Received: 2 October 2019 | Revised: 19 December 2019 | Accepted: 1 January 2020

DOI: 10.1111/jop.12990

ORIGINAL ARTICLE

Journal of Oral Pathology & Medicine WILEY

Grading of oral squamous cell carcinomas – Intra and interrater agreeability: Simpler is better?

Sonja E. Steigen[1,2] | Tine M. Søland[3,4] | Elisabeth Sivy Nginamau[5] | Helene Laurvik[4] | Daniela-Elena Costea[5,6] | Anne Christine Johannessen[5,6] | Peter Jebsen[4] | Inger-Heidi Bjerkli[1,7] | Lars Uhlin-Hansen[1,2] | Elin Hadler-Olsen[1,2,8]

[1]Department of Medical Biology, Faculty of Health Sciences, UiT – The Arctic University of Norway, Tromsø, Norway

[2]Department of Clinical Pathology, University Hospital of North Norway, Tromsø, Norway

[3]Department of Oral Biology, Faculty of Dentistry, University of Oslo, Oslo, Norway

[4]Department of Pathology, Rikshospitalet, Oslo University Hospital, Oslo, Norway

[5]Department of Pathology, Haukeland University Hospital, Bergen, Norway

[6]The Gade Laboratory of Pathology and Center for Cancer Biomarkers CCBIO, Department of Clinical Medicine, Faculty of Medicine, University of Bergen, Bergen, Norway

[7]Department of Otorhinolaryngology, University Hospital of North Norway, Tromsø, Norway

[8]Department of Clinical Dentistry, Faculty of Health Sciences, University of Tromsø, The Arctic University of Norway, Tromsø, Norway

Correspondence
Sonja E. Steigen, Department of Medical Biology, Faculty of Health Sciences, UiT – The Arctic University of Norway, Tromsø, Norway.
Email: sonja.eriksson.steigen@uit.no

Funding information
UiT The Arctic University of Norway

Abstract

Background: Numerous studies have been presented on histological grading of oral squamous cell carcinomas (OSCC) for predicting survival, but uncertainty of their usefulness rises due to discordances of results. A scoring system should be robust and well validated, and intra- and interrater agreement can be used as a tool to visualize the strength of reproducibility.

Methods: Here, we present an intra- and inter-observer study on evaluation of OSCC using some of the most common histopathological parameters. The observers were from different Norwegian university hospitals, and calibration to ensure accuracy was first performed. Percentage of the agreement was calculated for the score made by the individual observer at different times, as well as between pairs of observers.

Results: The evaluation made by the same observer at two different time points (intrarater) correlated better than observations made by different participants (interrater). In an attempt to increase the rate of agreement, many of the parameters were either dichotomized into simply low- and high grade, or to a three-tier system when more than three options in the original design. This increased the concurrence with 15.4% for the intrarater and with 23% for the interrater comparisons.

Conclusion: High agreement for histopathological parameters can be difficult to obtain on hematoxylin and eosin staining in scoring systems with many options. A simpler system might be more advantageous to achieve higher degree of reproducibility.

KEYWORDS

histopathological parameters, intrarater and interrater agreement, oral cancer, squamous cell carcinoma

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2020 The Authors. *Journal of Oral Pathology & Medicine* published by John Wiley & Sons Ltd

1 | INTRODUCTION

Oral cavity cancer originates almost exclusively from squamous cells (SC), and the histopathological evaluation of these tumors is the basis for their classification and further treatment.

The prediction of outcome and the selection of treatment for patients with oral squamous cell carcinomas (OSCC) are today based on the clinical tumor, nodes, and metastasis (TNM) staging. Tumor thickness, as measured during microscopic evaluation, was recently implemented in the T (size) variable.[1] Further, according to the WHO classification of head and neck (HN) tumors, the tumor differentiation should also be reported in order to predict prognosis.[2] The histological grading does not take into account the tumor-host interactions that modulate tumor progression and aggressiveness although several factors such as inflammation are likely to influence on prognosis.[3,4]

During the last decades, several histopathological grading systems for SC carcinomas in the HN region have been suggested and tested. The first grading systems only considered the morphological characteristics of the tumor, but later on, the tumor-host relationship also came into consideration.[5,6] For evaluation of tumor differentiation, nuclear polymorphism and keratinization have been important variables.[7,8] The characteristics of tumor invasion in the surrounding tissue have been implemented when evaluating the tumor-host relationship, as well as immune response (plasma-lymphocytic infiltration), vascular invasion, and perineural infiltration.[4,8] In particular, tumor budding (invading clusters of four or less tumor cells at the invasive front) has been proposed to be a simple and reliable prognostic marker for OSCC.[9]

Reproducibility in the scoring of histopathological parameters is essential if they are to be used as prognostic markers. The purpose

TABLE 1 Histopathological parameters evaluating tumor characteristics, invasion patterns, stromal reactions, and peritumoral tissue

	Score				
Variable	1	2	3	4	5
Differentiation, WHO whole tumor[7]	Well	Moderate	Poor		
Differentiation, WHO worst pattern[7]	Well	Moderate	Poor		
Degree, keratinization, whole tumor[8]	Highly keratinized (>50% of the cells)	Moderately keratinized (20%-50% of the cells)	Minimal keratinization (5%-20% of the cells)	No keratinization (0%-5% of the cells)	
Degree of keratinization, tumor front[15]	Highly keratinized (>50% of the cells)	Moderately keratinized (20%-50% of the cells)	Minimal keratinization (5%-20% of the cells)	No keratinization (0%-5% of the cells)	
Nuclear polymorphism, whole tumor[8]	Little nuclear polymorphism (>75% mature cells)	Moderately abundant nuclear polymorphism (50%-75% mature cells)	Abundant nuclear polymorphism (25%-50% mature cells)	Extreme nuclear polymorphism (0%-25% mature cells)	
Nuclear polymorphism tumor front[15]	Little nuclear polymorphism (>75% mature cells)	Moderately abundant nuclear polymorphism (50%-75% mature cells)	Abundant nuclear polymorphism (25%-50% mature cells)	Extreme nuclear polymorphism (0%-20% mature cells)	
Perineural infiltration	None	Nerves at invasive front	Nerves in tumor center		
Lymphocyte infiltration[4]	Marked/continuous band	Moderate/large patches	Little or none		
Worst pattern of invasion[4]	Pushing, well-delineated infiltrating borders	Infiltrating, solid cords, bands, and/or strands	Small groups of cords of infiltrating cells (n > 15 cells)	Marked and widespread cellular dissociation in small groups and/or in single cells (n < 15 cells)	Tumor satellites of any size with 1 mm or greater distance of intervening normal tissue (not fibrosis) at the tumor-host interface
Vascular infiltration	Present	Not present			
Infiltration	Subepithelial tissue (submucosa/lamina propria)	Muscle	Bone		

TABLE 2 Definition of categories

Tumor characteristics	Original variables	Categorization 1	Categorization 2	Categorization 3
WHO differentiation, whole tumor WHO pattern, worst pattern	Well	Low-grade		
	Moderate			
	Poor	High-grade		
Keratinization, whole tumor keratinization tumor front	Highly	Low-grade		
	Moderate			
	Minimal	High-grade		
	None			
Polymorphism, whole tumor polymorphism tumor front	Little/none	Low-grade		
	Moderate			
	Abundant	High-grade		
	Extreme			
Perineural infiltration	None	No		
	Invasive front	Yes		
	Tumor center			
Lymphocytic infiltrate	Marked	Marked	Abundant	
	Moderate	Not marked		
	Little		Little	
Pattern of invasion	Broad pushing	Low-grade	Low-grade	Low-grade
	Pushing fingers/large islands			
	Invasive islands >15 cells	Intermediate-grade	High-grade	
	Invasive islands <15 cells	High-grade		High-grade
	Tumor satellites			

of this study was to test the intra- and interrater agreement of a broad spectrum of parameters previously suggested as prognostic markers for OSCC.

2 | MATERIALS AND METHODS

2.1 | Observers and calibration

The observers were experienced pathologists/oral pathologist (TMS, EN, HL, ACJ, LUH, and SES), and two oral pathologist under training (DEC and EHO) from three university hospitals in Norway. Prior to the scoring, all the participants had taken part in two calibration workshops to agree on how to interpret the parameters.

One of the observers performed only one round of scoring, and one observer scored thickness and depth only once. The interrater observations were all calculated on the first set of scoring allowing all eight observers to participate.

2.2 | Slides for evaluation

Hematoxylin and eosin (HE) stained sections of 31 randomly selected formalin-fixed OSCC cases, representing various intraoral locations and different tumor stages, were distributed to each hospital. The participants reported each case with the assumption that the single slide available was representative of the whole lesion. No special stains were provided. The scoring was done independently by the observers for each variable at two different time points (3-6 month interval) permitting calculation of inter- and intrarater reliability. Thickness and tumor depth were measured in millimeter, but for statistical analyses, measurements were divided into three-triers; size ≤5 mm, 5.1-10.0 mm and >10 mm. This is according to The International Consortium for Outcome Research (ICOR) in Head and neck Cancer and TNM Classification of Malignant tumors.[10] Tumor budding was divided according to recommendations by Almangush et al[9]; <5 buds as low grade, 5-9 as intermediate grade, and ≥10 as high grade. The measurement of all other variables was on a nominal scale with three to six categories with no overlaps (mutually exclusive) as shown in Table 1.

2.3 | Ethics

The study was approved by the Northern Norwegian Regional Committee for Medical Research Ethics (REK Nord; 2013/1786 and 2015/1381).

TABLE 3 Mean percentage of agreement, prior to and after categorization

Tumor characteristics	Intrarater, mean agreement, prior to categorization (%)	Intrarater, mean agreement, after categorization (%)	Interrater, mean agreement before categorization (%)	Interrater, mean agreement, after categorization (%)
Differentiation, WHO whole tumor				
	77.0	92.5	53.9	84.4
Differentiation, WHO worst pattern				
	75.8	79.0	49.9	62.2
Degree, keratinization, whole tumor	60.7	77.1	46.6	71.4
Degree of keratinization, tumor front				
	67.3	90.5	54.0	74.6
Nuclear polymorphism, whole tumor				
	64.3	81.6	36.9	59.7
Nuclear polymorphism, tumor front				
	64.9	80.2	27.7	59.3
Perineural infiltration				
	79.9	88.9	55.9	66.4
Lymphocyte infiltration				
Category 1	69.2	87.3	50.9	78.6
Category 2		79.9		66.9
Worst pattern of invasion, categorized				
Category 1	67.3	71.5	50.1	58.2
Category 2		85.0		78.6
Category 3		83.5		74.7

2.4 | Statistics

Statistics was performed by using IBM SPSS statistics 24. We did statistical calculations both in percent agreement and Cohen's kappa (κ). The variability (spread of scoring) was low, and therefore, Cohen's kappa was of no/little value; thus, all correlations are given in percent.

3 | RESULTS

3.1 | Intrarater and interrater agreement

The first nine parameters in Table 1 had three to five different scoring options, and they were all categorized into new groups with fewer options (Table 2). The first seven were dichotomized, while the worst pattern of infiltration had two and three different scoring options (Table 2). Lymphocyte infiltration was dichotomized into two different groups according to different cut-offs. Dichotomizing variables increased the mean intrarater agreement from 68.3% (range 60.7%-77.0%) to 83.5% (range 77.1%-92.5%). Mean agreement for each variable prior to and after categorization is listed in Table 3. Prior to categorization, perineural infiltration showed the highest intrarater agreement, whereas differentiation was most agreed upon after categorization.

Some variables had predefined categories that were not changed (Table 4). These had a mean intrarater agreement of 85.4% (range 79.2%-93.3%), and vascular infiltration and infiltration into deeper tissues showed the highest intrarater agreement.

In order to evaluate interrater agreement, two observers from different hospitals were paired randomly. The average interrater agreement was lower than the intrarater agreement for all variables (Table 3). Dichotomizing variables increased the mean interrater agreement from 42.9% (range 26.7%-55.9%) to 70.6% (range 59.3%-84.4%). As for the intrarater agreement, perineural infiltration and differentiation had the highest agreement prior to and after categorization, respectively (Table 3). For the variables with predefined categories, the average interrater agreement was 72.7% (range 65.0%-78.9%), where vascular infiltration was the variable with the highest agreement (Table 4).

4 | DISCUSSION

This study was conducted to investigate how consistent an observer was at measuring the same histopathological variables at different times, as well as the consistency between different observers. We

TABLE 4 Intra- and interobserver agreement in variables with predefined categories, or unchanged variables

Tumor characteristics	Intrarater, mean agreement (%)	Interrater, mean agreement (%)
Number of buds, categorized		
Low, intermediate, high grade	79.2	72.8
Tumor thickness, categorized		
0-5.0 mm, 5.1-10 mm, >10 mm	80.7	69.6
Depth of invasion, categorized		
0-5.0 mm, 5.1-10 mm, >10 mm	80.7	65.0
Vascular infiltration		
Not found, present	93.3	78.9
Infiltration		
Subepithelial, muscle, bone	93.0	77.1

found that reducing the number of options for each variable by categorizing improved the correlation significantly.

Routine histopathological examination of OSCC is important for treatment decisions. Grading of differentiation is recommended by WHO, but the clinical value is limited. As a result, several other grading systems have been suggested. An optimal grading system should be reproducible, be applicable in all settings, and preferable be performed on standard HE stained sections by pathologist throughout the world.

Grading of epithelial dysplasia is poorly reproducible between observers, and to improve reproducibility, some advocate a binary system with only low- and high grade compared to mild, moderate, and severe dysplasia.[11] This has been evaluated in several studies on trials for oral epithelial dysplasia, but to a lesser extent in OSCC where the differentiation still is graded into well, moderate, and poor. In general, high-grade tumors (moderately and poorly differentiated) are related to higher degree of recurrence and shorter survival time than low-grade tumors.[12] Most OSCC are moderately or well differentiated, and the classification has not been found to correlate well with prognosis.[7] In this study, re-categorizing differentiation into low and high improved the reproducibility considerably.

In our study, the mean intrarater agreement was 70% for all variables before categorization compared to an interrater agreement of 48%. This indicates that the pairs of observers agreed on the tumor grading in less than half of the cases. Lack of reproducibility questions the reason for using a sophisticated grading system, and we, therefore, pooled scores into broader categories. This increased the mean of intra- and interrater agreement to 85% and 71%, respectively, using the best score for lymphocytic infiltration (categorization 1) and the best score for worst pattern of infiltration (categorization 2). The increase of agreement was less pronounced in the intrarater (14.9%) than in the interrater group (23.0%).

For the worst pattern of invasion, immunohistochemical staining for cytokeratin could ease the recognition of tumor cells among stromal cells, and thus, make scoring more reproducible. The low grade of the agreement for this variable in our study indicates that HE staining only is not sufficient. Likely, parameters such as number of buds, tumor thickness, depth of tumor, and infiltration into the underlying tissue could also benefit from special staining of epithelial cells. To our knowledge, no special staining has been promoted before scoring these variables in the different proposed scoring systems. The intrarater agreement for the budding, tumor thickness, and depth of infiltration was however 80% and 70% for the interrater groups, suggesting that HE stain is sufficient for categorized variables.

Other studies have shown a higher grade of interrater agreement compared to our results. In a study by Rodrigues et al,[13] 50 random samples of OSCC were selected and examined for worst pattern of invasion twice with a 2-week interval by calibrated observers. The intra- and interobserver agreement was strong in both cases (κ = 0.77-0.89 and 0.84). In a study evaluating tumor budding, Wang et al[14] found excellent intrarater and good interrater agreement (κ = 0.880/0.838 and 0.717). The high degree of agreement in these two studies might be due to the fact that they only had one parameter to score giving this full attention. Also, the time interval between first and second scoring in the first study was only 2 weeks with the possibility that the observer still could remember the previous scoring.

5 | CONCLUSION

To be of value, a tumor prognostic marker must be both reproducible and significantly associated with disease progression or survival. In this study, we have evaluated the reproducibility of a number of proposed histopathological prognostic markers in OSCC. Our findings suggest that simpler/uncomplicated scoring protocols will increase the reproducibility. However, we have not tested whether the new categorizations influence the prognostic value of the parameters. In our study, we included most of the previously proposed histopathological parameters and many observers, but a limited number of patient samples to avoid fatigue of the observers. We included tumors of different stages and from various intraoral locations; thus, the cohort was not suited for survival/prognostic analyses. The prognostic value of the revised categorization of the parameters should be tested in a larger, more homogenous cohort.

ACKNOWLEDGEMENTS

The publication charges for this article have been funded by a grant from the publication fund of UiT The Arctic University of Norway. All authors have filled out the ICMJE form for disclosure of potential conflicts of interest. ESN has declared that she has been a consultant for MSD Norway.

ORCID

Sonja E. Steigen https://orcid.org/0000-0001-8376-2489

REFERENCES

1. *The TNM Classification of Malignant Tumours*, 8th edn. Chichester, West Sussex, UK: Wiley; 2016.
2. *WHO Classification of Head and Neck Tumours*, 4th edn. Lyon: International Agency for Research on Cancer (IARC); 2017.
3. Anneroth G, Hansen LS. A methodologic study of histologic classification and grading of malignancy in oral squamous cell carcinoma. *Scand J Dent Res* 1984;92(5):448-468.
4. Brandwein-Gensler M, Teixeira MS, Lewis CM, et al. Oral squamous cell carcinoma: histologic risk assessment, but not margin status, is strongly predictive of local disease-free and overall survival. *Am J Surg Pathol*. 2005;29(2):167-178.
5. Broders AC. Squamous cell epithelioma of the lip. *J Am Med Assoc*. 1920;74:9.
6. Jakobsson PA, Eneroth CM, Killander D, Moberger G, Martensson B. Histologic classification and grading of malignancy in carcinoma of the larynx. *Acta Radiol*. 1973;12(1):1-8.
7. *Pathology and Genetics of Head and Neck Tumours (IARC WHO Classification of Tumours)*. Lyon: IARC Press; 2005.
8. Anneroth G, Batsakis J, Luna M. Review of the literature and a recommended system of malignancy grading in oral squamous cell carcinomas. *Scand J Dent Res*. 1987;95(3):229-249.
9. Almangush A, Pirinen M, Heikkinen I, Makitie AA, Salo T, Leivo I. Tumour budding in oral squamous cell carcinoma: a meta-analysis. *Br J Cancer*. 2018;118(4):577-586.
10. International Consortium for Outcome Research (ICOR) in Head and Neck Cancer, Ebrahimi A, Gil Z, et al. Primary tumor staging for oral cancer and a proposed modification incorporating depth of invasion: an international multicenter retrospective study. *JAMA Otolaryngol Head Neck Surg*. 2014;140(12):1138-1148.
11. Kujan O, Oliver RJ, Khattab A, Roberts SA, Thakker N, Sloan P. Evaluation of a new binary system of grading oral epithelial dysplasia for prediction of malignant transformation. *Oral Oncol*. 2006;42(10):987-993.
12. Padma R, Kalaivani A, Sundaresan S, Sathish P. The relationship between histological differentiation and disease recurrence of primary oral squamous cell carcinoma. *J Oral Maxillofac Pathol*. 2017;21(3):461.
13. Rodrigues PC, Miguel MC, Bagordakis E, et al. Clinicopathological prognostic factors of oral tongue squamous cell carcinoma: a retrospective study of 202 cases. *Int J Oral Maxillofac Surg*. 2014;43(7):795-801.
14. Wang C, Huang H, Huang Z, et al. Tumor budding correlates with poor prognosis and epithelial-mesenchymal transition in tongue squamous cell carcinoma. *J Oral Pathol Med*. 2011;40(7):545-551.
15. Bryne M, Boysen M, Alfsen CG, et al. The invasive front of carcinomas. The most important area for tumour prognosis? *Anticancer Res*. 1998;18(6B):4757-4764.

How to cite this article: Steigen SE, Søland TM, Nginamau ES, et al. Grading of oral squamous cell carcinomas – Intra and interrater agreeability: Simpler is better?. *J Oral Pathol Med*. 2020;00:1–6. https://doi.org/10.1111/jop.12990