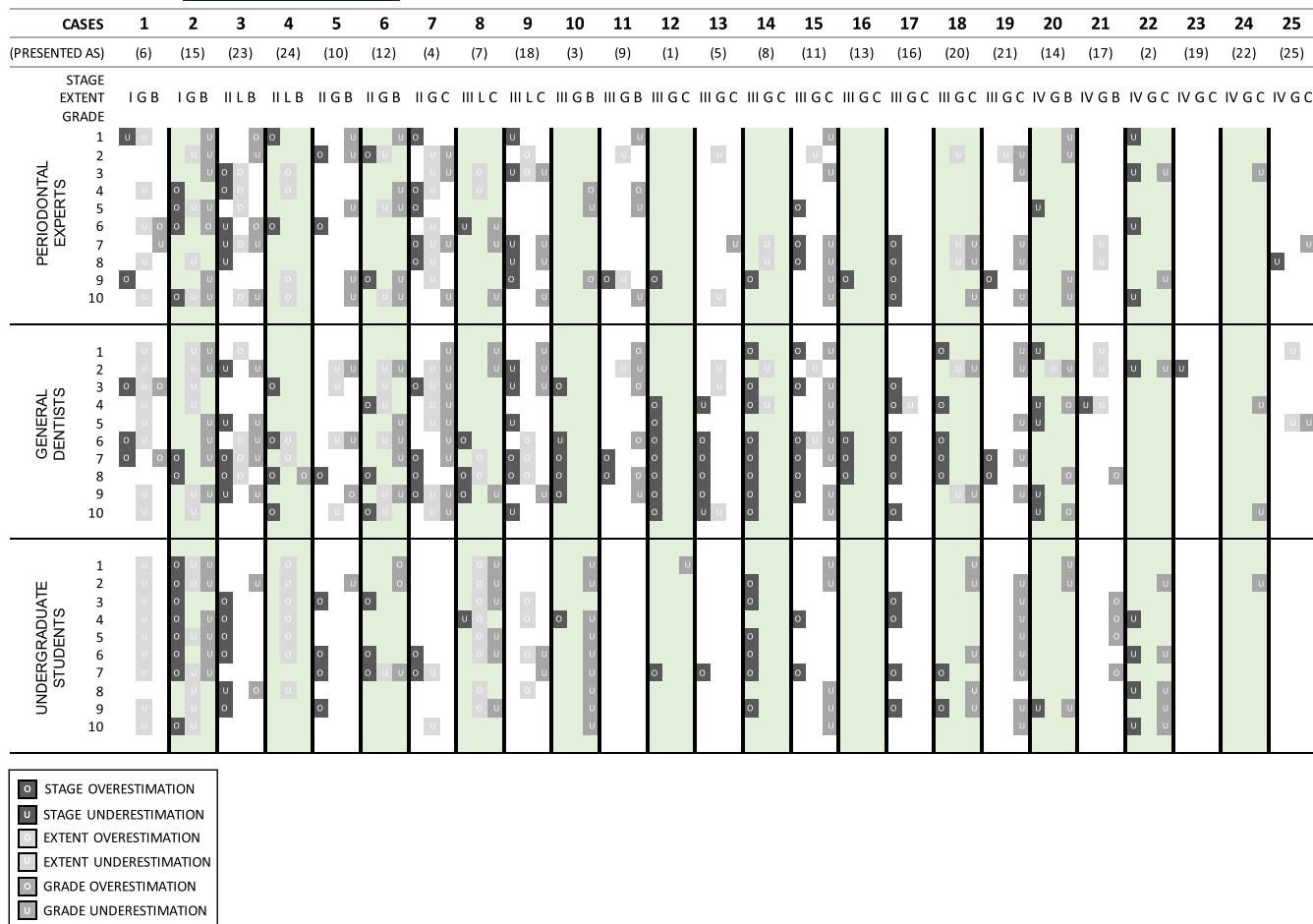


FIGURE 1 Individual stage, extent and grade of the twenty-five periodontitis cases defined by the gold-standard examiner and comparison against periodontal experts, general dentists and undergraduate students. Cases are progressively numbered according to the increasing severity of the disease. The number assigned to each case within the first presentation file is also provided

Agreement with the gold-standard examiner, who was assumed to provide the true definitions of stage, extent and grade is presented in Table 3. The quadratic weighted kappa values were higher for stage (almost perfect agreement) than for extent and grade (moderate agreement).

Frequencies and percentage of complete agreement with the gold-standard examiner are presented in Table 4. Consistency with the gold standard of general dentists was significantly lower than that of the other two groups for the overall diagnosis ($p < .001$) and, more in detail, for stage III ($p < .001$), extent ($p < .001$) and grade B ($p < .001$). Amongst all examiners, the more severe the stage and grade the greater the possibility to get the true diagnosis ($p < .001$ for both stage and grade).

A high percentage of complete agreement with the gold standard was reached for the discrimination between stage I and II vs III and IV, whilst a progressively lower percentage of agreement was achieved for the distinction between stage II vs III, I vs II and III vs IV (Figure 2).

Presence of modifying factors such as smoking and diabetes influenced agreement with the gold standard for grade. In particular,

the more severe the modifier, the higher the chance of obtaining agreement with the gold standard ($p < .001$) (Figure 2).

3.5 | Scoring time

The mean and SD of the time taken to evaluate all cases collected in the first presentation file by the three different groups of examiners are presented in Table 5. Periodontal experts were the fastest, followed by undergraduate students and finally by general dentists. The difference was statistically significant between the three groups ($p < .001$).

Table 5 shows minutes taken by all examiners for the overall diagnosis (definition of stage, extent and grade) according to the stage or the grade of the periodontitis cases (as assigned by the gold-standard examiner) and according to the accuracy of the complete diagnosis. Time for case definition was significantly shorter for cases that had a higher stage ($p < .001$) or grade ($p = .003$). Finally, cases properly diagnosed by examiners were evaluated in less amount of time compared to those that were misdiagnosed ($p < .001$).

TABLE 3 Frequency and percentage of agreements achieved by pairwise comparisons against gold-standard examiner

	Periodontal experts n (%)	General dentists n (%)	Undergraduate students n (%)	All examiners n (%)
Stage				
Slight (K = 0.01–0.2)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Fair (K = 0.21–0.4)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Moderate (K = 0.41–0.6)	0 (0.0%)	2 (20.0%)	0 (0.0%)	2 (6.6%)
Substantial (K = 0.61–0.8)	1 (10.0%)	4 (40.0%)	4 (40.0%)	9 (30.0%)
Almost perfect (K = 0.81–1.0)	9 (90.0%)	4 (40.0%)	6 (60.0%)	19 (63.3%)
Extent				
Slight (K = 0.01–0.2)	0 (0.0%)	2 (20.0%)	0 (0.0%)	2 (6.6%)
Fair (K = 0.21–0.4)	4 (40.0%)	3 (30.0%)	4 (40.0%)	11 (36.6%)
Moderate (K = 0.41–0.6)	3 (30.0%)	4 (40.0%)	5 (50.0%)	12 (40.0%)
Substantial (K = 0.61–0.8)	2 (20.0%)	1 (10.0%)	1 (10.0%)	4 (13.3%)
Almost perfect (K = 0.81–1.0)	1 (10.0%)	0 (0.0%)	0 (0.0%)	1 (3.3%)
Grade				
Slight (K = 0.01–0.2)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Fair (K = 0.21–0.4)	0 (0.0%)	2 (20.0%)	1 (10.0%)	3 (10.0%)
Moderate (K = 0.41–0.6)	5 (50.0%)	6 (60.0%)	4 (40.0%)	15 (50.0%)
Substantial (K = 0.61–0.8)	5 (50.0%)	2 (20.0%)	5 (50.0%)	12 (40.0%)
Almost perfect (K = 0.81–1.0)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

Note.: K, quadratic weighted kappa.

TABLE 4 Frequencies and percentage of stage, extent and grade definitions of periodontal experts, general dentists and undergraduate students consistent with those of the gold-standard examiner

Variable	Frequencies and % of complete agreement with gold-standard examiner				p value between examiners ^a	All examiners
	Periodontal experts	General dentists	Undergraduate students			
Stage (I–IV)	205 (82.0%)	161 (64.4%)	204 (81.6%)		<.001 [*]	570 (76.0%)
Stage^b						
I	15 (75.0%)	15 (75.0%)	12 (60.0%)		.489	42 (70.0%)
II	34 (68.0%)	33 (66.0%)	35 (70.0%)		.912	102 (68.0%)
III	101 (84.1%)	61 (50.8%)	102 (85.0%)		<.001 [*]	264 (73.6%)
IV	55 (90.0%)	52 (86.0%)	55 (91.0%)		.662	162 (89.4%)
p value between stages ^a	0.017 [*]	<0.001 [*]	0.001 [*]			<0.001 [*]
Extent	210 (84.0%)	191 (76.4%)	219 (87.6%)		.003 [*]	620 (82.6%)
Grade (A–C)	181 (72.4%)	169 (67.6%)	186 (74.4%)		.223	536 (71.4%)
Grade^b						
A	-	-	-		-	-
B	72 (60.0%)	63 (52.5%)	87 (72.5%)		.006 [*]	222 (61.7%)
C	109 (83.8%)	106 (81.5%)	99 (76.2%)		.275	314 (80.5%)
p value between grades ^a	<0.001 [*]	<0.001 [*]	0.563			<0.001 [*]
Overall diagnosis	126 (50.4%)	94 (37.6%)	134 (53.6%)		<.001 [*]	354 (47.2%)

^aChi-square test.

^bAs assigned by the gold-standard examiner.

*Statistically significant.

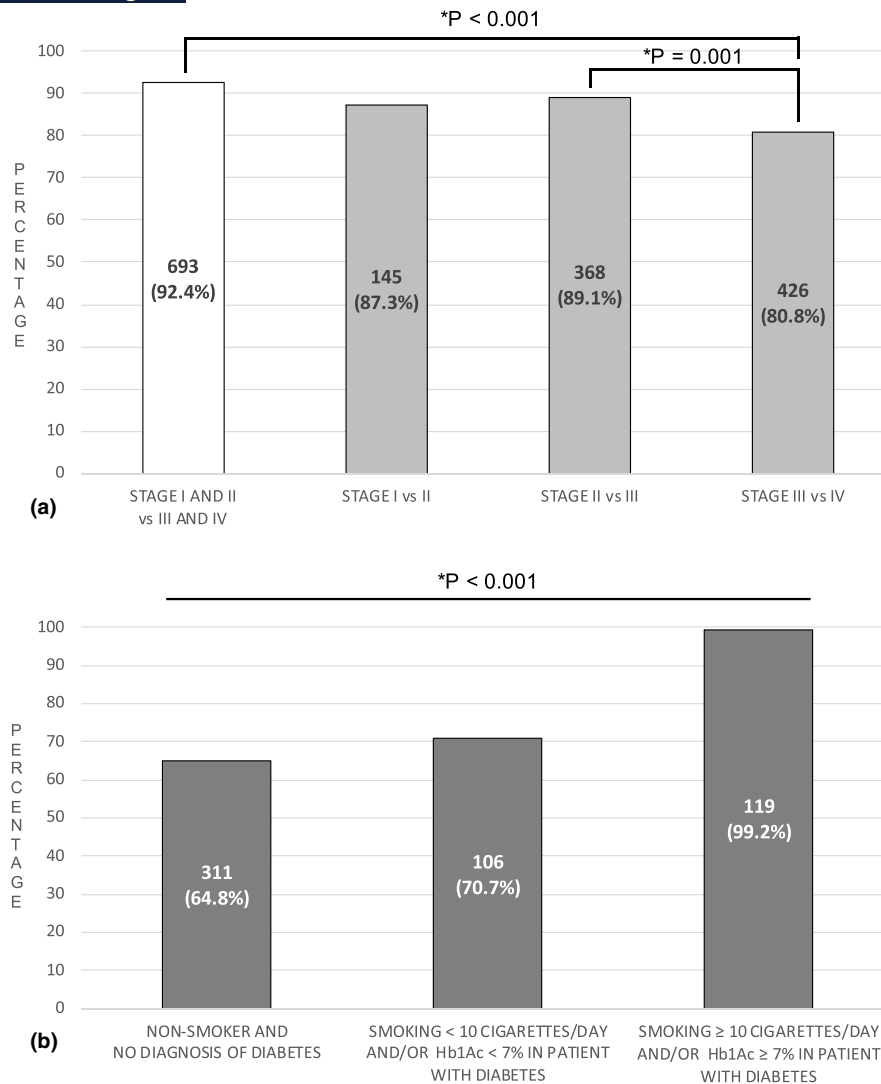


FIGURE 2 (a) Frequencies and percentage of complete agreement with the gold-standard examiner for stage distinction between I and II vs III and IV, I vs II, II vs III and III vs IV. (b) Frequencies and percentage of complete agreement with the gold-standard examiner for grade according to the presence of grade modifiers. *, statistically significant using chi-square test; HbA1c, haemoglobin A1c values

4 | DISCUSSION

The results of this study are noteworthy as they indicate that: (a) general dentists performed, in general, less well than either periodontists or senior dental students; (b) clinicians performed better in the staging component of the case definition than in the newly introduced grading or extent portion; (c) less consistent and accurate diagnoses were made for borderline cases; and (d) the bone loss by age component of grading was associated with less consistency and accuracy. Taken as a whole, these findings seem to indicate that the introduction of the new classification system requires significant additional training and specific clarifications aimed at aspects characterized by lower accuracy and consistency. The good performance of dental students indicates that training is possible. Training and implementation seem to be critical as imprecision and misclassification might limit the health gains that can be obtained from a new classification (Hefti & Preshaw, 2012).

In this study, consistency of the definitions of stage, extent and grade of 25 periodontitis cases across time was almost perfect, whilst across examiners was moderate. This observation may question the underlying knowledge of the raters. Accuracy of stage assessments was high and greater than that of extent and grade, which were moderate. In nearly half of the cases, a complete agreement was reached with the gold standard for all three components of the case definition.

This study offers the opportunity to assess performance of users with different level of knowledge and most likely exposure to training of the new classification system. The excellent performance of dental students shows what can be achieved with incorporation of the system into the undergraduate curriculum. Room for improvement of dental practitioners is evident and additional training seems necessary. Critical aspects for such training seem to be both extent and grade.

This analysis showed that clinicians are better at correctly discriminating more advanced stages of periodontitis (better

TABLE 5 Mean and SD of time taken for overall case definition (stage, extent and grade) according with the different groups of examiners, the stage and the grade assigned by the gold standard and the accuracy of the diagnosis.

Variable	Minutes, seconds (Mean ±SD)	p Value ^a
Examiners		
Periodontal experts (n = 10)	1:07 ± 0:43	<.001*
General dentists (n = 10)	2:04 ± 1:04	
Undergraduate students (n = 10)	1:51 ± 1:11	
Stage^b		
I (n = 2)	1:52 ± 1:02	<.001*
II (n = 5)	1:54 ± 1:03	
III (n = 12)	1:42 ± 1:05	
IV (n = 6)	1:24 ± 1:04	
Grade^b		
A (n = 0)	-	
B (n = 11)	1:44 ± 0:59	.003*
C (n = 14)	1:38 ± 1:09	
Complete diagnosis^b		
Accurate	1:31 ± 1:46	<.001*
Inaccurate	1:50 ± 1:42	

Abbreviations: SD, standard deviation.

^aKruskal–Wallis test

^bAs assigned by the gold-standard examiner

*Statistically significant.

accuracy for stage III and IV compared to stage I and II) but have difficulties in discriminating between stage III and IV. The clinical implications of this difficulty seem particularly important as it may affect communication with the patient of the complexity to manage their case.

Moderate or better agreement (0.41 based on Fleiss kappa) for stage, extent and grade, was consistently obtained only by dental students, whereas for stage and grade by periodontal experts and only for stage by general dentists. Extent obtained the lowest value of agreement amongst all examiners (Fleiss kappa = 0.37), probably because overall periodontitis sites distribution rather than percentage of teeth with the assigned stage was evaluated. It should be noted that the recently published clarification on how to apply the extent was not yet available to the examiners at the time of the assessments (Sanz, Herrera, et al., 2020). The reason why better consistency was achieved amongst students could be explained because they were recruited from the same institution and received uniform training.

In order to assess accuracy, each examiner's case definitions were compared with those provided by the gold-standard examiner. Given the importance of providing accurate diagnoses, one expected to obtain quadratic weighted kappa ≥ 0.61 for at least 50% of the

pairwise comparisons with gold standard for stage, extent and grade separately. However, it was only achieved by all examiners for stage and by periodontal experts and students for grade. With regards to the relatively low percentage of complete agreement for all three components of the case definition, it was not a surprising finding. Firstly, this may have been due to the fact that the new classification is rather 'young' and, secondly, it may have been due to the large number of cases that had to be assessed in a session.

Different case definitions can have a great impact on the prevalence and the extent rates of periodontitis. In this manner, the discrepancies may influence the results and the associations presented in studies as well as over or underestimating the real need for periodontal treatment (Costa et al., 2009). Although over or the underestimation of stage as well as of extent and grade can lead to different results, to date there is no data that suggests which of the two misalignments is worse.

In this study, none of the cases was classified as Grade A. However, this result offers an opportunity to remember how clinicians should initially assume the disease as Grade B and seek specific evidence to progress to Grade A. If in doubt, especially in the absence of direct evidence of lack of progression, clinicians should be discouraged from using Grade A at initial diagnosis.

Periodontal experts reached a diagnosis significantly faster than other groups, indicating that experience in periodontology may influence the speed in defining each periodontitis case. Although the scoring time generally seemed to be too short, the more a case showed obvious characteristics of a specific stage (in particular of stage IV) and grade (C), the less time was necessary for an exact diagnosis.

This study has several strengths. Mainly, this paper reports the first assessment of the consistency and accuracy of diagnosis that can be achieved with the new classification system. Cases were assembled in two presentation files in a randomized order after a one-week interval, to limit the effects of bias on the second examination. Documentation was shown in a uniform format that was easy to be examined. The pre-study training phase further ensured understanding of assessment methods. No time limit has been imposed for the evaluation. Data collection was simple and examiners were blinded by the case definitions of other participants. The number of examined cases was reasonably large and allowed to test the consistency and the accuracy through a wide range of manifestation of periodontitis and to perform a sub-analysis according with the case characteristics. However, further studies could require increased number of examiners.

The major limitation of this study was that all the information needed to define stage, grade and extent was assumed to be accurate and was not directly collected by each examiner. For these reasons, the effects of the individual skills in the periodontal anamnestic, clinical and radiographic examination, as well as the data selection, on the subsequent consistency in the case definition could not be estimated. However, the objective of this study was not to evaluate the diagnostic process as a whole, but rather

to assess the consistency and accuracy in defining a periodontitis case when all data are available and presumed to be correct. Another limitation was represented by the digital photographs in place of clinical inspection, even though this approach has been commonly validated in similar studies in various fields, including evaluation of aesthetic outcomes of periodontal plastic surgery (Cairo et al., 2010). Finally, the gold-standard examiner was arbitrarily designated. However, he was supposed to provide the most precise case definition as one of the authors of the newly developed staging and grading system.

5 | CONCLUSIONS

Education, practical skills and calibration might further increase both consistency and accuracy, in particular when an early periodontitis case or a borderline case in a non-smoker and/or non-diabetic patient is defined by general dentists. Further studies evaluating the ability of existing empiric decision-making tools or dedicated software to improve diagnostic skills are encouraged.

ACKNOWLEDGEMENTS

The authors are grateful to the general dentists and undergraduate dental students of the Sapienza University of Rome who participated in this study for their contribution in the role of examiners.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interests.

AUTHOR'S CONTRIBUTION

Lorenzo Marini: Conceptualization; data curation; formal analysis; investigation; methodology; project administration; resources; visualization; writing – original draft preparation; writing – review and editing. **Maurizio S. Tonetti:** Conceptualization; investigation; methodology; writing – original draft preparation; writing – review and editing. **Luigi Nibali:** Formal analysis; methodology; writing – review and editing. **Mariana A. Rojas:** Conceptualization; resources; writing – review and editing. **Mario Aimetti:** Investigation; writing – review and editing. **Francesco Cairo:** Investigation; writing – review and editing. **Raffaele Cavalcanti:** Investigation; writing – review and editing. **Alessandro Crea:** Investigation; writing – review and editing. **Francesco Ferrarotti:** Investigation; writing – review and editing. **Filippo Graziani:** Investigation; writing – review and editing. **Luca Landi:** Investigation; writing – review and editing. **Nicola M. Sforza:** Conceptualization; investigation; writing – review and editing. **Cristiano Tomasi:** Investigation; writing – review and editing. **Andrea Pilloni:** Conceptualization; investigation; methodology; supervision; writing – original draft preparation; writing – review and editing.



FUNDING INFORMATION

The study was self-funded by the authors and their institution.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Lorenzo Marini  <https://orcid.org/0000-0001-6662-2644>
 Maurizio S. Tonetti  <https://orcid.org/0000-0002-2743-0137>
 Luigi Nibali  <https://orcid.org/0000-0002-7750-5010>
 Mariana A. Rojas  <https://orcid.org/0000-0002-8712-0958>
 Mario Aimetti  <https://orcid.org/0000-0003-0657-0787>
 Francesco Cairo  <https://orcid.org/0000-0003-3781-1715>
 Filippo Graziani  <https://orcid.org/0000-0001-8780-7306>
 Cristiano Tomasi  <https://orcid.org/0000-0002-3610-6574>
 Andrea Pilloni  <https://orcid.org/0000-0002-6268-1863>

REFERENCES

- Albandar, J. M., Susin, C., & Hughes, F. J. (2018). Manifestations of systemic diseases and conditions that affect the periodontal attachment apparatus: Case definitions and diagnostic considerations. *Journal of Clinical Periodontology*, 45(Suppl 20), S171–S189. <https://doi.org/10.1111/jcpe.12947>
- Armitage, G. C. (1999). Development of a classification system for periodontal diseases and conditions. *Annals of Periodontology*, 4, 1–6. <https://doi.org/10.1902/annals.1999.4.1.1>
- Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 12: survival analysis. *Critical Care*, 8, 389–394. <https://doi.org/10.1186/cc2955>
- Cairo, F., Nieri, M., Cattabriga, M., Cortellini, P., De Paoli, S., De Sanctis, M., Fonzar, A., Francetti, L., Merli, M., Rasperini, G., Silvestri, M., Trombelli, L., Zucchelli, G., & Pini-Prato, G. P. (2010). Root coverage esthetic score after treatment of gingival recession: an interrater agreement multicenter study. *Journal of Periodontology*, 81, 1752–1758. <https://doi.org/10.1902/jop.2010.100278>
- Caton, J. G., Armitage, G., Berglundh, T., Chapple, I. L. C., Jepsen, S., Kornman, K. S., Mealey, B. L., Papapanou, P. N., Sanz, M., & Tonetti, M. S. (2018). A new classification scheme for periodontal and peri-implant diseases and conditions – Introduction and key changes from the 1999 classification. *Journal of Clinical Periodontology*, 45(Suppl 20), S1–S8. <https://doi.org/10.1111/jcpe.12935>
- Costa, F. O., Guimarães, A. N., Cota, L. O., Pataro, A. L., Segundo, T. K., Cortelli, S. C., & Costa, J. E. (2009). Impact of different periodontitis case definitions on periodontal research. *Journal of Oral Science*, 51, 199–206. <https://doi.org/10.2334/josnusd.51.199>
- Donner, A., & Rotondi, M. A. (2010). Sample size requirements for interval estimation of the kappa statistic for interobserver agreement studies with a binary outcome and multiple raters. *International Journal of Biostatistics*, 6(1), <https://doi.org/10.2202/1557-4679.1275>
- Fleiss, J. L. (1981). *Statistical methods for rates and proportions*, 2nd ed. (pp. 211–236). John Wiley & Sons.
- Hamp, S. E., Nyman, S., & Lindhe, J. (1975). Periodontal treatment of multirooted teeth. Results after 5 years. *Journal of Clinical Periodontology*, 2, 126–135. <https://doi.org/10.1111/j.1600-051x.1975.tb01734.x>
- Hefti, A. F., & Preshaw, P. M. (2012). Examiner alignment and assessment in clinical periodontal research. *Periodontology*, 2000(59), 41–60. <https://doi.org/10.1111/j.1600-0757.2011.00436.x>
- Herrera, D., Retamal-Valdes, B., Alonso, B., & Feres, M. (2018). Acute periodontal lesions (periodontal abscesses and necrotizing periodontal diseases) and endo-periodontal lesions. *Journal of Clinical Periodontology*, 45(Suppl 20), S78–S94. <https://doi.org/10.1111/jcpe.12941>
- Isaia, F., Gyurko, R., Roomian, T. C., & Hawley, C. E. (2018). The root coverage esthetic score: Intra-examiner reliability among dental

- students and dental faculty. *Journal of Periodontology*, 89, 833–839. <https://doi.org/10.1002/JPER.17-0556>
- Karanicolas, P. J., Bhandari, M., Kreder, H., Moroni, A., Richardson, M., Walter, S. D., Norman, G. R., & Guyatt, G. H., on Behalf of the Collaboration for Outcome Assessment in Surgical Trials (COAST) Musculoskeletal Group (2009). Evaluating agreement: Conducting a reliability study. *Journal of Bone and Joint Surgery*, 91, 99–106. <https://doi.org/10.2106/JBJS.H.01624>
- Koran, L. M. (1975). The reliability of clinical methods, data, and judgments (first of two parts). *The New England Journal of Medicine*, 293, 642–646. <https://doi.org/10.1056/NEJM197509252931307>
- Kornaman, K. S., & Papapanou, P. N. (2020). Clinical application of the new classification of periodontal diseases: Ground rules, clarifications and "gray zones". *Journal of Periodontology*, 91, 352–360. <https://doi.org/10.1002/JPER.19-0557>
- Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Robersts, C., Shoukri, M., & Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, 64(1), 96–106. <https://doi.org/10.1016/j.jclinepi.2010.03.002>
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Lang, N., Bartold, P. M., Cullinan, M., Jeffcoat, M., Mombelli, A., Murakami, S., Page, R., Papapanou, P., Tonetti, M., & Dyke, T. V. (1999). Consensus report: aggressive periodontitis. *Annales of Periodontology*, 4, 53. <https://doi.org/10.1902/annals.1999.4.1.53>
- McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22, 276–282. <https://doi.org/10.11613/BM.2012.031>
- Miller, S. C. (1950). *Textbook of Periodontia*, 3rd ed. Blakiston Co.
- O'Leary, T. J., Drake, R. B., & Naylor, J. E. (1972). The plaque control record. *Journal of Periodontology*, 43, 38. <https://doi.org/10.1902/jop.1972.43.1.38>
- Rotundo, R., Nieri, M., Bonaccini, D., Mori, M., Lamberti, E., Massironi, D., Giachetti, L., Franchi, L., Venezia, P., Cavalcanti, R., Bondi, E., Farneti, M., Pinchi, V., & Buti, J. (2015). The Smile Esthetic Index (SEI): A method to measure the esthetics of the smile. An intra-rater and inter-rater agreement study. *European Journal of Oral Implantology*, 8, 397–403.
- Sanz, M., Herrera, D., Kebschull, M., Chapple, I., Jepsen, S., Beglundh, T., Sculean, A., Tonetti, M. S., Merete Aass, A., Aimetti, M., Kuru, B. E., Belibasakis, G., Blanco, J., Bol-van den Hil, E., Bostanci, N., Bozic, D., Bouchard, P., Buduneli, N., Cairo, F., ... Wennström, J. (2020). Treatment of stage I–III periodontitis -The EFP S3 level clinical practice guideline. *Journal of Clinical Periodontology*, 47, 4–60. <https://doi.org/10.1111/jcpe.13290>
- Sanz, M., Papapanou, P. N., Tonetti, M. S., Greenwell, H., & Kornman, K. S. (2020). Guest Editorial: Clarifications on the use of the new classification of periodontitis. *Journal of Clinical Periodontology*, 47, 658–659. <https://doi.org/10.1111/jcpe.13286>
- Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85, 257–268. <https://doi.org/10.1093/ptj/85.3.257>
- Streiner, D. L., & Norman, G. R. (2003). *Health measurement scales: a practical guide to their development and use*, 3rd ed. Oxford University Press.
- Tonetti, M. S., Greenwell, H., & Kornman, K. S. (2018). Staging and grading of periodontitis: Framework and proposal of a new classification and case definition. *Journal of Clinical Periodontology*, 45(Suppl 20), S149–S161. <https://doi.org/10.1111/jcpe.12945>
- Tonetti, M. S., & Sanz, M. (2019). Implementation of the new classification of periodontal diseases: Decision-making algorithms for clinical practice and education. *Journal of Clinical Periodontology*, 46, 398–405. <https://doi.org/10.1111/jcpe.13104>

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Marini L, Tonetti MS, Nibali L, et al. The staging and grading system in defining periodontitis cases: consistency and accuracy amongst periodontal experts, general dentists and undergraduate students. *J Clin Periodontol*. 2020;00:1–11. <https://doi.org/10.1111/jcpe.13406>