Estimating the reliability of salivary gland ultrasound scoring in Sjögren's disease: the outcome of an international training workshop

M.A. Suludere[1], N.R.F. Sluijpers[1], S. Arends[2], A.J. Stel[2],
S. Jousse-Joulin[3], A. Hočevar[4], H. Bootsma[2], A. Vissink[1,5], K. Delli[1,5]

[1]Department of Oral and Maxillofacial Surgery, University Medical Center Groningen, the Netherlands; [2]Department of Rheumatology and Clinical Immunology, University Medical Center Groningen, the Netherlands; [3]Department of Rheumatology, Brest University Hospital Centre, Université de Bretagne Occidentale, University of Brest, France; [4]Department of Rheumatology, University Medical Centre Ljubljana, Slovenia; [5]Center for Dentistry and Oral Hygiene, University of Groningen, University Medical Center Groningen, the Netherlands.

Abstract

Objective

Current evidence on how training influences the reliability of salivary gland ultrasound (SGUS) image scoring is scarce, particularly in the context of Sjögren's disease (SjD). This study aimed to address this gap by evaluating the effect of a structured training workshop on inter-observer reliability in SGUS scoring among clinicians assessing patients with SjD.

Methods

25 healthcare professionals from 10 countries, with varying SGUS expertise participated. In random order, SGUS images of 20 suspected SjD patients were assessed before and after the workshop. Images included grey-scale (GS) and colour Doppler (CD) scans of the submandibular and parotid glands and were scored using the OMERACT GS and CD scoring systems. Intraclass correlation coefficients (ICC) assessed overall inter-observer reliability, and participant vs. SGUS-expert reliability (gold standard-participant agreement). Analyses were stratified by SGUS experience (none vs. ≥1 year).

Results

The inter-observer reliability ICC for the total OMERACT score was 0.68 pre-workshop vs. 0.79 post-workshop for GS, and 0.73 pre-workshop vs. 0.72 post-workshop for CD. Training significantly improved the gold standard-participant ICC GS (0.06±0.12, p=0.020), particularly for the submandibular glands, while the CD ICC showed a minor, non-significant improvement (0.03±0.09, p=0.129). Inexperienced participants (n=11) showed significant ICC improvement for the total GS OMERACT score (0.13±0.13, p=0.012), whereas experienced participants (n=14) showed a negligible change (0.01±0.09, p=0.624). No significant differences were observed for CD scoring.

Conclusion

A training workshop was associated with improvements for inter-observer reliability for GS SGUS, particularly in submandibular gland assessment and among inexperienced participants. The effects on CD scoring were minimal.

Key words

ultrasound, salivary gland, Sjögren's disease, reliability, training

Mehmet A. Suludere, MD
Niels R.F. Sluijpers, MD
Suzanne Arends, PhD
Alja J. Stel, MD, PhD
Sandrine Jousse-Joulin, MD, PhD
Alojzija Hočevar, MD, PhD
Hendrika Bootsma, MD, PhD
Arjan Vissink, DDS, MD, PhD
Konstantina Delli, DDS, Dr Med Dent, PhD

Please address correspondence to:
Mehmet Suludere
Department of Oral and
Maxillofacial Surgery,
University of Groningen,
University Medical Center Groningen,
Hanzeplein 1, HPC BB70,
Postbus 30001,
9700 RB, Groningen, The Netherlands.
E-mail: m.a.suludere@umcg.nl

Received on April 25, 2025; accepted in revised form on September 22, 2025.

Competing interests: none declared.

Introduction

Sjögren's disease (SjD) is a systemic autoimmune disease, with an estimated prevalence of 61 cases per 100.000 individuals (1). SjD is characterised by lymphocytic infiltration in the exocrine glands, especially the salivary and lacrimal glands, and experienced symptoms such as dry eyes and mouth (2-4). SjD has a negative impact on the quality of life due to, amongst others, symptoms of fatigue, dryness and pain, depression and anxiety, and a decreased physical ability, which all affect daily activities and social well-being (5, 6).

Diagnosing SjD is challenging due to its complex nature, and frequently nonspecific and variable symptoms (7). The current 2016 American College of Rheumatology - European Alliance of Associations for Rheumatology (ACR-EULAR) classification criteria (8), although intended to classify patients for studies, are commonly used as a guide to assess whether the patient has SjD or not. Studies suggest that salivary gland ultrasound (SGUS) could be a valuable tool to add to the ACR-EULAR classification criteria. Namely, SGUS has a promising diagnostic accuracy for SjD, offering a good sensitivity and specificity (9-11). Furthermore, Jousse-Joulin *et al.* reported that incorporating SGUS into classification criteria improves their sensitivity, while Le Goff *et al.* demonstrated that SGUS improved the performance of the criteria, especially in patients with atypical presentations (12, 13). In our diagnostic cohort study, the accuracy of the ACR-EULAR classification criteria remained excellent for the clinical diagnosis of SjD after incorporating the SGUS OMERACT score, providing a more balanced set of objective glandular items (14).

Ultrasound has several advantages: it is a non-invasive, non-irradiating, nonexpensive, and widely available imaging technique that can provide real time information about the glands and their morphological abnormalities (11, 15-18). SGUS is a dynamic examination, and the outcome relies on the skills and experiences of the operator, which might result in variability and challenges with consistency (17, 19). Until now, studies regarding the reliability of SGUS were performed with well-trained and experienced ultrasonographers. There is limited information on the reliability of less experienced observers and the potential effect of training on their ability to read ultrasound images.

A standardised training pathway for inexperienced health care professionals could improve SGUS reliability. Ultimately, the intention is to broaden the pool of clinicians using SGUS, thereby facilitating its wider and more consistent application. Quéré *et al.* demonstrated that videoconference training helped sonographers to interpret greyscale images (20). However, their study did not assess the training effect on the observers and did not incorporate colour Doppler SGUS, uncovered areas that merit further investigation. Therefore, we performed a study to assess whether an international training workshop to standardise ultrasound scoring in SjD could enhance the reliability of observers. The primary objective was to determine whether attending the training improved inter-observer reliability of participants. The secondary objective was to assess whether the training produced a measurable learning effect, evaluated by comparing the inter-observer reliability between participants and experienced SGUS operators.

Material and methods

Participants and workshop training

The study was conducted among healthcare professionals attending the salivary gland ultrasound pre-conference workshop at the 16th International Symposium Sjögren's Disease, on the 21st April 2024, Egmond aan Zee, the Netherlands. Each participant provided before the exercise the following information: gender, age, experience in SjD, number of years with experience in SGUS, and experience with ultrasound in general.

In the first round, participants were asked to score anonymised SGUS images from 20 patients suspected of having SjD two weeks to at latest one hour prior to attending the workshop using the OMERACT grey-scale (GS) and colour Doppler (CD) scoring systems, in a Microsoft Teams environment (21, 22). While the images were viewed in the same online environment, the display conditions like screen size, resolution, or light conditions were not standardised among the participants, reflecting real life practice. The participants received detailed written information via e-mail, explaining how to access the images, how to score them, and how to upload their scoring. Every participant filled out a standardised scoring chart.

Then all participants attended the ultrasonography workshop, as planned. The workshop was a structured, 2-hour session focusing on enhancing the skills of the participants in assessing ultrasonographic images of the major salivary glands and interpreting the findings. The workshop's content included instructions on the basics of salivary gland ultrasound, recognition of the most common ultrasonographic glandular abnormalities, practical advice on standardising image assessment, and information about the scoring systems. In the last part of the workshop, participants were shown a different set of ultrasonographic images of the major salivary glands (not included in the reliability exercise described in this article to avoid influencing the results of this study) and were asked to score them by using an online voting platform. The correct scoring of each image was then revealed and discussed through an online platform to enhance learning.

Within a timeframe of two weeks to one month following the workshop, participants completed a second-round assessment of the ultrasound images. The series of SGUS images were from the same patients as in the first round, but presented in a different, random order to avoid evaluation bias.

Patients

20 subjects suspected of having SjD with a variety of ultrasonographic characteristics were selected. All patients had visited the department of Rheumatology and Clinical Immunology, University Medical Center Groningen between July 2023 and November 2023, and underwent SGUS imaging as part of the routine diagnostic workup. Their images were selected from an existing dataset. Eight high-resolution static ultrasound images from each patient were collected, anonymised for all patient information, *i.e.* name, gender, date of examination, diagnosis and clinical characteristics. The images were organised into PowerPoint presentations. For each patient, a separate PowerPoint file was created for the pre-workshop assessment, and for the post-workshop assessment. Each PowerPoint file consisted of 8 slides: one ultrasound image per slide for GS and CD assessment: the left submandibular (LSm), the left parotid (LPar), the right submandibular (RSm), and the right parotid (RPar) glands. These PowerPoint files were subsequently uploaded to a safe Microsoft Teams environment, where the study participants could access and evaluate them.

Ultrasound equipment

The patients included in this reliability study were examined using the same ultrasonography machine (Esaote MyLabSeven, Genova, Italy), equipped with a high-resolution linear probe operating at a frequency range of 3–13 MHz. Patients were instructed to refrain from eating or drinking one hour prior to the ultrasonography appointment. The machine settings that were applied for the CD examination of the salivary glands were as follows: image depth 2.5 cm, one focus point at 1.0 cm below the surface of the skin, CD frequency up to 8.3 MHz (range 3.6–8.3 MHz) and pulse repetition frequency of 750 Hz. The submandibular glands were examined in the longitudinal plane, while the parotid glands were examined in both longitudinal and transverse planes. Both GS and CD images of the glands were collected (23).

Image scoring system

The participants of the workshop were asked to assess the set of images using the OMERACT GS and OMERACT CD scoring systems, which are both established methods for grading glandular abnormalities (21, 22).

The OMERACT grey-scale scoring system uses an ordinal scale from 0 to 3 as follows: grade 0, normal appearing salivary gland parenchyma; grade 1, minimal change: mild inhomogeneity without hypo/anechoic areas; grade 2, moderate change: moderate inhomogeneity with focal hypo/anechoic areas; and grade 3, severe change: diffuse inhomogeneity with hypo/anechoic areas occupying the entire gland surface (1). The OMERACT CD system similarly uses an ordinal scale from 0 to 3: grade 0, no vascular signals; grade 1, focal vascular signals; grade 2, diffuse vascular signals detected in <50% of the gland; and grade 3 diffuse vascular signals detected in >50% of the gland (22).

Data analysis

- Inter-observer reliability

The overall inter-observer reliability was evaluated among all participants for the pre- and post-workshop rounds. We hypothesised that inter-observer reliability would be lower in the first round (*i.e.* pre-workshop round), due to non-standardised image assessment. We further hypothesised the inter-observer reliability would increase post-workshop (in a second round), reflecting increased consistency among the participants following the standardised instructions during the workshop training.

- Gold standard-participant agreement

To evaluate the SGUS image scoring ability of the participants, a 'gold standard-participant agreement' was used. The workshop trainers, K.D. and A.S. were considered the gold standard, they had previously scored the images and demonstrated excellent inter-observer agreement based on previous studies (10). The gold standard-participant agreement was determined by calculating inter-observer reliability between each participant and the workshop trainers, K.D. or A.S., by comparing their ultrasound scores for each image. This agreement was calculated for each participant at two timepoints, *i.e.* pre- and post-workshop. The pre- and post-workshop values were compared to determine whether the scoring ability of the participants had improved after the workshop.

A sub-analysis was conducted to evaluate if the effects of the workshop training were larger for less experienced observers. The participants were divided into two groups: participants with no prior SGUS experience (0 years) and participants with prior SGUS experience (at least 1 year).

To assess the effect of training on the reliability of ultrasonographers in distinguishing whether SGUS findings are compatible with SjD or not, we performed an additional analysis using previously established OMERACT GS cut-offs for SjD that have demonstrated diagnostic value in earlier studies. SGUS was considered compatible with SjD based on the following cut-off criteria: (i) compatibility with SjD if the highest single-gland score was ≥2 across the four salivary glands (24, 25); and (ii) compatibility with SjD if the total OMERACT GS score was ≥5 (25).

Statistical analysis

Mean (±SD), median (IQR) or n (%) was used for descriptive statistics of normally distributed data, non-normally distributed data and categorical data, respectively. Graphical interpretation of histograms Q-Q plots were used to determine the distribution of the data.

Fleiss' kappa (FK) was utilised to quantify overall inter-observer reliability (26). FK values were interpreted as follows: <0.00; poor agreement, 0.00–0.20; slight agreement, 0.21–0.40; fair agreement, 0.41–0.60; moderate agreement, 0.61–0.80; good agreement, and 0.81–1.00; excellent agreement (10). Weighted Cohen's kappa (WCK) was used to evaluate the gold standard-participant agreement for each individual gland. Cohen's kappa (CK) was used to calculate the gold standard-participant agreement for the OMERACT greyscale diagnostic cut-offs. The same interpretation as for FK was applied for interpreting the WCK and CK values. To assess overall inter observer reliability and overall gold standard-participant agreement across all four glands, we calculated the sum score (range: 0–12) by adding the scores of the LSm, the LPar, the RSm, and the RPar glands. Intraclass correlation coefficients (ICCs; two-way mixed effects model, single measures, absolute agreement) were used to assess the overall inter observer reliability and overall gold standard-participant agreement on the sum score of GS and CD assessment (10).

Table I. Overall inter-observer reliability among participants pre- and post-workshop.

Gland	Pre-workshop	Post-workshop
Left Sm GS	0.23	0.28
Left Par GS	0.32	0.38
Right Sm GS	0.14	0.25
Right Par GS	0.37	0.40
Total GS **(ICC)**	0.68	0.79
Left Sm CD	0.35	0.36
Left Par CD	0.54	0.47
Right Sm CD	0.34	0.29
Right Par CD	0.59	0.45
Total CD **(ICC)**	0.73	0.72

Sm: submandibular gland; Par: parotid gland; GS: grey-scale; CD: colour Doppler.
Kappa values for individual glands are reported as weighted Cohen's kappa, while kappa values for total OMERACT scores are presented as Intraclass Correlation Coefficients (ICC).

The gold standard-participants kappa values followed a normal distribution. To assess the effects of the workshop training on the scoring ability of the participants, the pre- and post-workshop gold standard-participant agreements were compared using a paired sample t-test. Additionally, a sub-analysis was performed by stratifying participants by the level of experience into 2 groups; and the gold standard-participants agreements pre- and post-workshop were compared separately within each group using a paired sample t-test.

Statistical analysis was made using IBM SPSS Statistics 28 (SPSS, Chicago, IL, USA).

Results

Participants of the workshop

The workshop was attended by 110 individuals, and 37 of them accepted the invitation to participate in this study. Of the 37 participants that accepted the invitation, 10 did not complete both rounds, and two had incomplete assessments. This resulted in a total of 25 participants from 10 countries (Brazil, China, Denmark, Italy, Japan, the Netherlands, Norway, USA, South Korea, and Switzerland) who agreed to participate in this study and completed the both the pre- and post-workshop assessments. Participants included 13 rheumatologists, four oral medicine experts, three physician assistants, two dentists, two oral and maxillofacial surgeons, and one PhD student. The mean age of participants was 47.0±12.3 years. Nearly all participants had previous experience with diagnosing SjD patients (96%, n=24). Eleven (44%) did not have any previous SGUS experience, and 14 (56%) had at least one year of SGUS experience. Among the observers with SGUS experience, the median experience was 1.5 (1.0–7.0) years.

Patients

The median patient age was 58 (44–66) years, the majority was female (90%, n=18), and 40% (n=8) had a Hočevar score greater than 15 (16). The mean total OMERACT grey-scale score was 5.1±4.0, while the mean OMERACT CD score was 5.9±2.6. A more detailed overview of the patients' characteristics can be found in Supplementary Table S1.

Overall inter-observer reliability

After the workshop, there were improvements in inter-observer reliability in scoring GS images for all four glands. Kappa values were as follows: the LSm improved from 0.23 pre-workshop to 0.28 post-workshop, LPar from 0.32 to 0.38, RSm from 0.14 to 0.25, and RPar from 0.37 to 0.40. The total OMERACT score also showed improvement, the ICC increased from 0.68 to 0.79. For the CD images, the improvements were limited. Kappa value increased only for LSm (from 0.35 to 0.36). The kappa values decreased for the other glands, with the LPar dropped from 0.54 to 0.47, RSm from 0.34 to 0.29, RPar from 0.59 to 0.45. The total OMERACT score showed a slight change, the ICC decreasing from 0.73 to 0.72. An overview of the results is shown in Table I.

Figure 1 shows the modified Bland-Alt-

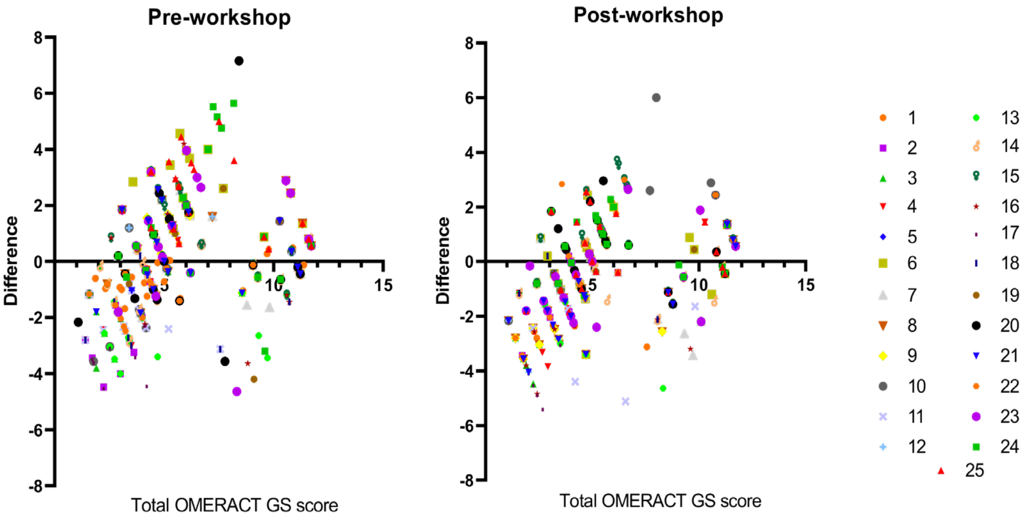

Fig. 1. Systematic differences in ultrasound scores using the OMERACT Grey-scale (GS) scoring system. For each of the 20 patients, the mean score of 25 observers was calculated. The differences between each observation and the mean were plotted against the mean. The left plot shows the pre-workshop results, while the right plot displays the post-workshop results. Each symbol in the legend represents an individual observer.

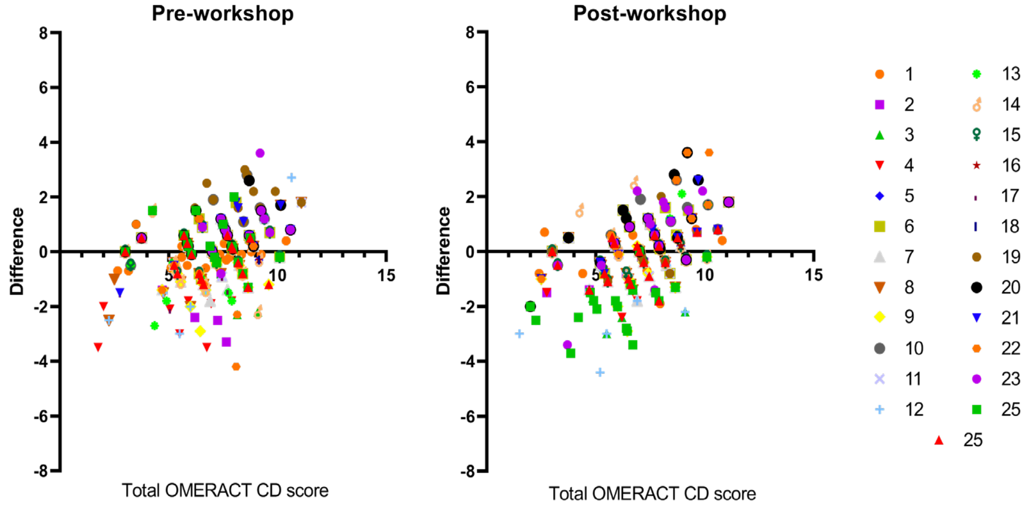

Fig. 2. Systematic differences in ultrasound scores using the OMERACT Colour Doppler (CD) scoring system. For each of the 20 patients, the mean score of 25 observers was calculated. The differences between each observation and the mean were plotted against the mean. The left plot shows the pre-workshop results, while the right plot displays the post-workshop results. Each symbol in the legend represents an individual observer.

man plots illustrating the total OMERACT GS scores and shows larger differences between the 25 observers for the first pre-workshop round compared to the second post-workshop round, particularly for the mid-range scores (3-7). The post-workshop plot shows clustering of scores around the mean and fewer outliers.

Figure 2 shows the modified Bland Altman plots illustrating the total OMERACT CD scores, showing no substantial changes in the distribution of pre- and post-workshop scores.

Effect of workshop training on gold standard - participant agreement

The training workshop improved the agreement of participants with the gold standard. For the grey scale images, the total OMERACT score showed a significant increase in ICC by 0.06±0.12 (*p*=0.020). Among the individual glands, the LSm and RSm showed a significant increase in WCK by 0.08±0.15 (*p*=0.015) and 0.06±0.14 (*p*=0.050) respectively. Although improvement was observed for the WCKs of LPar and RPar by 0.05±0.16 (*p*=0.12), and 0.03±0.11 (*p*=0.13) respectively, these changes were not statistically significant.

For the CD images, the total OMERACT score showed a non-significant improvement in ICC of 0.03±0.09 (*p*=0.129). Among the individual glands, a significant WCK increase of 0.08±0.10 (*p*<0.001) was observed for the LSm and the WCK for LPar stayed the same 0.00±0.13 (*p*=0.903). The RSm and RPar showed a non-significant decrease in WCK by -0.02±0.14 (*p*=0.583) and -0.01±0.12 (*p*=0.700), respectively. Table II presents an overview of the gold standard-participant reliability pre- and post-workshop.

Effect of the workshop on gold standard participant agreement using cut-off scores

When applying the OMERACT GS cut-offs, the gold standard-participant agreement showed some improvement after the workshop. Specifically, using the cut-off score of the OMERACT GS score ≥2 in any single gland, the mean CK increased by 0.08±0.25 (*p*=0.062), which did not reach statistical significance. In contrast, when using the cut-off of the total OMERACT GS score ≥5 across the four glands, the CK showed a small but significant improvement by 0.10±0.21 (*p*=0.031) after the workshop.

Effect of the workshop on gold standard participant agreement based on participant's experience

For grey-scale images, observers with no experience demonstrated a significant improvement in the ICC for to-

Table II. Gold standard-participant agreement: pre- and post-workshop.

	Median score (Q1-Q3) by gold standard	Median score (Q1-Q3) scored by participants before the training	Median score (Q1-Q3) as scored by participants after the training	Pre-workshop kappa	Post-workshop kappa	Changes in kappa values	Two-sided *p*
Left Sm GS	1.5 (0.75;2.0)	2.0 (1.0;3.0)	1.0 (1.0;2.0)	0.36 ± 0.14	0.44 ± 0.14	0.08 ± 0.15	0.015
Left Par GS	1.0 (0.0;2.25)	1.0 (0.0;2.0)	1.0 (0.0;2.0)	0.60 ± 0.15	0.65 ± 0.11	0.05 ± 0.16	0.124
Right Sm GS	1.0 (0.0;2.0)	2.0 (1.0;2.0)	2.0 (1.0;2.0)	0.28 ± 0.16	0.34 ± 0.14	0.06 ± 0.14	0.050
Right Par GS	1.0 (0.0;2.25)	1.0 (1.0;2.0)	1.0 (0.0;2.0)	0.59 ± 0.10	0.63 ± 0.11	0.03 ± 0.11	0.132
Total GS (ICC)*	5.1 ± 4.0	5.9 ± 3.4	5.3 ± 3.5	0.72 ± 0.14	0.79 ± 0.08	0.06 ± 0.12	0.020
Left Sm CD	2.0 (1.0;2.0)	2.0 (2.0;2.0)	2.0 (2.0;2.0)	0.30 ± 0.08	0.38 ± 0.07	0.08 ± 0.10	<0.001
Left Par CD	1.0 (1.0;2.0)	2.0 (1.0;2.0)	2.0 (1.0;2.0)	0.44 ± 0.10	0.44 ± 0.11	0.00 ± 0.13	0.903
Right Sm CD	2.0 (1.0;2.0)	2.0 (2.0;2.0)	2.0 (1.0;2.0)	0.35 ± 0.10	0.33 ± 0.10	-0.02 ± 0.14	0.583
Right Par CD	1.0 (1.0;2.0)	1.0 (1.0;2.0)	1.0 (1.0;2.0)	0.46 ± 0.10	0.46 ± 0.09	-0.01 ± 0.12	0.700
Total CD (ICC)*	5.9 ± 2.6	6.9 ± 2.1	6.5 ± 2.3	0.57 ± 0.08	0.60 ± 0.00	0.03 ± 0.09	0.129

Sm: submandibular gland; Par: parotid gland; GS: grey-scale; CD: colour Doppler.
Kappa values for individual glands are reported as weighted Cohen's kappa, while kappa values for total OMERACT scores are presented as Intraclass Correlation Coefficients (ICC). *For total scores, mean ±standard deviation (SD) is presented.

Table III. Gold standard-participant agreement: pre- and post-workshop based on experience.

	0-Year experience (n=11)						**Experience ≥1 year (n=14)**					
	Pre-workshop		Post-workshop		**Changes**		Pre-workshop		Post-workshop		**Changes**	
	Median score (Q1-Q3)	Kappa	**Median score (Q1-Q3)**	Kappa	Mean ± standard deviation	***p*-value**	**Median score (Q1-Q3)**	Kappa	**Median score (Q1-Q3)**	Kappa	Mean ± standard deviation	***p*-value**
Left Sm GS	2.0 (1.0-3.0)	0.30 ±0.15	2.0 (1.0-2.0)	0.39 ±0.14	0.09 ±0.17	**0.079**	1.0 (1.0-2.0)	0.42 ±0.10	1.0 (1.0-2.0)	0.48 ±0.13	0.06 ±0.14	**0.113**
Left Par GS	1.0 (0.0-2.0)	0.54 ±0.17	1.0 (0.0-2.0)	0.65 ±0.07	0.11 ±0.19	**0.077**	1.0 (0.0-2.0)	0.64 ±0.11	1.0 (0.0-2.0)	0.64 ±0.14	0.01 ±0.13	**0.880**
Right Sm GS	2.0 (1.0-3.0)	0.20 ±0.10	2.0 (1.0-2.0)	0.23 ±0.11	0.03 ±0.13	**0.383**	2.0 (1.0-2.0)	0.34 ±0.17	1.0 (1.0-2.0)	0.42 ±0.09	0.08 ±0.16	**0.085**
Right Par GS	1.0 (1.0-2.0)	0.54 ±0.10	1.0 (0.0-2.0)	0.61 ±0.08	0.07 ±0.12	**0.068**	1. (0.0-2.0)	0.63 ±0.09	1.0 (0.0-2.0)	0.64 ±0.13	0.01 ±0.09	**0.895**
Total GS (ICC)*	6.3 ± 3.2	0.63 ±0.14	5.7 ± 3.3	0.76 ±0.06	0.13 ±0.13	**0.012**	5.5 ± 3.5	0.80 ±0.10	5.1 ± 3.6	0.81 ±0.09	0.01 ±0.09	**0.624**
Left Sm CD	2.0 (2.0-3.0)	0.31±0.08	2.0 (2.0-3.0)	0.38 ±0.07	0.07 ±0.08	**0.010**	2.0 (2.0-2.0)	0.28 ±0.08	2.0 (2.0-2.0)	0.38 ±0.08	0.10 ±0.12	**0.010**
Left Par CD	2.0 (1.0-2.0)	0.44 ±0.11	1.0 (1.0-2.0)	0.42 ±0.09	-0.02 ±0.15	**0.642**	1.0 (1.0-2.0)	0.45 ±0.09	1.0 (1.0-2.0)	0.46 ±0.12	0.01 ±0.12	**0.735**
Right Sm CD	2.0 (2.0-3.0)	0.34 ±0.06	2.0 (2.0-3.0)	0.34 ±0.11	0.00 ±0.13	**0.959**	2.0 (1.0-2.0)	0.36 ±0.13	2.0 (1.0-2.0)	0.33 ±0.10	-0.03 ±0.15	**0.485**
Right Par CD	1.0 (1.0-2.0)	0.45 ±0.11	1.0 (1.0-2.0)	0.41 ±0.08	-0.04 ±0.14	**0.383**	1.0 (1.0-2.0)	0.47 ±0.09	1.0 (1.0-2.00	0.49 ±0.0	0.02 ±0.10	**0.607**
Total CD (ICC)*	7.2 ± 1.9	0.55 ±0.09	6.9 ± 2.5	0.57 ±0.08	0.02 ±0.11	**0.609**	6.7 ± 2.2	0.59 ±0.08	6.7 ± 2.2	0.62 ±0.07	0.01 ±0.09	**0.094**

Sm: submandibular gland; Par: parotid gland; GS: greyscale; CD: colour Doppler.
Kappa values for individual glands are reported as weighted Cohen's kappa, while kappa values for total OMERACT scores are presented as Intraclass Correlation Coefficients (ICC). *For total scores, mean ±standard deviation (SD) is presented.

tal OMERACT score, increasing by 0.13±0.13 (*p*=0.012), compared to a negligible improvement of 0.01±0.09 (*p*=0.624) among the group with experience. Similarly, improvement in the WCK values for the individual glands in the group without experience were observed, with greater increases in reliability observed in the more experienced group, however, these improvements were not significant.

For CD images, there were no notable differences in effects of the workshop between the two groups. In both groups, there was an increase in the LSm WCK, in the group without experience by 0.07±0.08 (*p*=0.010), compared to 0.10±0.12 (*p*=0.010) in the experienced group. However, for the other glands the changes in WCK or ICC values were minimal and not statistically significant in either group. Table III shows an overview of the observer-reference agreement pre- and post-workshop based on the level of experience.

Discussion

Our results suggest that training has a positive effect on overall inter-observer reliability and gold standard-participant agreement, with the most notable benefits for less experienced participants. This positive effect suggests that workshop training could help ultrasonographers to improve their scoring reliability. Our findings align with the study of Quéré *et al*. who suggested that video-conferencing training could be a tool to train sonographers (20). They reported post-training inter-observer reliability kappa values ranging from 0.23 to 0.54, measured between participants and the most experienced observer designated as the reference score. In our study the mean WCK values ranged between 0.34±0.14 and 0.65±0.11 for the GS individual glands, measured between the participants and the gold standard. Quéré *et al*. did not report an overall inter-observer reliability, instead they measured inter-observer reliability pairwise, *i.e.* between each and every participant separately, with WCK values ranging from 0.23 to 0.87. In contrast, our study found a fair agreement on the overall inter-observer reliability among all participants in GS scoring of the individual glands using FK, ranging from 0.25 to 0.40. The pooled glandular assessment demonstrated good inter-observer reliability (ICC =0.79), the FK values for the individual glands were lower. While modest, the increases in gland-level kappa may still be meaningful, especially when baseline agreement is fair to moderate. One possible explanation might be that assessing each gland in isolation might be more challenging for observers compared to assessing the overall status of the four major salivary glands. Additionally, these results could also be due to the different statistical properties of ICC and FK. ICC tolerates minor differences in the larger (0–15) pooled range score, while FK adjusts for chance in a narrower (0–3) categorical range, resulting in a stricter reliability measure (27-29). However, unlike our study, Quéré *et al.* did not report pre-training kappa values or kappa values for the individual glands, making it difficult to objectify and compare the effects of training (20).

In addition to GS scoring, we also assessed the impact of training on the scoring of CD images. To our knowledge, this is the first study investigating the impact of training on the reliability of scoring of CD images. While the improvements in GS scoring are promising, it is noteworthy that the effects of the workshop were limited on the scoring of CD images, as there were no significant improvements in the CD scoring after the workshop. Although overall inter-observer reliability was generally higher for CD scoring compared to GS scoring, the FK values for GS scoring improved across all glands post-workshop. In contrast, the FK values for CD images post-workshop stayed rather stable. The limited effect of the workshop on CD scoring may be due to the relatively simple and straightforward nature of the CD scoring system, which already had a high level of baseline agreement. The CD OMERACT scoring system is less complex, since only vascular signals are assessed, compared to GS images where multiple elements of the gland are assessed.

The discrepancy in the impact of training on GS and CD scoring may lie in the differences in complexity between the two scoring systems. Jousse-Joulin *et al.* outlined challenges in achieving reliability in GS SGUS assessments due to the wide range of abnormalities that can be evaluated, including echogenicity, homogeneity, hypoechoic areas, and calcifications. These observations emphasised the need of a standardised scoring system and expert training (9). The difference in the reliability between CD and GS is further exposed in our study: especially in the pre-workshop kappa values, with CD kappas being higher than the GS kappas. A study by Hočevar *et al.* reported high inter-observer reliability for CD SGUS in an exercise involving only expert sonographers and static images, with a Light's kappa of 0.80 across all four glands (22). In contrast Sluijpers *et al.* observed lower agreement, with overall inter-observer FK values between 0.46–0.66 in the first round and 0.35–0.50 in the second round, conducted two weeks apart (23). These results are very similar to our results, as we found FK values between 0.46 and 0.63 pre-workshop, and between 0.30 and 0.50 post-workshop. Sluijpers *et al.* highlighted the need for training as they showed that inexperienced observers had a lower intra-observer reliability compared to the most experienced observers (WCKs of respectively 0.23–0.48 *vs.* 0.72–0.81).

Applying cut-off scores for the OMERACT GS showed that training particularly improved agreement for the total OMERACT score ≥5, while the single gland ≥2 definition only showed a non-significant trend. This could suggest that training may be more effective in enhancing recognition of overall glandular morphology pattern than isolated gland changes, supporting the use of SGUS as a cumulative measure in SjD diagnosis.

In addition, our results also suggest that the effectiveness of workshop training varies per salivary gland evaluated. Specifically, the workshop particularly improved the scoring of the submandibular glands, while the effects were less pronounced for the parotid glands. This difference may be due to anatomical and structural variations between the two glands, which makes the submandibular glands more challenging to assess. Parotid glands are larger, have more homogenous structures that make abnormalities easier to detect and make consistent scoring easier compared to submandibular glands (9). In contrast, submandibular glands are anatomically more complex, smaller, and not infrequently have more heterogenous echotexture, all factors that contribute to lower reliability. These glandular specific findings provide further insight into where training may be most impactful.

One of the main strengths of our study is its design that enables comparison of the scoring reliability over time (pre and post intervention). Furthermore, inclusion of a diverse group of participants enhances the representativeness of the sample and provides valuable insights into how the training can be beneficial for less experienced sonographers. Additionally, by evaluating both GS and CD images, as well as assessing and reporting on individual glands, this study delivers a comprehensive analysis of SGUS reliability.

Among the limitations of this study is the evaluation of static images only. As Jousse-Joulin *et al.* noted, static images differ from live image acquisition, and thus the use of static images could lead to the underestimation of the variability observed in real life (31). Accordingly, the applicability of our findings to real-time SGUS in clinical settings will require further validation, ideally through studies assessing reliability during live ultrasonography. However, using static images is a common practice when assessing the reliability of SGUS, and its use during reliability exercises allows standardised conditions for the participants and reduced variability, that is potentially caused by differences in ultrasound technique (10, 20, 23, 31). Also, only a subgroup of the workshop participants, 25 out of 110 (22.7%), participated in the study, which raises the possibility of selection bias. Those who agreed to participate may have had a particular interest in SGUS, potentially having greater knowledge and motivation, which could have resulted in an overestimation of the results. Nevertheless, the diversity of the participants, representing 10 countries worldwide and various healthcare professions, and the inclusion of both inexperienced and experienced participants enhance the generalisability of the findings. Lastly, the use of a single ultrasound device may be a limiting factor for generalisability in centres with different equipment with varying technical specifications (*e.g.* resolution, sensitivity, imaging software). The goal of this study, however, was to assess the effectiveness of SGUS training using a particular model or set of specifications, which may be widely used or representative of a standard device. Additionally, any potential differences in imaging quality due to variations in equipment could be mitigated by proper training for operators and by applying imaging protocols. In summary, a training workshop was associated with improvements for inter-observer reliability for GS SGUS, particularly in submandibular gland assessment and among inexperienced participants. The effects on CD scoring were minimal.

Collaborators

Akaluck Thatayatikom, Alan Baer, Alexandre Dumusc, Ana Carolina Fragoso Motta, André Silva Franco, Annalisa Marino, Chiara Baldini, Elisabeth Maeland, Fabiola Reis de Oliveira, Gaetano La Rocca, Giovanni Fulvio, Heidi Munk, Janice Gales, Jing He, Joanita van Santen, Kyung-Ann Lee, Malin Jonsson, Maria Lucia Lemos Lopes, Minako Tomiita, Peter Olsson, Sarah Pringle, Seunghee Cha, Silvia Liefers, Vesna Risso, Virginia Fernandes Moça Trevisani, Yuebo Jin, Yuzaburo Inoue.

References

1. QIN B, WANG J, YANG Z *et al.*: Epidemiology of primary Sjögren's syndrome: a systematic review and meta-analysis. *Ann Rheum Dis* 2015; 74(11): 1983-89. https://doi.org/10.1136/annrheumdis-2014-205375
2. KREISBERG MK, TURNER J: Dental causes of referred otalgia. *Ear Nose Throat J* 1987; 66(10): 398-408.
3. KROESE FGM, BOOTSMA H: Biomarkers: new biomarker for Sjögren's syndrome--time to treat patients. *Nat Rev Rheumatol* 2013; 9(10): 570-72. https://doi.org/10.1038/nrrheum.2013.143
4. MARIETTE X, CRISWELL LA: Primary Sjögren's syndrome. *N Engl J Med* 2018; 378(10): 931-39. https://doi.org/10.1056/nejmcp1702514
5. LENDREM D, MITCHELL S, MCMEEKIN P *et al.*: Health-related utility values of patients with primary Sjögren's syndrome and its predictors. *Ann Rheum Dis* 2014; 73(7): 1362-68. https://doi.org/10.1136/annrheumdis-2012-202863
6. TROMBY FMV, CHATZIS LG, ARENDS S *et al.*: Clinical manifestations, imaging and treatment of Sjögren's disease: one year in review 2024. *Clin Exp Rheumatol* 2024; 42(12): 2322-35. https://doi.org/10.55563/clinexprheumatol/5xq3fb
7. SLUIJPERS NRF, PRINGLE S, BOOTSMA H, SPIJKERVET FKL, VISSINK A, DELLI K: Connecting salivary gland inflammation to specific symptoms in Sjögren's disease. *Expert Rev Clin Immunol* 2024; 20(10): 1169-78. https://doi.org/10.1080/1744666x.2024.2377616
8. SHIBOSKI CH, SHIBOSKI SC, SEROR R *et al.*: 2016 American College of Rheumatology/European League Against Rheumatism classification criteria for primary Sjögren's syndrome: a consensus and data-driven methodology involving three international patient cohorts. *Ann Rheum Dis* 2017; 76(1): 9-16. https://doi.org/10.1136/annrheumdis-2016-210571
9. JOUSSE-JOULIN S, MILIC V, JONSSON MV *et al.*: Is salivary gland ultrasonography a useful tool in Sjögren's syndrome? A systematic review. *Rheumatology* (Oxford) 2016; 55(5): 789-800. https://doi.org/10.1093/rheumatology/kev385
10. DELLI K, ARENDS S, VAN NIMWEGEN JF *et al.*: Ultrasound of the major salivary glands is a reliable imaging technique in patients with clinically suspected primary Sjögren's syndrome. *Ultraschall Med* 2018; 39(3): 328-33. https://doi.org/10.1055/s-0043-104631
11. RAMSUBEIK K, MOTILAL S, SANCHEZ-RAMOS L, RAMRATTAN LA, KAELEY GS, SINGH JA: Diagnostic accuracy of salivary gland ultrasound in Sjögren's syndrome: A systematic review and meta-analysis. *Ther Adv Musculoskelet Dis* 2020; 12. https://doi.org/10.1177/1759720x20973560
12. LE GOFF M, CORNEC D, JOUSSE-JOULIN S *et al.*: Comparison of 2002 AECG and 2016 ACR/EULAR classification criteria and added value of salivary gland ultrasonography in a patient cohort with suspected primary Sjögren's syndrome. *Arthritis Res Ther* 2017; 19(1): 269. https://doi.org/10.1186/s13075-017-1475-x
13. JOUSSE-JOULIN S, GATINEAU F, BALDINI C *et al.*: Weight of salivary gland ultrasonography compared to other items of the 2016 ACR/EULAR classification criteria for Primary Sjögren's syndrome. *J Intern Med* 2020; 287(2): 180-88. https://doi.org/10.1111/joim.12992
14. VAN NIMWEGEN JF, MOSSEL E, DELLI K *et al.*: Incorporation of salivary gland ultrasonography into the American College of Rheumatology/European League Against Rheumatism Criteria for Primary Sjögren's syndrome. *Arthritis Care Res* (Hoboken) 2020; 72(4): 583-90. https://doi.org/10.1002/acr.24017
15. LIANG H-D, BLOMLEY MJK: The role of ultrasound in molecular imaging. *Br J Radiol* 2003; 76 Spec No 2: S140-50. https://doi.org/10.1259/bjr/57063872
16. HOCEVAR A, AMBROZIC A, ROZMAN B, KVEDER T, TOMSIC M: Ultrasonographic changes of major salivary glands in primary Sjogren's syndrome. Diagnostic value of a novel scoring system. *Rheumatology* (Oxford) 2005; 44(6): 768-72. https://doi.org/10.1093/rheumatology/keh588
17. SAIED F, WŁODKOWSKA-KORYTKOWSKA M, MAŚLIŃSKA M *et al.*: The usefulness of ultrasound in the diagnostics of Sjögren's syndrome. *J Ultrason* 2013; 13(53): 202-11. https://doi.org/10.15557/JoU.2013.0020
18. NIETO-GONZÁLEZ JC, OVALLES-BONILLA JG, ESTRADA E *et al.*: Salivary gland ultrasound is linked to the autoimmunity profile in patients with primary Sjögren's syndrome. *J Int Med Res* 2020; 48(1). https://doi.org/10.1177/0300060518767031
19. PATIL P, DASGUPTA B: Role of diagnostic ultrasound in the assessment of musculoskeletal diseases. *Ther Adv Musculoskelet Dis* 2012; 4(5): 341-55. https://doi.org/10.1177/1759720x12442112
20. QUÉRÉ B, SARAUX A, CARVAJAL-ALEGRIA G *et al.*: Reliability exercise of ultrasound salivary glands in Sjögren's disease: an international web training initiative. *Rheumatol Ther* 2024; 11(2): 411-23. https://doi.org/10.1007/s40744-024-00645-6
21. JOUSSE-JOULIN S, D'AGOSTINO MA, NICOLAS C *et al.*: Video clip assessment of a salivary gland ultrasound scoring system in Sjögren's syndrome using consensual definitions: an OMERACT ultrasound working group reliability exercise. *Ann Rheum Dis* 2019; 78(7): 967-73. https://doi.org/10.1136/annrheumdis-2019-215024
22. HOČEVAR A, BRUYN GA, TERSLEV L *et al.*: Development of a new ultrasound scoring system to evaluate glandular inflammation in Sjögren's syndrome: an OMERACT reliability exercise. *Rheumatology* (Oxford) 2022; 61(8): 3341-50. https://doi.org/10.1093/rheumatology/keab876
23. SLUIJPERS NRF, FADHIL M, STEL AJ *et al.*: Reliability of colour Doppler ultrasonography of the major salivary glands in Sjögren's disease. *Clin Exp Rheumatol* 2024; 42(12): 2476-82. https://doi.org/10.55563/clinexprheumatol/ku5evp
24. FANA V, DOHN UM, KRABBE S, TERSLEV L: Application of the OMERACT Grey-scale Ultrasound Scoring System for salivary glands in a single-centre cohort of patients with suspected Sjögren's syndrome. *RMD Open* 2021; 7(2): 1-7. https://doi.org/10.1136/rmdopen-2020-001516
25. REBEL D, DE WOLFF L, DELLI K *et al.*: Added value of the salivary gland ultrasonography OMERACT score in the ACR/EULAR classification criteria for Sjögren's disease. *Semin Arthritis Rheum* 2024; 67: 152473. https://

doi.org/10.1016/j.semarthrit.2024.152473

26. LANDIS JR, KOCH GG: The measurement of observer agreement for categorical data. *Biometrics* 1977; 33(1): 159-74.
27. FLEISS JL: Measuring nominal scale agreement among many raters. *Psychol Bull* 1971; 76(11-1971): 378-82.
28. SHROUT PE, FLEISS JL: Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979; 86(2): 420-28. https://doi.org/10.1037//0033-2909.86.2.420
29. MCGRAW KO, WONG SP: Forming inferences about some intraclass correlation coefficients. *Psychol Methods* 1996; 1(1): 30-46.
30. JOUSSE-JOULIN S, NOWAK E, CORNEC D *et al.*: Salivary gland ultrasound abnormalities in primary Sjögren's syndrome: consensual US-SG core items definition and reliability. *RMD Open* 2017; 3(1): e000364. https://doi.org/10.1136/rmdopen-2016-000364
31. ZABOTTI A, ZANDONELLA CALLEGHER S, TULLIO A *et al.*: Salivary gland ultrasonography in Sjögren's syndrome: a European Multicenter Reliability Exercise for the HarmonicSS Project. *Front Med* (Lausanne) 2020; 7: 581248. https://doi.org/10.3389/fmed.2020.581248