Dentomaxillofacial Radiology (2018) 47, 20180099

birpublications.org/dmfr

RESEARCH ARTICLE

Spatial agreement of demineralized areas in quantitative light-induced fluorescence images and digital photographs

[1]Rosalia Tatano, [1]Benjamin Berkels, [2]Eva E Ehrlich, [3]Thomas M Deserno and [2]Ulrike B Fritz

[1]Aachen Institute for Advanced Study in Computational Engineering Science (AICES), RWTH Aachen University, Aachen, Germany; [2]Klinik für Kieferorthopädie, Uniklinik RWTH Aachen, Aachen, Germany; [3]Peter L. Reichertz Institute for Medical Informatics, University of Braunschweig - Institute of Technology and Hannover Medical School, Braunschweig, Germany

Objectives: Previous work has shown qualitatively that detection of demineralized tooth areas (white spot lesions, WSLs) is more reliable in digital photographs (DP) as in quantitative light-induced fluorescence (QLF) images. Based on non-rigid, multimodal image registration, we now quantitatively compare manual and automatic markings in both modalities.
Methods: After braces removal, pairs of DP and QLF were acquired from 124 teeth of 31 patients. Three experienced raters marked the WSL on both DP and QLF images, each of which was presented twice in randomized order. For each tooth and each modality, a ground truth (GT) was established using the simultaneous truth and performance level estimation algorithm on the total of six manual markings per image. DP and QLF image pairs were spatially registered, by aligning the outline of the tooth area in DPs to that of the corresponding tooth area in QLF. Between all pairs of markings for all teeth, position and size were compared quantitatively by the Dice coefficient and the novel coefficient of inclusion.
Results: Our hypotheses: (i) the clinical inspection supported by DP is more sensitive to WSL as that by QLF, disregarding whether the automatic analysis or the experts' manual assessment of QLF is applied, and (ii) detected lesions in QLF are included in those of DP, were confirmed and not confirmed, respectively.
Conclusion: DP and QLF are valuable methods to detect WSL in demineralized teeth. Combining both modalities can provide additional information on early lesion assessment.

Dentomaxillofacial Radiology (2018) **47**, 20180099. doi: 10.1259/dmfr.20180099

Cite this article as: Tatano R, Berkels B, Ehrlich EE, Deserno TM, Fritz UB. Spatial agreement of demineralized areas in quantitative light-induced fluorescence images and digital photographs. *Dentomaxillofac Radiol* 2018; **47**: 20180099.

Keywords: Demineralization; QLF; digital photography; image registration; ground truth

Introduction

White spot lesions (WSLs) are an undesirable iatrogenic side effect of orthodontic treatment with fixed appliances.[1,2] A recently published meta-analysis including 14 clinical studies confirmed that 68.4% of patients undergoing orthodontic treatment had WSLs.[3]

Both in clinical inspection and in digital photographs (DP), WSLs appear as white-opaque enamel areas. Beside the aesthetic impairment,[4] WSLs are characterized by mineral loss[5] and considered as primary stage of caries (incipient carious lesions).

In addition to clinical diagnosis, various non-invasive methods were developed to quantify WSLs, summarized by Angmar-Månsson[6] and Gomez et al:[7] laser fluorescence, fibre-optic transillumination, electrical conductance and quantitative light-induced fluorescence (QLF). The QLF diagnostic capacity relies on the different fluorescence properties of sound and demineralized enamel. In particular, the QLF system is based on the principle that the change in the intensity of natural fluorescence of a tooth is due to dental

Correspondence to: Mrs Rosalia Tatano, E-mail: tatano@aices.rwth-aachen.de
Received 09 March 2018; revised 24 May 2018; accepted 25 May 2018

hard tissue changes caused by a caries lesion.[8] However, the systematic review of Gomez et al revealed a large variation in sensitivity and specificity for the considered four methods, and a lack of consistency in definition of the disease and analytical assessment. In addition, the review indicated that EC and QLF are promising methods for early detection of lesions. However, visual methods should remain the standard method for clinical early lesion assessment in dental practice.[7]

QLF findings provide information about presence/absence and regression/progression of WSLs. Moreover, QLF presents qualitative and quantitative information about location, extent and severity of demineralization. While in visual inspection, demineralization is seen as white opaque enamel areas, WSLs appear in QLF images as dark areas surrounded by bright green fluorescing tooth tissue.[9] Enamel fluorescence is directly correlated with the mineral content of enamel.[10]

Several clinical studies compared QLF and clinical inspection (supported by DP) for quantification of enamel demineralization. These studies have focused on determining the presence/absence of lesions;[11,12] detecting changes in lesion mineral content;[13] or monitoring WSL development.[14]

However, to the best of our knowledge, no publication deals with quantitatively correlating both diagnostic methods.

Due to the different acquisition method, the tooth areas of the two images are not directly superimposable. A registration procedure is required to compare the size and the extension of the demineralization. Moreover, the reliability of humans (or computer algorithms) performing image segmentation is difficult to quantify because the true area that needs to be marked, also referred to as ground truth (GT) or gold-standard,[15] is usually unknown.[16] In this context, the simultaneous truth and performance level estimation (STAPLE) algorithm has been introduced to manage the multiobserver variability and iteratively determining an estimated GT. The algorithm has been successfully used in many applications of medical image segmentation.[17,18] Recently, STAPLE has been used also in dentistry to compare the reproducibility of the detection of demineralized tooth areas in DP and QLF.[19]

The present study is a follow up to our previous study[19] and uses the same database of QLF and DP images, as well as the same GT markings of the demineralized areas, estimated using STAPLE on manual markings from experts. The previous study was just concerned with the reproducibility of the manual markings and thus, the assessment of which modality is less rater-dependent for early lesion detection.

The present study, instead, aims at a direct and quantitative comparison of the demineralized areas as detected in QLF and DP taking into account the spatial agreement of the detected areas. To this end, GT markings of the DP are compared to those of the QLF images, both estimated using STAPLE on markings by human raters. In addition, the GT markings of the DP and QLF images are compared to the demineralized areas detected by the QLF software. The demineralized areas are compared in terms of position and extension of the detected lesions, while the mineral content and the depth of the demineralization are not taken into consideration. Our key contributions are as follows

- a direct, quantitative comparison of the demineralized areas in QLF and DP using a non-rigid, multi-modal registration method[20] that aligns QLF and DP images, by transforming the outlines of the tooth areas from the one to the other modality;
- a comparison that quantifies the agreement of two detected lesions in terms of their spatial position and size;
- a quantification of the agreement via a distance-based Dice coefficient (DC), which unlike in previous work,[21] here is defined in a symmetric fashion, *i.e.* using a dilation of both sets involved in the calculation of the coefficient;
- a quantification of the amount of information detected in one image modality that is also detected in the other modality, using a new measure, which we call *coefficient of inclusion* (CI).

Our hypotheses are that (i) the clinical inspection supported by DP is more sensitive to WSL as that by QLF, disregarding whether the automatic analysis or the experts' manual assessment of QLF is applied, and (ii) detected lesions in QLF are included in those of DP.

Methods and materials

To test (i), a comparison of the size of the detected demineralized areas, manually (for QLF and DP) and with QLF-software, is performed and the position and extension of the detected lesions are compared. To test (ii), the amount of information given by one modality and contained in the other modality is quantified.

Subjects

Patients have been examined after orthodontic treatment with a fixed appliance after giving informed, written consent. The study was approved by the Ethics Committee of the RWTH Aachen University (ethical review committee number Aachen IORG0006299) prior to conducting from November 2013 to November 2014.

In total, 124 DPs and QLF images of upper and lower incisors and canines were acquired from 31 patients (16 female, 15 male), exhibiting WSLs in varying extent and severity at one or more teeth after debonding (Figure 1a, first line). Compared to our previous work,[19] here 15 DP/ QLF image pairs were excluded from the database, since the registration of these pairs was not sufficiently accurate. The average age of the subjects at the day of DP and QLF acquisition was 16.8 years (range: 11.1–35.8 years). The

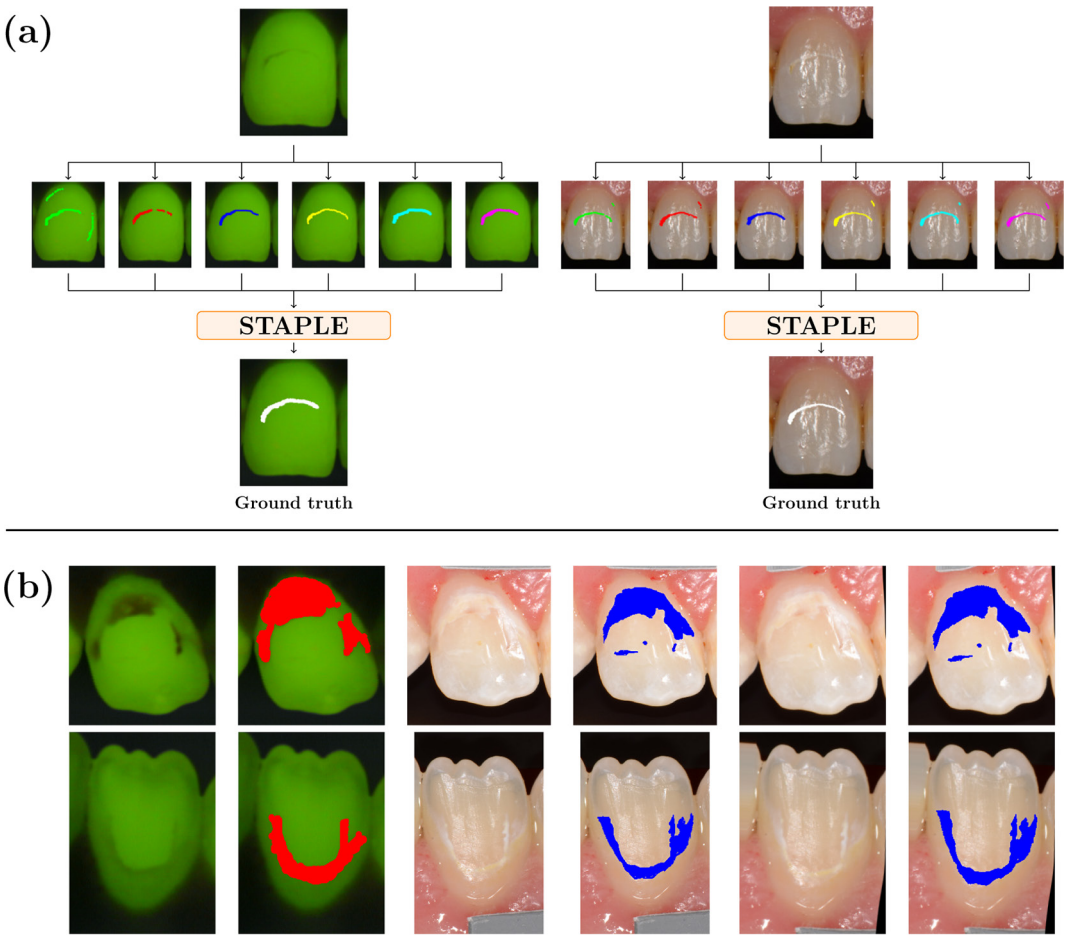

Figure 1 (a) Markings of the demineralized area from three raters in two cycles (six markings in total) and corresponding ground truth computed with STAPLE for a QLF image (left) and a digital photo (right) of an upper central incisor; (b) registration of QLF/DP pairs. From left to right: QLF images, QLF images with marked ground truth area (red), digital photographs, digital photographs with marked ground truth area (blue), deformed photographs, deformed photographs with deformed marked ground truth area (blue). DP, digital photographs; QLF, quantitative light-induced fluorescence; STAPLE, simultaneous truth and performance level estimation.

treatment time lasted on average 26 months (range: 12–45 months). The images were acquired on average 3.6 months (range: 0–31.5 months) after debonding. Weingart pliers were used to debond the fixed appliances. The adhesive was removed using carbide burs and the teeth were finally polished using polishing brush and paste.

Images acquisition

The DP were acquired using a digital single lens reflex camera (D7000, Nikon, Japan) under standardized conditions. A macro lens (Nikkor 105 mm, 1:2.8, Nikon, Japan) with a ring flash (Sigma EM-140 DG, Kawasaki, Japan) was used. Manually specified camera settings were used: shutter speed 1/250 s, aperture f/29, film speed ISO 100, and flash setting $\frac{1}{4}$ s. Before acquiring the photographs, the teeth were carefully cleaned with polishing cup and paste (Zircate Prophy, Dentsply Sirona International, York, PA) and air-dried. The camera was positioned at right angles to the teeth longitudinal axis. A better visualization of the tooth contour was achieved by positioning a black contrastor behind the teeth.

The Inspector Pro system, composed of QLF (WV-KS 152 QLF-clin, Panasonic, Japan) and Inspector Pro Software v. 2.0.0.49 (Inspector Research System BV, Netherlands), was used to record the QLF images. In this system, the surface of the tooth is exposed to a xenon-lamp (13 mW cm^{-2}) of wavelength 370 ± 80 nm and the reflected light is recorded by the camera integrated in a toothbrush handle. Furthermore, a filter that detects the emitted fluorescence light of wavelength 520 nm is used by the system. An auxiliary attachment to the toothbrush handle (ambient light shield) was used to avoid additional exposures and ensure the hygienic conditions on the patient. The tooth surface was again cleaned and dried before the measurement. The examination room was darkened. The handle was carefully positioned such that the incident light meets the tooth surface at a right angle.

However, this setting does not allow to control the perspective projection of both imaging modalities such

that the geometry (size, location, projection) is equal in the obtained images. Therefore, contours marked in either of the imaging modalities cannot be compared directly.

Markings by QLF software
The QLF images were analysed using the system's software. First, the region of interest to be analysed by the QLF software was selected by a dental expert, such that it includes all suspicious demineralized areas and the nearby healthy enamel. Afterwards, the QLF software analysed the area by calculating the percentage of fluorescence loss in the lesion area. The loss of fluorescence at a point is calculated as the difference between the intensity of the green fluorescence and the intensity of a virtual reconstruction of sound enamel at the same point.[22] When exceeding a specific threshold at a certain pixel, the fluorescence loss is indicated by a colour code ranging from dark blue (minor fluorescence loss) to yellow (severe loss) and altered intensity. The resulting markings will be denoted by M_{QLF}.

Manual markings
Three experienced raters identified the demineralized area on the examined tooth surface of DPs and QLF images by manual markings. The markings were repeated after 2 weeks. Thus, for each tooth and each modality, a total of six manual markings were collected. The images were presented to the raters in random order. Particular attention was paid to showing the investigators only the QLF images without the marking provided by the QLF system, to avoid any influence on the rater's judgement.

Ground truth estimation
The STAPLE algorithm was applied to all six markings for each modality and each tooth to estimate a GT marking for each modality (Figure 1a). The STAPLE algorithm is an expectation–maximization algorithm that uses a set of manual markings to compute a probabilistic estimate of the hidden GT.[16] Given a set of binary segmentations, STAPLE estimates iteratively an optimal-weighted combination of the input segmentations. The weights indicate the rater's performance and are updated at every iteration. Thus, for each modality, the six manual markings were given as input to the algorithm, which gives as output the estimated best WSL marking, which then is considered as GT. In this sense, STAPLE provides a rater-independent estimate of the demineralized areas based on the existing manual markings. This estimated GT is considered as gold standard of the demineralized area. The resulting markings for DP and QLF and will be denoted by GT_{DP} and GT_{QLF}, respectively.

Image and marking registration
Due to the different acquisition modalities of the QLF/DP pairs, a model-free (elastic) registration method is used to correctly superimpose DP and QLF as well as their markings,[20,23] such that the geometry of both modalities correspond as if the images have been acquired using mechanical fixation (Figure 1b). The registration method aligns the contour of the tooth area in QLF to the contour of the tooth area shown in the DP. To this end, before the registration, the QLF images and the photos were classified into tooth area and background. To achieve the best possible outcome in the alignment of QLF and DP, the tooth area in both QLF and DP were manually segmented by three raters. Then, STAPLE was used to estimate a GT contour for the tooth areas in both modalities. Finally, the registration algorithm was applied to the GT contours to get the QLF/DP alignment. In other words, we transform the contours such that they match the perspective geometry of the opposite imaging modality.

Measures of agreement
Aim of this study is to assess whether the visual inspection supported by photography is more sensitive than QLF analysis or QLF-software-based lesion detection. To this end, a comparison of the markings of the detected demineralized areas (GT and QLF-system obtained) in the two modalities is performed by evaluating the size of the markings and the agreement in extension and position of the detected lesions.

To quantify the size of the lesions detected from human markings and from the QLF-software, the size of the markings (A_M) is calculated as the ratio between the size of the marked area ($|A|$) and the size of the tooth area ($|T|$), *i.e.* $A_M = \frac{|A|}{|T|}$. $A_M = 0$ indicates that no demineralization is detected, while $A_M = 1$ indicates that the detected demineralization extends to the whole tooth.

The spatial agreement between two markings is measured by the Dice coefficient (DC that quantifies the spatial overlap of two sets A and B and is defined as the ratio between the intersection of the two sets ($A \cap B$) and their average volume.[24] $DC = 0$ and $DC = 1$ indicate that the two sets are separated and completely overlapped, respectively. The classification in "poor" ($DC \leq 0.4$), "moderate" ($0.4 < DC \leq 0.6$), "good" ($0.6 < DC \leq 0.8$), and "excellent" ($DC > 0.8$) is used.[25]

In previous work,[21,26] a tolerance zone that accounts for small inaccuracies at the boundary of the markings and improves the value of the DC was introduced. Unlike previous work, we define a distance-based DC between two sets A and B that accounts for a tolerance zone constructed in a symmetric fashion using dilations of both sets as

$$DC_\rho(A,B) = \frac{2\left|Dil_B(A,\rho) \cap Dil_A(B,\rho)\right|}{\left|Dil_B(A,\rho)\right| + \left|Dil_A(B,\rho)\right|}$$

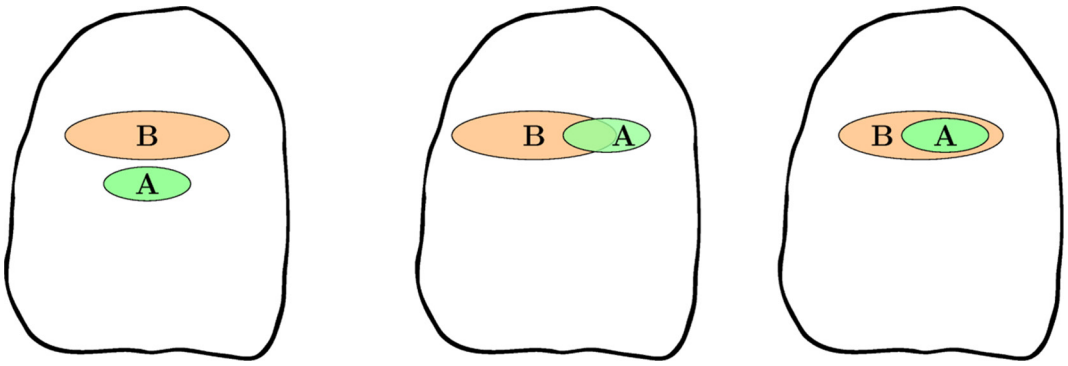

Figure 2 CI and corresponding Dice coefficient (DC_ρ) for $\rho = 0$ pixels. From left to right: $CI(A, B) = 0$ ($DC_\rho(A, B) = 0$), $CI(A, B) \approx 0.5$ ($DC_\rho(A, B) \approx 0.3$), $CI(A, B) = 1$ ($DC_\rho(A, B) \approx 0.6$). CI, coefficient of inclusion; DC, Dice coefficient.

where $Dil_B(A, \rho) = \{A \cup (Dil(A, \rho) \cap B)\}$ and $Dil(A, \rho)$ denotes the morphological dilation of A using a disc of diameter ρ.

Based on A_M and the DC, the QLF system marking is compared with the GT generated using STAPLE of the QLF and the registered DP for all 124 teeth. In addition, the DC of the GT of DP and QLF is calculated and used to quantify the agreement of the detected demineralization in the two modalities.

To test whether the QLF findings are included in the DP ones, *i.e.* the markings in the DP contain the same information (or more) provided by the QLF markings, we calculate the CI as the fraction of elements in one set that are contained in the other set as

$$CI(A, B) = \frac{|A \cap Dil_A(B, \rho)|}{|A|}$$

It is equal to 1 if $A \subseteq B$ and 0 if the two sets are disjoint (Figure 2). Again, small inaccuracies at the boundaries are accounted for by considering the dilation of the set B.

For each tooth, the CI is calculated for every possible pair of markings.

Descriptive statistics (mean and standard deviation) of A_M are presented in Table 1. In addition, means and standard deviations of the DC and the CI, each computed with a tolerance zone ($\rho = 2$ pixels) for each pair of markings, are reported in this table.

Statistically significant differences among the means of A_M were assessed using a one-way repeated measures analysis of variance (ANOVA). The test was performed using IBM SPSS 25 with significance level $\alpha = 0.05$.

Results

Examples of excellent and poor agreement for the GT from DP (GT_{DP}), the GT from QLF (GT_{QLF}), and the markings obtained by the QLF-software (M_{QLF}) are shown in Figure 3. The descriptive statistics are summarized for A_M, DC and CI in Table 1.

The one-way repeated measures ANOVA determined a statistically significant difference among the means of A_M for the three demineralization detection methods ($F(2, 246) = 141.09, p < 0.001$). *Post hoc* tests using Bonferroni correction were carried out to compare the means pairwise. The tests revealed statistically significant differences between the pairs of means (Table 2).

There is only a moderate agreement (mean DC_ρ = 0.572) between the GTs derived from QLF and DP. In particular, poor, moderate, good, and excellent

Table 1 Descriptive statistics of size of the markings, Dice coefficient and of the coefficient of inclusion

Measure	*Mean*	*SD*	*Range*
A_M			
GT_{QLF}	0.139	0.112	0.00–0.58
GT_{DP}	0.165	0.109	0.02–0.59
M_{QLF}	0.032	0.048	0.00–0.27
Dice coefficient			
$DC_\rho(GT_{QLF}, GT_{DP})$	0.572	0.225	0.00–0.94
$DC_\rho(GT_{QLF}, M_{QLF})$	0.331	0.282	0.00–0.96
$DC_\rho(GT_{DP}, M_{QLF})$	0.267	0.261	0.00–0.83
Coefficient of Inclusion			
$CI(GT_{QLF}, GT_{DP})$	0.642	0.207	0.00–1.00
$CI(GT_{DP}, GT_{QLF})$	0.537	0.274	0.00–0.98
$CI(M_{QLF}, GT_{QLF})$	0.892	0.246	0.00–1.00
$CI(GT_{QLF}, M_{QLF})$	0.245	0.243	0.00–1.00
$CI(M_{QLF}, GT_{DP})$	0.786	0.291	0.00–1.00
$CI(GT_{DP}, M_{QLF})$	0.187	0.213	0.00–0.74

A_M, size of the markings; CI, coefficient of inclusion; DC_ρ, Dice coefficient; GT_{DP}, ground truth marking for the DP; GT_{QLF}, ground truth marking for the QLF; M_{QLF}, marking from the QLF-software; GT, ground truth; QLF, quantitative light-induced fluorescence; DP, digital photographs; SD, standard deviation.

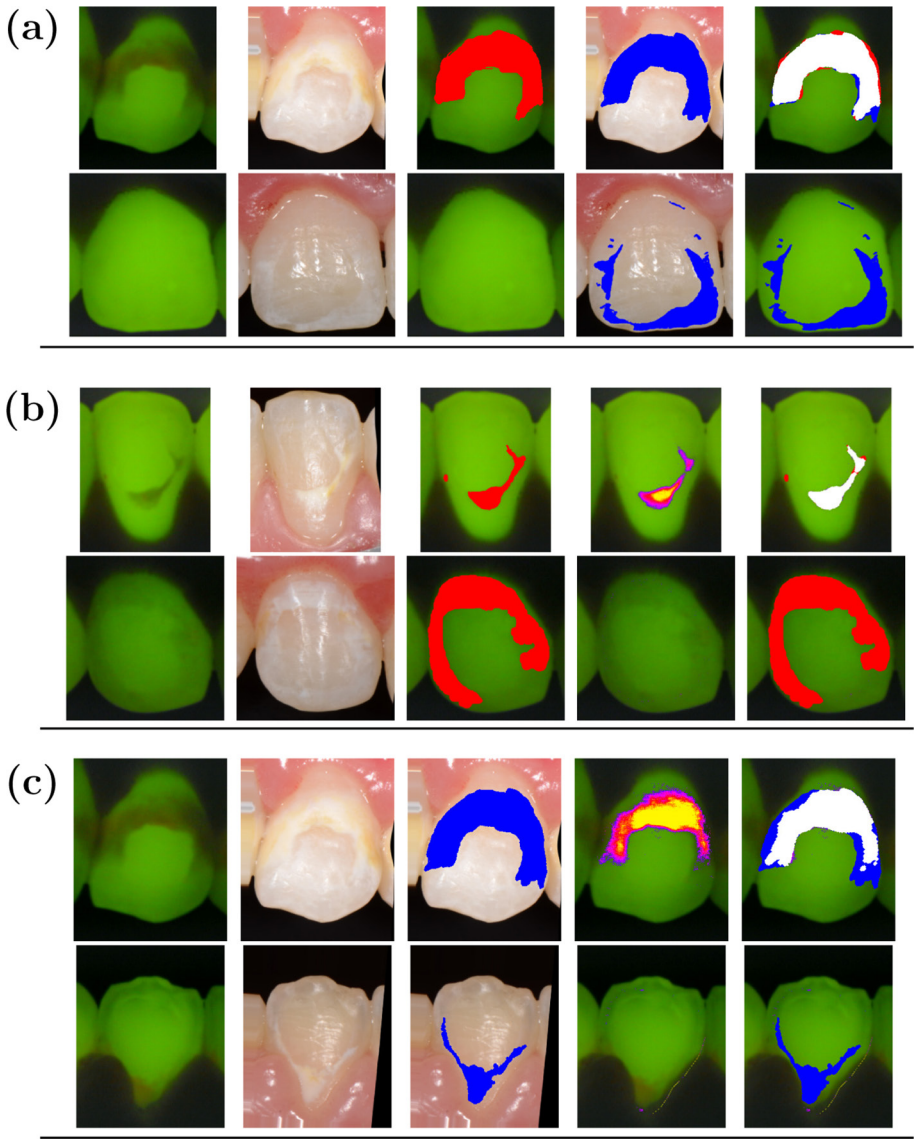

Figure 3 (a) Agreement of GT from QLF images (GT_{QLF}) and digital photographs (GT_{DP}). From top to bottom: excellent (DC_ρ (GT_{QLF}, GT_{DP}) = 0.94, CI (GT_{QLF}, GT_{DP}) = 0.932, CI (GT_{DP}, GT_{QLF}) = 0.937) and poor (DC_ρ (GT_{QLF}, GT_{DP}) = 0.16, CI (GT_{QLF}, GT_{DP}) = 0.295, CI (GT_{DP}, GT_{QLF}) = 0.089) agreement. From left to right: QLF, registered DP, QLF with GT (red), registered DP with GT (blue), superimposed agreement (white); (b) Agreement of QLF ground truth and QLF-software marking (M_{QLF}). From Top to bottom: excellent (DC_ρ (GT_{QLF}, M_{QLF}) = 0.96, CI (M_{QLF}, GT_{QLF}) = 0.982, CI (GT_{QLF}, M_{QLF}) = 0.931) and poor (DC_ρ (GT_{QLF}, M_{QLF}) = 0, CI (M_{QLF}, GT_{QLF}) = 1.0, CI (GT_{QLF}, M_{QLF}) = 0.0) agreement. From left to right: QLF, registered DP, QLF with GT (red), QLF with QLF-software marking (purple), superimposed agreement (white); (c) Agreement of DP ground truth and QLF-software marking. From Top to bottom: excellent (DC_ρ (GT_{DP}, M_{QLF}) = 0.83, CI (M_{QLF}, GT_{DP}) = 0.979, CI (GT_{DP}, M_{QLF}) = 0.713) and poor (DC_ρ (GT_{DP}, M_{QLF}) = 0, CI (M_{QLF}, GT_{DP}) = 0, CI (GT_{DP}, M_{QLF}) = 0) agreement. From left to right: QLF, registered DP, DP with GT (blue), QLF with QLF-software marking (purple), superimposed agreement (white). CI, coefficient of inclusion; DC_ρ, Dice coefficient; DP, digital photographs; GT, ground truth; QLF, quantitative light-induced fluorescence.

agreement was obtained in 24.19, 25.81, 33.06 and 16.93% of the cases, respectively (Figure 4, blue bars).

The agreement between the QLF ground truth and QLF-software markings is poor on average (mean DC_ρ = 0.331). As shown in Figure 4 (orange bars), poor and moderate agreement was obtained in 58.87 and 19.35% of cases, respectively, while good and excellent agreement in 16.93 and 4.84%, respectively.

The agreement between the DP ground truth and QLF-software markings is also poor on average (mean DC_ρ= 0.267). In particular, poor agreement was obtained in 70.96% of total cases, while moderate and good agreement was found in 14.52 and 11.29% of

Table 2 Results of the post hoc tests for the pairwise comparison of the means of A_M

Pair	*Mean*	*SD*	*p*
$GT_{DP} - GT_{QLF}$	0.027	0.089	0.003
$GT_{DP} - M_{QLF}$	0.133	0.100	0.000
$GT_{QLF} - M_{QLF}$	0.106	0.089	0.000

GT_{DP}, ground truth marking for the DP; GT_{QLF}, ground truth marking for the QLF; M_{QLF}, marking from the QLF-software; A_M, size of the markings; GT, ground truth; DP, digital photographs; QLF, quantitative light-induced fluorescence; SD, standard deviation; *p*, probability.

cases, respectively; the remaining 3.22% of all cases were classified as excellent agreement (Figure 4, green bars).

The CI values in Table 1 indicate that, on average, 64.2% of the demineralized area detected by the experts in the QLF images are also detected in the photos (mean CI $(GT_{QLF}, GT_{DP}) = 0.642$), while 53.7% of lesion area in DP is also detected in QLF (mean CI $(GT_{DP}, GT_{QLF}) = 0.537$). The amount of the lesion area shown by the QLF software that is detected on DP and QLF images by the experts are 78.6 and 89.2% respectively (mean CI $(M_{QLF}, GT_{DP}) = 0.786$ and mean CI $(M_{QLF}, GT_{QLF}) = 0.892$), while only 18.7 and 24.5% of the lesion area detected respectively in DP and QLF images by the human raters are also detected by the QLF software markings (mean CI $(GT_{DP}, M_{QLF}) = 0.187$ and mean CI $(GT_{QLF}, M_{QLF}) = 0.245$).

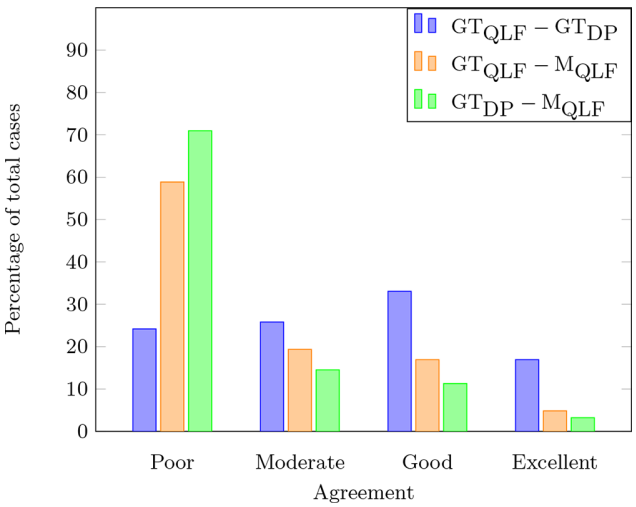

Figure 4 Histogram of the Dice coefficients (DC_ρ). The three coloured bars represent the percentage of cases where the agreement between pairs of markings was considered poor, moderate, good, and excellent. DC, Dice coefficient; GT_{DP}, ground truth marking for the DP; GT_{QLF}, ground truth marking for the QLF; M_{QLF}, marking from the QLF-software; GT, ground truth; DP, digital photographs; QLF, quantitative light-induced fluorescence.

Discussion

The results for the size of the markings A_M indicate that the detected demineralized areas in the DP is, on average, slightly larger than those marked by the human raters in the QLF. The difference in size between the GT markings in both DP and QLF, and the QLF-software markings is, on average, much larger (Table 1). Repeated measures ANOVA and *post hoc* comparison of the means determined a statistically significant difference in the means of the markings' size for all three markings (Table 2). From this, we conclude that clinical inspection supported by photographs is more sensitive to detection of demineralization than QLF in both, automatic and manual assessments. Therefore, our first hypothesis is confirmed.

In addition, the CI results for GT_{DP} and GT_{QLF}, suggest that the information about the lesion area is complementary to some degree, *i.e.* DP shows lesion parts not visible in the QLF and vice versa, while DP on average shows a slightly larger lesion area, which again confirms our first hypothesis. The standard deviation for the CI has high value, showing high variability in the inclusion of the findings by one modality into the other.

On the other hand, the agreement between the manual GT markings of QLF and DP is moderate, while the agreement between the photo GT markings and the QLF-software markings is rather poor (Table 1). The former indicates that both modalities, QLF and DP, do not correspond but detect different areas of demineralized tooth. Therefore, our second hypothesis is not confirmed. Contrarily, these results suggest that the combined use of QLF and DP can be beneficial for early lesions assessment.

Moreover, the high values of the DCs' standard deviation indicate a large variability for the agreement between the marking pairs. Therefore, automatic analysis of QLF shall be used carefully, in particular with the software's default settings. The CI for GT_{DP} and M_{QLF} indicates that the QLF software evaluates "conservatively" in the selected region of interest. Since the system seems to mark only areas with pronounced mineral loss, it likely has only a few false-positives, at the price of non-negligible areas of false-negatives. Note that that the software was used at its default parameters. Changing the threshold may change the results. Nevertheless, with the recent advances in machine learning, it may be possible to create an automatic detection that matches the human ratings.

Other works showed that diagnosis of non-cavitated lesions might be more accurate combining visual inspection and the use of electrical methods or QLF for monitoring purposes.[7–19,22] Gomez concluded for both cost and practicality considerations, visual methods should remain the standard for clinical assessment in dental practice.[7] While Boersma reported that the number of lesions found by QLF far outnumbered that found by visual examination,[12] the present study does not confirm this outcome.

Conclusions

Our study indicates that there are differences in position and/or extent of the demineralized areas detected by the two modalities. While, on average, DP shows a slightly larger lesion area, the CI indicates that findings in DP do not include all findings in QLF, *i.e.* DP shows lesion parts that are not visible in QLF and vice versa. Therefore, it can be concluded that QLF and DP are valuable methods to detect demineralized areas, and the combination of both modalities could be beneficial for the assessment of early lesions. In addition, our study indicates that the QLF-software, used at its default parameters, is very conservative in marking demineralization in the selected region of interest, as it seems to mark mostly areas with pronounced mineral loss.

Funding

The authors at AICES RWTH Aachen were funded in part by the Excellence Initiative of the German Federal and State Governments. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Informed consent

Informed consent was obtained from all individuals participating in the trial.

References

1. Enaia M, Bock N, Ruf S. White-spot lesions during multibracket appliance treatment: A challenge for clinical excellence. *Am J Orthod Dentofacial Orthop* 2011; **140**: e17–e24. doi: https://doi.org/10.1016/j.ajodo.2010.12.016
2. Gorelick L, Geiger AM, Gwinnett AJ. Incidence of white spot formation after bonding and banding. *Am J Orthod* 1982; **81**: 93–8. doi: https://doi.org/10.1016/0002-9416(82)90032-X
3. Sundararaj D, Venkatachalapathy S, Tandon A, Pereira A. Critical evaluation of incidence and prevalence of white spot lesions during fixed orthodontic appliance treatment: a meta-analysis. *J Int Soc Prev Community Dent* 2015; **5**: 433–9. doi: https://doi.org/10.4103/2231-0762.167719
4. Knösel M, Attin R, Becker K, Attin T. External bleaching effect on the color and luminosity of inactive white-spot lesions after fixed orthodontic appliances. *Angle Orthod* 2007; **77**: 646–52. doi: https://doi.org/10.2319/060106-224
5. Palamara J, Phakey PP, Rachinger WA, Orams HJ. Ultrastructure of the intact surface zone of white spot and brown spot carious lesions in human enamel. *J Oral Pathol Med* 1986; **15**: 28–35. doi: https://doi.org/10.1111/j.1600-0714.1986.tb00560.x
6. Angmar-Månsson B, al-Khateeb S, Tranaeus S. Monitoring the caries process. Optical methods for clinical diagnosis and quantification of enamel caries. *Eur J Oral Sci* 1996; **104**: 480–5. doi: https://doi.org/10.1111/j.1600-0722.1996.tb00116.x
7. Gomez J, Tellez M, Pretty IA, Ellwood RP, Ismail AI. Non-cavitated carious lesions detection methods: a systematic review. *Community Dent Oral Epidemiol* 2013; **41**: 55–66. doi: https://doi.org/10.1111/cdoe.12021
8. Angmar-Månsson B, ten Bosch JJ. Quantitative light-induced fluorescence (QLF): a method for assessment of incipient caries lesions. *Dentomaxillofac Radiol* 2001; **30**: 298–307. doi: https://doi.org/10.1038/sj.dmfr.4600644
9. de Josselin de Jong E, Sundström F, Westerling H, Tranaeus S, ten Bosch JJ, Angmar-Månsson B. A new method for in vivo quantification of changes in initial enamel caries with laser fluorescence. *Caries Res* 1995; **29**: 2–7. doi: https://doi.org/10.1159/000262032
10. Gmür R, Giertsen E, van der Veen MH, de Josselin de Jong E, ten Cate JM, Guggenheim B. In vitro quantitative light-induced fluorescence to measure changes in enamel mineralization. *Clin Oral Investig* 2006; **10**: 187–95. doi: https://doi.org/10.1007/s00784-006-0058-z
11. Heinrich-Weltzien R, Kühnisch J, Ifland S, Tranaeus S, Angmar-Månsson B, Stösser L. Detection of initial caries lesions on smooth surfaces by quantitative light-induced fluorescence and visual examination: an in vivo comparison. *Eur J Oral Sci* 2005; **113**: 494–8. doi: https://doi.org/10.1111/j.1600-0722.2005.00255.x
12. Boersma JG, van der Veen MH, Lagerweij MD, Bokhout B, Prahl-Andersen B. Caries prevalence measured with QLF after treatment with fixed orthodontic appliances: influencing factors. *Caries Res* 2005; **39**: 41–7. doi: https://doi.org/10.1159/000081655
13. Cochrane NJ, Walker GD, Manton DJ, Reynolds EC. Comparison of quantitative light-induced fluorescence, digital photography and transverse microradiography for quantification of enamel remineralization. *Aust Dent J* 2012; **57**: 271–6. doi: https://doi.org/10.1111/j.1834-7819.2012.01706.x
14. Beerens MW, Boekitwetan F, van der Veen MH, ten Cate JM. White spot lesions after orthodontic treatment assessed by clinical photographs and by quantitative light-induced fluorescence imaging; a retrospective study. *Acta Odontol Scand* 2015; **73**: 441–6. doi: https://doi.org/10.3109/00016357.2014.980846
15. Cardoso JR, Pereira LM, Iversen MD, Ramos AL. What is gold standard and what is ground truth? *Dental Press J Orthod* 2014; **19**: 27–30. doi: https://doi.org/10.1590/2176-9451.19.5.027-030.ebo
16. Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging* 2004; **23**: 903–21. doi: https://doi.org/10.1109/TMI.2004.828354
17. Dewalle-Vignion AS, Betrouni N, Baillet C, Vermandel M. Is STAPLE algorithm confident to assess segmentation methods in PET imaging? *Phys Med Biol* 2015; **60**: 9473–91. doi: https://doi.org/10.1088/0031-9155/60/24/9473
18. Gordon S, Lotenberg S, Long R, Antani S, Jeronimo J, Greenspan H. Evaluation of uterine cervix segmentations using ground truth from multiple experts. *Comput Med Imaging Graph* 2009; **33**: 205–16. doi: https://doi.org/10.1016/j.compmedimag.2008.12.002
19. Tatano R, Ehrlich EE, Berkels B, Sirazitdinova E, Deserno TM, Fritz UB. Quantitative light-induced fluorescence images and digital photographs - Reproducibility of manually marked demineralisations. *J Orofac Orthop* 2017; **78**: 137–43. doi: https://doi.org/10.1007/s00056-016-0069-6
20. Tatano R, Berkels B, Deserno TM. Mesh-to-raster region-of-interest-based nonrigid registration of multimodal images. *J Med Imaging* 2017; **4**: 044002. doi: https://doi.org/10.1117/1.JMI.4.4.044002
21. Dauguet J, Peled S, Berezovskii V, Delzescaux T, Warfield SK, Born R, et al. Comparison of fiber tracts derived from in-vivo DTI tractography with 3D histological neural tract tracer reconstruction on a macaque brain. *Neuroimage* 2007; **37**: 530–8. doi: https://doi.org/10.1016/j.neuroimage.2007.04.067
22. Gomez J. Detection and diagnosis of the early caries lesion. *BMC Oral Health* 2015; **15**(Suppl 1): S3. doi: https://doi.org/10.1186/1472-6831-15-S1-S3

23. Berkels B, Deserno TM, Ehrlich EE, Fritz UB, Sirazitdinova E, Tatano R. Curve-to-image based non-rigid registration of digital photos and quantitative light-induced fluorescence images in dentistry. In: Tolxdorff T, Deserno T, Handels H, Meinzer HP, eds. *Bildverarbeitung für die Medizin 2016, Informatik aktuell*. Berlin, Heidelberg: Springer Vieweg; 2016. pp. 80–5.
24. Dice LR. Measures of the amount of ecologic association between species. *Ecology* 1945; **26**: 297–302. doi: https://doi.org/10.2307/1932409
25. Khmelinskii A, Baiker M, Kaijzel EL, Chen J, Reiber JH, Lelieveldt BP. Articulated whole-body atlases for small animal image analysis: construction and applications. *Mol Imaging Biol* 2011; **13**: 898–910. doi: https://doi.org/10.1007/s11307-010-0386-x
26. Strumia M, Schmidt FR, Anastasopoulos C, Granziera C, Krueger G, Brox T. White matter MS-lesion segmentation using a geometric brain model. *IEEE Trans Med Imaging* 2016; **35**: 1636–46. doi: https://doi.org/10.1109/TMI.2016.2522178