ACTA OTORHINOLARYNGOLOGICA ITALICA 2022;42:26-33; doi: 10.14639/0392-100X-N1643

HEAD AND NECK

Systematic interpretation of confocal laser endomicroscopy: larynx and pharynx confocal imaging score

Interpretazione sistematica dell'endomicroscopia confocale laser: punteggio endomicroscopico confocale laringeo e faringeo

Matti Sievert[1], Konstantinos Mantsopoulos[1], Sarina K. Mueller[1], Markus Eckstein[2], Robin Rupp[1], Marc Aubreville[3], Florian Stelzle[4], Nicolai Oetter[4], Andreas Maier[5], Heinrich Iro[1], Miguel Goncalves[1,6]

[1] Department of Otorhinolaryngology, Head and Neck Surgery, Friedrich-Alexander-Universität Erlangen–Nürnberg, University Hospital, Erlangen, Germany; [2] Institute of Pathology, Friedrich-Alexander-Universität Erlangen–Nürnberg, University Hospital, Erlangen, Germany; [3] Technische Hochschule, Ingolstadt, Germany; [4] Department of Maxillofacial Surgery, Friedrich-Alexander-Universität Erlangen–Nürnberg, University Hospital, Erlangen, Germany; [5] Pattern Recognition Laboratory, Computer Science, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany; [6] Department of Otorhinolaryngology, Plastic Head and Neck Surgery, Rheinische Westfälische Technische Hochschule Aachen, University Hospital, Aachen, Germany

SUMMARY

Objective. Development and validation of a confocal laser endomicroscopy (CLE) classification score for the larynx and pharynx.

Methods. Thirteen patients (154 video sequences, 9240 images) with laryngeal or pharyngeal SCC were included in this prospective study between October 2020 and February 2021. Each CLE sequence was correlated with the gold standard of histopathological examination. Based on a dataset of 94 video sequences (5640 images), a scoring system was developed. In the remaining 60 sequences (3600 images), the score was validated by four CLE experts and four head and neck surgeons who were not familiar with CLE.

Results. Tissue homogeneity, cell size, borders and clusters, capillary loops and the nucleus/cytoplasm ratio were defined as the scoring criteria. Using this score, the CLE experts obtained an accuracy, sensitivity, and specificity of 90.8%, 95.1%, and 86.4%, respectively, and the CLE non-experts of 86.2%, 86.4%, and 86.1%. Interobserver agreement Fleiss' kappa was 0.8 and 0.6, respectively.

Conclusions. CLE can be reliably evaluated based on defined and reproducible imaging features, which demonstrate a high diagnostic value. CLE can be easily integrated into the intraoperative setting and generate real-time, in-vivo microscopic images to demarcate malignant changes.

KEY WORDS: confocal laser endomicroscopy, head and neck cancer, classification system, non-invasive histological imaging, larynx, pharynx

RIASSUNTO

Obiettivo. *Sviluppo e validazione di un punteggio di classificazione dell'endomicroscopia laser confocale (CLE) per la laringe ed il faringe.*

Metodi. *Tredici pazienti (154 sequenze video, 9240 immagini) con SCC laringeo o faringeo sono stati inclusi in questo studio prospettico tra ottobre 2020 e febbraio 2021. Ogni sequenza CLE è stata correlata con il gold standard, ovvero l'esame istopatologico fissato con ematossilina ed eosina (H&E). Sulla base di un set di dati di 94 sequenze video (5640 immagini), è stato sviluppato un sistema di punteggio. Nelle restanti 60 sequenze (3600 immagini), il punteggio è stato convalidato da quattro esperti di CLE e quattro chirurghi della testa e del collo che non avevano familiarità con la CLE.*

Risultati. *L'omogeneità dei tessuti, le dimensioni delle cellule, i bordi e i cluster, i loop capillari e il rapporto nucleo/citoplasma sono stati definiti come criteri di punteggio. Utilizzando questo punteggio, gli esperti di CLE hanno ottenuto un'accuratezza, una sensibilità e una specificità del 90,8%, 95,1% e 86,4%, rispettivamente, e i non esperti di CLE dell'86,2%, 86,4% e 86,1%. L'accordo tra osservatori Fleiss'kappa era rispettivamente di 0,8 e 0,6.*

Received: April 28, 2021
Accepted: September 28, 2021
Published online: February 7, 2022

Correspondence
Miguel Goncalves
Department of Otorhinolaryngology, Plastic Head and Neck Surgery, RWTH University of Aachen
Pauwelsstraße 30, 52074 Aachen, Germany
Tel. +49 024180-89360. Fax +49 0241 80-82465
E-mail: mgoncalves@ukaachen.de

How to cite this article: Sievert M, Mantsopoulos K, Mueller SK, et al. Systematic interpretation of confocal laser endomicroscopy: larynx and pharynx confocal imaging score. Acta Otorhinolaryngol Ital 2022;42:26-33. https://doi.org/10.14639/0392-100X-N1643

Conclusioni. CLE può essere valutato in modo affidabile sulla base di caratteristiche di imaging definite e riproducibili, che dimostrano un alto valore diagnostico. CLE può essere facilmente integrato nell'impostazione intraoperatoria e generare immagini microscopiche in tempo reale e in vivo per delimitare i cambiamenti maligni.

PAROLE CHIAVE: *endomicroscopia laser confocale, cancro della testa e del collo, sistema di classificazione, imaging istologico non invasivo, laringe, faringe*

Introduction

Non-invasive methods to characterise mucosal changes of the head and neck region, such as narrow-band imaging, confocal laser imaging, optical coherence tomography and fluorescence endoscopy, have been emerging recently [1-4]. One method that has shown a potential advance regarding its broad application outside of experimental studies is confocal laser endomicroscopy (CLE), which has demonstrated its value in the fairly similar oesophageal, gastric and intestinal epithelium in a wide range of clinical applications [5-8]. CLE is usually performed with the systemic application of fluorescein, which stains the intercellular spaces and allows real-time, non-invasive examination of the surface epithelium, cell features and fine capillary system to resemble a histological examination in detail and resolution [9]. Our group and others have published several reports in the last few years demonstrating the feasibility of CLE and emphasising its potential in distinguishing benign from malignant lesions of the vocal cords and oral and pharyngeal mucosa, as well as its application in the intraoperative assessment of free margins during oncological surgery [10-15]. The high accuracy of CLE reported in the latter studies, varying between 70 to 95%, was, however, obtained by experts in most cases and based on subjective analysis of the images without a criterial classification system. The primary objective of this study was to develop criteria to help differentiate benign from malignant lesions in the pharyngeal and laryngeal regions. The secondary objective was to examine the practicability of the scoring system carried out by head and neck surgeons inexperienced in CLE and to determine its diagnostic value.

Materials and methods

Study design

We conducted a prospective pilot diagnostic study with two different cohorts, one to develop the scoring system for assessing the accuracy of CLE and one to validate this scoring system. The study was performed at a tertiary hospital and academic cancer centre (Department of Otorhinolaryngology, Head and Neck Surgery, Friedrich Alexander University of Erlangen-Nuremberg, Erlangen, Germany). Based on the German Guidelines for laryngeal carcinoma, our treatment was performed on the basis of clinical parameters independent of confocal laser endomicroscopy findings.

Eligibility criteria

A total of 13 consecutive patients were included in this study. Inclusion criteria were all sequential patients with a histological diagnosis of laryngeal and pharyngeal SCC (i.e. after panendoscopy and biopsy) in which open surgery was required and appropriate for the oncological treatment. Exclusion criteria were prior treatment of any head and neck cancer, distant metastasis, radiotherapy in the head and neck area, pregnancy, thyroid dysfunction, underaged patients, severe kidney failure and allergy to fluorescein.

Technical details

We used cold instruments to reduce thermal damage to the resection margin. After accessing the pharynx or larynx as required, we exposed the tumour to initiate CLE imaging. Image acquisition was performed using a GastroFlex probe combined with a 488 nm Cellvizio laser scanning system (Mauna Technologies, Paris, France). The 2.6 mm diameter probe has a field of view of 240 µm and a resolution of 1 µm with a penetration depth of 55-65 µm. Subsequently, 2.5 ml fluorescein alcon 10% was injected intravenously, followed by an additional 2.5 ml fluorescein administration after 8 minutes. In order to correlate the CLE imaging with the gold standard of histopathology, the regions recorded were marked with suture, or a separate biopsy was performed. The complete resection was histologically confirmed intraoperatively following the standard protocol using frozen sections around the circumference of the defect. The histopathological assessment followed a standard protocol with haematoxylin and eosin (H&E) staining.

Development of the evaluation criteria

We developed the evaluation criteria according to a 4-step protocol. For the creation and validation of the scoring criteria, we used different sequences of different patients. For data processing, we analysed the sequences using Cellvizio Viewer software 1.6.2 (Mauna Kea Technologies, Paris, France).
Step 1 - Out of 28306 CLE frames from eight consecutive patients, we selected a dataset of 94 video sequences (5640 images; 47 sequences of SCC and 47 sequences of benign mucosa) from four examiners for analysis to create

dedicated evaluation criteria. The blinded examiners were asked to make a binary statement on dignity by assessing the cell morphology and vascular situation. In addition, the raters were allowed to comment on the individual findings.
Step 2 - The authors analysed the raters' results and checked for regularities between the different raters' decisions and their comments.
Step 3 - The authors obtained a consensus to develop a score (Larynx and Pharynx Confocal Imaging Score – LF-CIS) based on the results and comments. The data set analysis aimed to define simple, effective and easily applicable parameters for daily routine practice.
Step 4 - We validated the score by four experienced raters ("experts") consisting of one pathologist (P) and three surgeons (S) (who had independently performed or analysed a minimum of 20 CLE examinations) and four inexperienced raters (non-experts, without any previous experience in this technique). The pathologist who rated the CLE images was not involved and had no information about the histopathological examination of the probes and specimens. The histopathological examination, as reference standard, was performed independently by other pathologists not directly involved in the study.
Out of 12,065 new images (recorded in 5 additional consecutive patients), 60 sequences (3600 images; 33 sequences of SCC and 27 sequences of benign mucosa) were selected and evaluated as representative in acceptable quality. This dataset was kept separate from the development of the score to reduce overfitting and increase validity, to ensure that the score was tested in yet unseen images. We matched each of these sequences (60 images, 5 seconds) with a corresponding sample by H&E staining to determine the diagnostic accuracy of the score.

Data analysis
Absolute values are presented with mean and standard deviation (SD). Relative values are presented in absolute and relative frequencies and were compared using the chi-square test. Sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and accuracy were calculated for each investigator, as well as for each investigator group (expert and non-expert) and presented as percentage and 95% confidence interval (95% CI). We tested the interrater reliability/agreement using the Cohen's kappa and Fleiss' kappa coefficient. According to Landis and Koch, values of κ between 0 and 20 are defined as low, between 0.21 and 0.40 as adequate, between 0.41 and 0.60 as moderate, between 0.61 and 0.80 as substantial, and between 0.81 and 1.0 almost perfect [16]. We performed a receiver operating characteristic (ROC) analysis to calculate the area under the curve (AUC) as a quality measure for the DOC score as a classifier. Spearman's rank correlation coefficient is used to calculate the relationship between the score value and the probability of a malignant finding. A p-value of less than $p \leq 0.05$ was considered statistically significant. We performed statistical analysis using SPSS version 22.0 (IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY, USA).

Results

Patient characteristics
Between March 2020 and February 2021, we enrolled 13 patients (one female and 12 males; mean age 64.7 years (SD = 8.6) to undergo pharyngeal in vivo CLE during planned transcervical tumour resection in the area of the hypopharynx and larynx. In six patients (46.2%), the tumour mass was located in the larynx. Seven patients (53.8%) additionally presented with involvement of the hypopharyngeal mucosa. The tumour resection was performed via an open approach in each case. We performed a partial laryngectomy via vertical incision of the thyroid cartilage in two patients (15.4%; No. 1 and 2). In addition, we dissected an apron flap for total laryngectomy in 11 patients (84.6%; No. 3-13). Microvascular defect reconstruction was necessary in 7 patients (53.8%). Regarding tumour grading, three patients (23.1%) were confirmed as having an intermediate grade (G2) and 10 patients (76.9%) poor grade differentiation (G3). Patient characteristics, including stage, are shown in Table I. In all cases, safe margin resection could be performed independent of the use of CLE.

Definition of diagnosis-related parameters and creation of a scoring system
The details of all four raters' results regarding the CLE morphological criteria are presented in Table II. Based on the four raters' evaluation, we attributed a high value to the abnormal tissue structure with an inhomogeneous cell pattern ($\kappa = 0.67$) and variance in cell size and shape ($\kappa = 0.64$) for the assessment of dignity. We therefore decided to evaluate both criteria (an inhomogeneous cell pattern and various cells in terms of size, shape) with one point (+1), respectively (Fig. 1c, e, g). In addition, we defined small, dark cells as an indication of an abnormal nuclear-cytoplasmic correlation and cells of different greyscales as suspicious and rated them as +1 point (Fig. 1g). The normal squamous epithelium of the hypopharynx and larynx shows a honeycomb-like pattern (Fig. 1a) with regular intercellular spaces. The presence of these honeycombs was defined by the authors as a benign criterion and was thus scored as -1 point. Disrupted cell borders were assigned to malignant findings and were therefore assigned values of +1 point

Table I. Characteristics of the patient cohort.

Case No.	Age (years)	Tumour stage	Location	Grade	Surgery	CLE frames (n)	Recording time (seconds)	Selected sequences (n)
Score creation								
1.	67	T1	Larynx	G2	Partial vertical LE	7586	948	10
2.	60	T1	Larynx	G2	Partial vertical LE	2841	355	13
3.	68	T4a	Larynx	G3	Total LE	1919	239	11
4.	73	T4a	Larynx	G3	Total LE	2413	301	12
5.	57	T3	Larynx	G3	Total LE	6513	814	11
6.	69	T3	Larynx Hypopharynx	G3	Total LE, partial pharyngectomy	2398	299	11
7.	63	T3	Larynx Hypopharynx	G3	Total LE, partial pharyngectomy	2896	362	14
8.	57	T3	Larynx Hypopharynx	G3	Total LE, partial pharyngectomy	1743	217	12
				Total		28309	3535	94
Score validation								
9.	71	T4a	Larynx Hypopharynx	G3	Total LE, partial pharyngectomy	1468	183	14
10.	56	T4a	Larynx Hypopharynx	G3	Total LE, partial pharyngectomy	2204	275	12
11.	86	T4a	Larynx	G2	Total LE, partial pharyngectomy	2191	273	10
12.	61	T2	Larynx Hypopharynx	G3	Total LE, partial pharyngectomy	3311	413	14
13.	53	T3	Larynx Hypopharynx	G3	Total LE, partial pharyngectomy	2891	361	10
				Total		12065	1505	60

LE: laryngectomy.

(Fig. 1e). Concerning blood vessels, there was a lower consensus of the raters (κ = 0.45). This may be caused by the fact that, in some tissue sections, no blood vessels were visible. On the other hand, dilated vessels with vascular leakage can also be found in inflammatory altered mucosa. However, capillary loops (Fig. 1b) are considered a benign criterion by the authors, who rated it as -1 point.

Based on the current analysis and the authors' consensus and experience, we defined a six-category score. In general, benign criteria are scored as -1. In contrast, criteria suspected to be malignant are scored as +1 (see details in Table III), resulting in a rating scale with a minimum of -2 points and a maximum of +5 points. All images were evaluated by the authors MS and MG according to the current score, and an optimal cut-off value of ≥ 3 points was defined in the ROC analysis on the score creation set (AUC = 0.86; 95%CI: 0.81-0.92).

Diagnostic metrics in score-validation

When reviewing the 60 sequences, the value -2 was assigned in 82 (17.0%), the value -1 in 79 (16.4%), the value 0 in 36 (7.5%), the value 1 in 22 (5.6%), the value 2 in 14 (2.9%), the value 3 in 35 (7.3%), the value 4 in 48 (10%), and the value 5 in 164 cases (34.2%). The rank correlation was calculated as 0.77 ($p < 0.001$).

The performance of the scoring classification is illustrated in Table IV. We considered a score of ≥ 3 points to be malignant. The examiners achieved an accuracy, sensitivity, specificity, PPV and NPV ranging from 83.3% to 96.7%, 72.7% to 100%, 63.0% to 100%, 76.7% to 100% and 74.3% to 100%, respectively. In summary, we showed an accuracy, sensitivity, specificity, PPV and NPV of 88.5%, 87.1%, 90.3%, 91.6% and 85.1%, respectively. The observers' overall agreement was calculated with a Fleiss' κ of 0.66 and can therefore be considered substantial (Tab. V).

The expert group's selective analysis revealed an accuracy, sensitivity, specificity, PPV and NPV of 90.8%, 95.1%, 86.4%, 87.9% and 94.4%, respectively, with an inter-rater reliability of κ = 0.78. In the non-expert group, an accuracy, sensitivity, specificity, PPV and NPV of 86.2%, 86.4%,

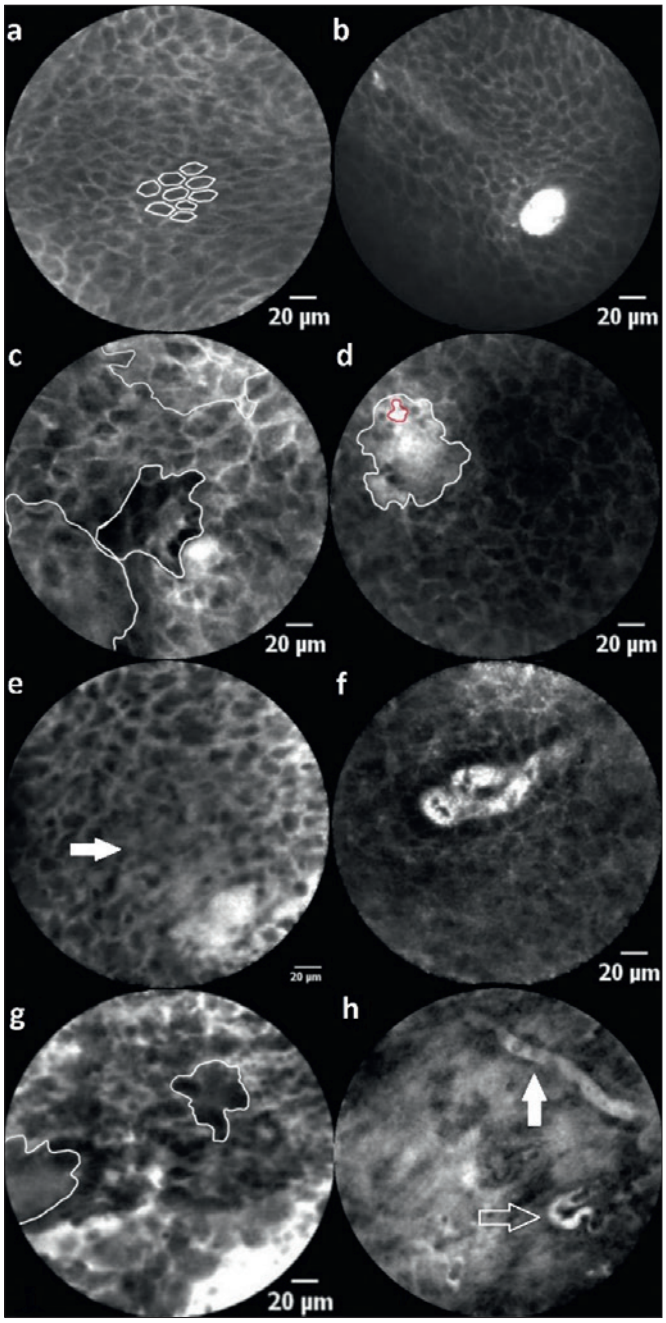

Figure 1. Description: (**a**) inconspicuous squamous epithelium with a classic "honeycomb" pattern and (**b**) capillary loops (rated as -2 points). (**c**) Inhomogenous global impression with areas of different grey levels. (**d**) Substantial leakage of fluorescein (white area) around a capillary (red area) indicates malignantly differentiated or altered inflammatory mucosa. (**e**) The arrow marks elapsed cell borders. Besides, there is vascular leakage (lower right) and an inhomogeneous cell pattern with varying cell sizes. In conclusion, this image is scored as 4 points. (**f**) Dilated, atypical capillary loop, which is interpreted as suspicious. (**g**) Cell conglomerates (circled areas), inhomogeneous cell pattern and size adjacent to dilated vessels and leakage (scored 5 points). (**h**) Atypical blood vessels with a horizontal course in the superficial epithelial layer (white arrow) and a corkscrew-like shape as a sign of neoangiogenesis (transparent arrow). This image is scored as 5 points in total.

86.1%, 88.4% and 83.8% could be observed ($\kappa = 0.58$). We calculated interrater reliability between the experienced and inexperienced raters of $\kappa = 0.77$.

In the ROC analysis, we calculated an AUC of 0.91 (95%CI: 0.88-0.94) considering the results of all investigators (Fig. 2), confirming an optimal cut-off value of 3 points to achieve the best sensitivity and specificity.

Discussion

A classification system, LP-CIS, was developed to help interpretate the CLE examination and objectively differentiate between malignant and benign mucosal lesions. It was developed based on 94 sequences (eight patients, 5640 images, respectively 50% benign and malignant sequences). Tissue homogeneity, cell size variation, presence of cell clusters and honeycomb pattern, presence of capillary loops, presence of small dark cells as well as the presence of clear cell borders were regarded as adequate characteristics from a statistical and practicability standpoint to differentiate benign from malignant lesions using CLE (Tabs. II and III). Blinded to the gold standard of histopathological examination, differentiation of benign and malignant lesions was possible with an accuracy, sensitivity, and specificity of 90.8%, 95.1%, 86.4%, respectively, for the CLE-experienced group on a new set of 5 different patients, to exclude bias (60 sequences, 3600 images). The diagnostic metrics were comparable for CLE-inexperienced head and neck surgeons, albeit slightly inferiorly, with an accuracy, sensitivity and specificity of 86.2%, 86.4% and 86.1%, respectively. Interrater reliability between the experts was

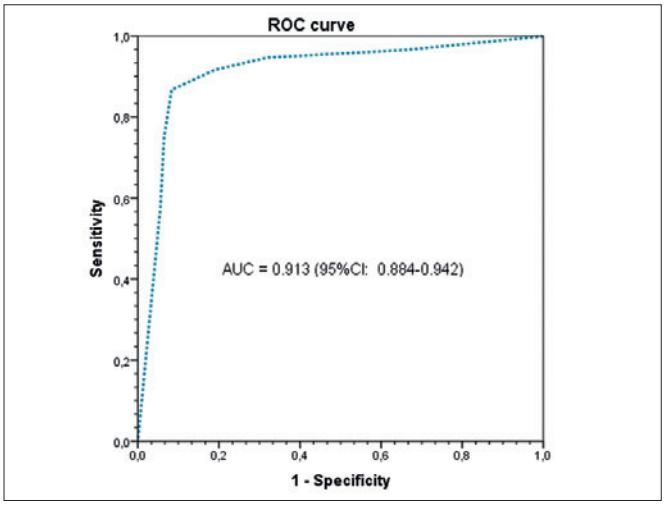

Figure 2. Receiver operating characteristic (ROC) analysis and the area under the curve (AUC) of the larynx and pharynx confocal imaging score. The optimal cut-off value is calculated as 3 points.

Table II. Descriptive analysis of typical characteristics for differentiating between SCC and benign squamous epithelium by CLE.

Image characteristics	SCC (n = 188)	Healthy mucosa (n = 188)	p value	Fleiss' kappa
Tissue inhomogeneity	147 (78.2%)	39 (20.7%)	$p < 0.001$	0.67
Different greyscale value and shape of cells	151 (80.3%)	41 (21.8%)	$p < 0.001$	0.64
Atypical vessels/ absence of capillary loops	103 (54.8%)	22 (11.7%)	$p < 0.001$	0.45
Elapsed or non-differentiable cell borders	144 (76.6%)	40 (21.3%)	$p < 0.001$	0.58

SCC: squamous cell carcinoma; the data refer to the evaluation of 47 sequences each from SCC and benign mucosal areas, each assessed by 4 raters. In total, 188 sequences of SCC and 188 sequences of healthy mucosa were presented.

very good and was almost perfect. In the CLE-naive group, interrater reliability was nevertheless substantial. Our group's previously published data, with a similar technical methodology, although solely by experts without systematic classification, obtained accuracy, sensitivity and specificity values of 80-86%, 72-90% and 72-89% respectively [10,11]. The classification of images in previous publications was based on published reports for oral cancer and oesophageal cancer, and was then transferred and applied to the pharyngeal and laryngeal regions. Our results suggest that a systematic classification system, which was specifically developed and refined for the laryngeal and pharyngeal segments of the upper airway tract, can significantly improve the diagnostic accuracy of the experts and, more surprisingly, enable CLE-inexperienced surgeons to identify malignant changes on the same level as a CLE expert without the new score.

Interpretation of CLE images is subjective and a crucial factor in making a correct assessment that, up to now, was based solely on the experience of examiners and the transferability of knowledge from other areas such as the oesophageal mucosa and oral mucosa [12,13]. The present study demonstrates that a systematic classification based on the pharyngeal and laryngeal regions can help improve experts' diagnostic accuracy in the same anatomic region and enable non-experts to achieve satisfactory accuracy comparable to that of experts up to this point.

Tissue homogeneity, clear cell borders with regular, similarly-sized cells with a small nucleus/cytoplasm ratio, which ideally form a honeycomb pattern, are typical features of benign mucosal tissue, both in histology as well as in CLE throughout the upper airway tract (Tab. III) [10,11,13,17,18]. Atypical vessels with fluorescein leakage as a sign of a distorted barrier are also a feature, which is much more frequent in malignant lesions than in benign lesions (Tab. II) [19]. These features were also evaluated from the standpoint of practicability and the validated Spearman's rank correlation coefficient ($p < 0.001$; Tab. II) [13,17].

While helping the intraoperative characterisation of free margins, we emphasise that this method cannot replace the time-consuming but still, up to this point, demonstrably more reliable frozen sections. The intraoperative frozen section has a reported accuracy of 98.6-88.2%, and despite increasing criticism of this method in recent years is broadly used in oncologic surgery [20,21]. The main argument against the frozen sections for safe margin evaluation is the inadequacy of the tissue sent for intraoperative analysis due to the under-sampling of biopsies, which are usually small, fragmented, and not oriented [21]. We demonstrated an accuracy of 91%; however, further research is needed to verify the value of CLE in cancer demarcation compared with fresh frozen sections before it is suitable for broad clinical use in this regard. Studies in the head and neck and other epithelial areas of the intestinal tract, such as rectal surgery, with similar methodology, show a feasibility and accuracy of up to 94% [22]. CLE is not limited to under-sampling and enables a software-based reconstruction of the whole plane of tissue around the resection area. CLE is, however, limited

Table III. Larynx and pharynx confocal imaging score (LP-CIS).

Tissue homogeneity	Cell size (+/-20%)	Cell cluster	Blood vessels	Nucleus/cytoplasm ratio	Cell borders
0 = homogeneous	0 = uniform	-1 = honeycomb	-1 = capillary loops	0 = no small dark cells	0 = clear cell borders
1 = inhomogeneous	1 = different	1 = no honeycombs	0 = no capillary loops and/or atypical vessels/ horizontally running vessels/leakage	1 = small and or dark cells	1 = disrupted cell borders and/or conglomerates

≥ 3 points are considered suspicious (range: minimum -2 to maximum 5 points).

Table IV. Data analysis by the various raters (larynx and pharynx confocal imaging score).

Diagnostic metrics	Expert 1 (P)	Expert 2 (S)	Expert 3 (S)	Expert 4 (S)	Nonexpert 1 (S)	Nonexpert 2 (S)	Nonexpert 3 (S)	Nonexpert 4 (S)	Experts (1-4)	Nonexperts (1-4)	All raters
	Value (95%CI)	Value (95%CI)	Value (95%CI)	Value (95%CI)	Value (95%CI)	Value (95%CI)	Value (95%CI)	Value (95%CI)	Value (95%CI)	Value (95%CI)	Value (95%CI)
Accuracy	86.7% (75.4-94.1)	90.0% (79.5-96.2)	96.7% (88.5-99.6)	95% (86.1-99.0)	95.0% (86.1-99.0)	83.3% (71.5-91.7)	83.3% (71.5-91.7)	83.3% (71.5-91.7)	90.8% (86.4-94.2)	86.2% (81.2-90.3)	88.5% (85.3-91.2)
Sensitivity	75.8% (57.7-88.9)	90.9% (75.7-98.1)	93.9% (79.8-99.3)	90.1% (75.7-98.1)	90.9% (75.7-98.1)	100% (89.4-100)	81.8% (64.5-93.0)	72.7% (54.5-86.7)	95.1% (89.6-98.2)	86.4% (79.3-91.7)	87.1% (82.5-90.9)
Specificity	100% (87.2-100)	88.9% (70.8-97.6)	100% (87.2-100)	100% (87.2-100)	100% (87.2-100)	63.0% (42.4-80.6)	85.2% (66.3-95.8)	96.3% (81.0-99.9)	86.4% (78.9-92.0)	86.1% (78.1-92.0)	90.3% (85.5-93.9)
PPV	100%	90.9% (77.4-96.7)	100%	100%	100%	76.7% (66.9-84.4)	87.1% (72.9-94.4)	96.0% (77.6-99.4)	87.9% (82.1-92.0)	88.4% (82.5-92.4)	91.6% (87.9-94.3)
NPV	77.1% (64.9-86.0)	88.9% (72.9-96.0)	93.1% (77.9-98.1)	90.0% (75.4-96.4)	90.0% (75.4-96.4)	100%	79.3% (64.6-88.9)	74.3% (62.2-83.5)	94.4% (88.6-97.4)	83.8% (78.0-88.9)	85.1% (80.7-88.7)
ROC(AUC)	0.92 (0.86-0.92)	0.92 (0.85-0.99)	0.98 (0.96-1.00)	0.94 (0.89-1.00)	0.93 (0.86-1.00)	0.81 (0.71-0.91)	0.90 (0.83-0.97)	0.95 (0.90-1.00)	0.94 (0.90-0.97)	0.89 (0.84-0.94)	0.91 (0.88-0.94)

S: surgeon; P: pathologist; PPV: positive predictive value; NPV: negative predictive value; 95%CI: 95% Confidence Interval; ROC: receiver operating characteristic; AUC: area under the curve.

by a fixed tissue penetration depth of 55-65 µm, which is determined by the probe used in our study. To characterise the morphological features of regular epithelium and carcinoma, we used IV fluorescein as a contrast agent. This allowed the sufficient resolution of in vivo cellular architecture within a few minutes of fluorescein administration and did not result in any complications or side effects. Fluorescein is also the only contrast agent with regulatory approval as a drug-device combination, granted a 510 (k) clearance by the FDA in January 2020. Other contrast agents, such as the topic application of acriflavine, provide further contrast and a better depiction of the nuclei. However, because it is carcinogenic, an application outside of clinical studies up to this point is unforeseeable [18,23]. Recently, molecular imaging has also increased in interest using complementary contrasting agents such as glucose analog 2-NBDG or EGFR antibodies to enhance extra- and intercellular architecture visualisation and improve the delineation of tumor margins [24,25]. Our scoring system is based on fluorescein alone, as this is the only approved drug-device combination with CLE and will probably remain so for years to come.

Conclusions

CLE is a promising imaging technology that may improve the intraoperative management of laryngeal and pharyngeal carcinoma by generating real-time, in-vivo microscopic images with the reliable demarcation of cancer from benign mucosa. In this study, a scoring system for the diagnosis of laryngeal and pharyngeal carcinoma was developed for the first time. LP-CIS enables a reliable classification of physiological mucosa and cancerous lesions even for CLE nonexperts in their daily clinical practice. It is an essential step to integrating CLE into broad clinical practice outside of the clinical-experimental setting.

Table V. Inter-observer agreement in score validation.

Observer pair	Fleiss-Kappa / Cohens-Kappa*
Experts	0.78
Non-experts	0.58
Experts *vs* non-experts	0.77*
All raters	0.66

Inter-observer agreement for multiple raters (Fleiss' kappa) and observer groups (Cohen's kappa).*

Conflict of interest statement

The authors declare no conflict of interest.

Funding

This project was supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under

grant number GO 3182/2-1, MA 4898/17-1, OE 743/1-1, STE 1877/7-1; Project Number 439264659.

Authors' contributions

MS and MG: conception and design, acquisition of data, analysis and interpretation of data, wrote the manuscript, final approval of the version to be published. KM, SM, ME, RR: acquisition of data, analysis and interpretation of data, critical revision of the article, final approval of the version to be published.MA, FS, NO, AM, HI: critical revision of the article, final approval of the version to be published. MG: drafting and critical revision of the article, final approval of the version to be published.

Ethical consideration

The study was approved by the local institutional ethics committee (approval number 60_14 B) and carried out in accordance with the Declaration of Helsinki. We obtained written informed consent from all study participants.

References

1 Chabrillac E, Dupret-Bories A, Vairel B, et al. Narrow-band imaging in oncologic otorhinolaryngology: state of the art. Eur Ann Otorhinolaryngol Head Neck Dis 2021;12:S1879-7296(21)00051-X. https://doi.org/10.1016/j.anorl.2021.03.004

2 Sievert M, Auberville M, Oetter N, et al. Konfokale Laser-Endomikroskopie des Kopf-Hals-Plattenepithelkarzinoms: eine systematische Übersicht [Confocal laser endomicroscopy of head and neck squamous cell carcinoma: a systematic review]. Laryngorhinootologie 2021;100:875-881. https://doi.org/10.1055/a-1339-1635

3 De Leeuw F, Abbaci M, Casiraghi O, et al. Value of full-field optical coherence tomography imaging for the histological assessment of head and neck cancer. Lasers Surg Med 2020;52:768-778. https://doi.org/10.1002/lsm.23223

4 Hart ZP, Nishio N, Krishnan G, et al. Endoscopic fluorescence-guided surgery for sinonasal cancer using an antibody-dye conjugate. Laryngoscope 2020;130:2811-2817. https://doi.org/10.1002/lary.28483

5 Yang X, Liu W. Current evidence on confocal laser endomicroscopy for noninvasive head and neck cancer imaging. Acta Otorhinolaryngol Ital 2020;40:396-398. https://doi.org/10.14639/0392-100X-N0801

6 Taunk P, Atkinson CD, Lichtenstein D, et al. Computer-assisted assessment of colonic polyp histopathology using probe-based confocal laser endomicroscopy. Int J Colorectal Dis 2019;34:2043-2051. https://doi.org/10.1007/s00384-019-03406-y

7 Tahara T, Horiguchi N, Terada T, et al. Diagnostic utility of probe-based confocal laser endomicroscopy in superficial non-ampullary duodenal epithelial tumours. Endosc Int Open 2019;7:E1515-E1521 https://doi.org/10.1055/a-0999-5282

8 Napoleon B, Krishna SG, Marco B, et al. Confocal endomicroscopy for evaluation of pancreatic cystic lesions: a systematic review and international Delphi consensus report. Endosc Int Open 2020;8:E1566-E1581. https://doi.org/10.1055/a-1229-4156

9 Neumann H, Kiesslich R, Wallace MB, et al. Confocal laser endomicroscopy: technical advances and clinical applications. Gastroenterology 2010;139:388-392. https://doi.org/10.1053/j.gastro.2010.06.029

10 Sievert M, Stelzle F, Aubreville M, et al. Intraoperative free margins assessment of oropharyngeal squamous cell carcinoma with confocal laser endomicroscopy: a pilot study. Eur Arch Otorhinolaryngol 2021;278:4433-4439. https://doi.org/10.1007/s00405-021-06659-y

11 Sievert M, Oetter N, Aubreville M, et al. Feasibility of intraoperative assessment of safe surgical margins during laryngectomy with confocal laser endomicroscopy: a pilot study. Auris Nasus Larynx 2021;48:764-769. https://doi.org/10.1016/j.anl.2021.01.005

12 Goncalves M, Aubreville M, Mueller SK, et al. Probe-based confocal laser endomicroscopy in detecting malignant lesions of vocal folds. Acta Otorhinolaryngol Ital 2019;39:389-395. https://doi.org/10.14639/0392-100X-2121

13 Oetter N, Knipfer C, Rohde M, et al. Development and validation of a classification and scoring system for the diagnosis of oral squamous cell carcinomas through confocal laser endomicroscopy. J Transl Med 2016;14:159. https://doi.org/10.1186/s12967-016-0919-4

14 Moore C, Mehta V, Ma X, et al. Interobserver agreement of confocal laser endomicroscopy for detection of head and neck neoplasia. Laryngoscope 2016;126:632-637. https://doi.org/10.1002/lary.25646 10.1002/lary.25646

15 Nathan CA, Kaskas NM, Ma X, et al. Confocal laser endomicroscopy in the detection of head and neck precancerous lesions. Otolaryngol Head Neck Surg 2014;151:73-80. https://doi.org/10.1177/0194599814528660 10.1177/0194599814528660

16 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.

17 Liu H, Li YQ, Yu T, et al. Confocal endomicroscopy for in vivo detection of microvascular architecture in normal and malignant lesions of upper gastrointestinal tract. J Gastroenterol Hepatol 2008;23:56-61. https://doi.org/10.1111/j.1440-1746.2007.05221.x

18 Shinohara S, Funabiki K, Kikuchi M, et al. Real-time imaging of head and neck squamous cell carcinomas using confocal micro-endoscopy and applicable dye: a preliminary study. Auris Nasus Larynx 2020;47:668-675. https://doi.org/10.1016/j.anl.2020.02.001

19 Wallace M, Lauwers GY, Chen Y, et al. Miami classification for probe-based confocal laser endomicroscopy. Endoscopy 2011;43:882-891. https://doi.org/10.1055/s-0030-1256632

20 Layfield EM, Schmidt RL, Esebua M, et al. Frozen section evaluation of margin status in primary squamous cell carcinomas of the head and neck: a correlation study of frozen section and final diagnoses. Head Neck Pathol 2018;12:175-180. https://doi.org/10.1007/s12105-017-0846-6

21 Kubik MW, Sridharan S, Varvares MA, et al. Intraoperative margin assessment in head and neck cancer: a case of misuse and abuse? Head Neck Pathol 2020;14:291-302. https://doi.org/10.1007/s12105-019-01121-2

22 Liu Z, Luo X, Jiang W, et al. Real-time in vivo optical biopsy using confocal laser endomicroscopy to evaluate distal margin in situ and determine surgical procedure in low rectal cancer. Surg Endosc 2019;33:2332-2338. https://doi.org/10.1007/s00464-018-6519-z

23 Paull PE, Hyatt BJ, Wassef W, et al. Confocal laser endomicroscopy: a primer for pathologists. Arch Pathol Lab Med 2011;135:1343-1348. https://doi.org/10.5858/arpa.2010-0264-RA

24 Watermann A, Gieringer R, Bauer AM, et al. Fluorescein- and EGFR-antibody conjugated silica nanoparticles for enhancement of real-time tumour border definition using confocal laser endomicroscopy in squamous cell carcinoma of the head and neck. Nanomaterials (Basel) 2019;9:1378. https://doi.org/10.3390/nano9101378

25 Belykh E, Jubran JH, George LL, et al. Molecular imaging of glucose metabolism for intraoperative fluorescence guidance during glioma surgery. Mol Imaging Biol 2021;23:586-596. https://doi.org/10.1007/s11307-021-01579-z