Reliability of colour Doppler ultrasonography of the major salivary glands in Sjögren's disease

N.R.F. Sluijpers[1], M. Fadhil[1], A.J. Stel[2], P.U. Dijkstra[1,3,4], F.K.L. Spijkervet[1], S. Arends[2], H. Bootsma[2], A. Vissink[1], K. Delli[1]

[1]Department of Oral and Maxillofacial Surgery, [2]Department of Rheumatology and Clinical Immunology, [3]Department of Rehabilitation Medicine, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; [4]Sirindhorn School of Prosthetics and Orthotics, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand.

Abstract

Objective

To analyse intraobserver and interobserver reliability of colour Doppler (CD) ultrasonography of the major salivary glands (SGUS) in patients clinically suspected of Sjögren's disease (SjD).

Methods

One hundred consecutive outpatients visiting the University Medical Center Groningen for a diagnostic trajectory because of a suspicion of SjD were evaluated using CD ultrasonography of the submandibular and parotid salivary glands. All images were independently assessed by four observers (two experienced observers, one lesser experienced resident, one inexperienced trainee) in two sessions using the Outcome Measures in Rheumatology (OMERACT) CD scoring system (scale 0-3). A total score was calculated as the sum of the scores of the 4 glands (scale 0-12). Intra- and interobserver reliability, and reliability of live versus static scores were determined. Factors influencing variability in scores were analysed.

Results

Intraobserver weighted Cohen's kappa's of individual glands ranged from 0.23 (inexperienced observer) to 0.81 (experienced observer). Intraclass Correlation Coefficients (ICCs) for intraobserver reliability of the total CD score ranged from 0.53 (inexperienced observer) to 0.90 (experienced observer). The ICC for intraobserver reliability of live scoring compared to static images was 0.72. ICCs for interobserver reliability of the total CD score were 0.81 for session 1 and 0.71 for session 2. Patient variance was 74.1%, whereas residual variance contributed 15.5% to the total variance.

Conclusion

CD SGUS is a reliable imaging technique to visualise intraparenchymal vasculature in patients suspected of SjD, and therefore could be an asset in daily clinical practice. It requires, however, experience and prior training.

Key words

Sjögren's disease, major salivary glands, ultrasonography, colour Doppler, intraobserver reliability, interobserver reliability

Niels R.F. Sluijpers, MD
Mina Fadhil
Alja J. Stel, MD, PhD
Pieter U. Dijkstra, PT, MT, PhD
Frederik K.L. Spijkervet, DDS, PhD
Suzanne Arends, PhD
Hendrika Bootsma, MD, PhD
Arjan Vissink, DDS, MD, PhD
Konstantina Delli, DDS, Dr Med Dent, PhD

Please address correspondence to:
Niels Sluijpers
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: n.r.f.sluijpers@umcg.nl

Received on August 30, 2024; accepted in revised form on November 25, 2024.

Competing interests: none declared.

Introduction

Sjögren's disease (SjD) is a chronic systemic autoimmune disorder with a female-to-male ratio of 9:1 (1, 2). The prevalence is 61 cases per 100,000 inhabitants, with onset typically in middle age, but SjD can develop at any age, either independently or alongside with other autoimmune diseases like rheumatoid arthritis or systemic lupus erythematosus (1, 3). SjD is characterised by immune-mediated damage to the exocrine glands, particularly the lacrimal and salivary glands, leading to symptoms such as dry eyes and dry mouth (sicca symptoms) (1, 3, 4).

Until now, a gold standard to diagnose the disease, *i.e.* a single test with high sensitivity and high specificity which can successfully discriminate patients with SjD from non-SjD controls, is lacking. As a result, a variety of diagnostic tests is implemented in the diagnostic work up (5, 6).

Ultrasonography of the major salivary glands (SGUS) is increasingly used due to its non-invasive nature and cost-effectiveness for diagnosing SjD, showing good to excellent intra- and interobserver reliability (7-9). SGUS is not yet but has been suggested to be added as additional item in the 2016 American College of Rheumatology/European League Against Rheumatism (ACR/EULAR) classification criteria (10, 11). Colour Doppler (CD) ultrasonography offers potentially additional diagnostic and monitoring benefits by evaluating glandular vascularisation, which may indicate the level of inflammation (12-14). Assessment of inflammatory activity using CD ultrasonography can possibly aid in evaluating disease progression and effect of therapeutic interventions. However, before assessing its clinical benefit in SjD, reliability of the technique should be studied first.

Recently, the Outcome Measures in Rheumatology (OMERACT) ultrasound working group developed the CD OMERACT scoring system (13). A pilot study showed good to excellent reliability for vascularisation evaluation in major salivary glands, but limitations like small sample size and inclusion of only experts in SGUS to score CD images necessitate additional research (13). Therefore, the primary aim of this study was to analyse the intra- and interobserver reliability of CD ultrasonography in evaluating the major salivary glands in patients clinically suspected of SjD as well as to identify potential sources of variation in outcomes. The secondary aim was to assess the reliability of CD ultrasound when comparing live scoring of the salivary glands with static image assessments.

Materials and methods

Patients

This cross-sectional study was conducted in the Sjögren's expertise centre at the University Medical Centre Groningen (UMCG), a tertiary referral centre. Between July 2023 and November 2023, a total of 100 consecutive patients clinically suspected of SjD, *i.e.* patients with sicca symptoms, swollen salivary glands, and/or systemic disease manifestations such as fatigue and arthralgia, who visited the outpatient clinic for a diagnostic and ultrasonographic evaluation, was included in this study. A sample size of 100 patients was used to obtain acceptable confidence intervals for the reliability parameters (15). Authorisation for the utilisation of research materials was secured from the Medical Ethics Review Committee (METc) (approval no. 016/120) at the UMCG.

Procedures

Each patient underwent examination by an experienced ultrasonographer (AS or KD) using an ultrasonographic scanner (Esaote MyLabSeven, Genova, Italy), which was equipped with a high-resolution linear scanner operating at a frequency range of 3–13 MHz (8). The following baseline settings were applied for the examination of the parotid and submandibular glands: image depth 2.5 cm, one focus point at 1 cm below the skin's surface, CD frequency up to 8.3 MHz (range 3.6–8.3 MHz) and pulse repetition frequency of 750 Hz. All patients underwent scanning in a supine position with the neck slightly extended and the head slightly turned to the contralateral side. Patients were instructed not to eat or drink one hour before ultrasonographic evaluation.

Table I: CD score table of both sessions.

		Observer 1	Observer 2	Observer 3	Observer 4	Total
Session 1						
LSm	Grade 0	4	2	3	0	9 (2.3%)
	Grade 1	51	37	39	43	170 (42.5%)
	Grade 2	39	45	41	56	181 (45.3%)
	Grade 3	5	16	17	1	39 (9.8%)
LPar	Grade 0	9	4	3	4	20 (5%)
	Grade	62	60	60	65	247 (61.8%)
	Grade 2	25	31	33	30	119 (29.8%)
	Grade 3	4	5	4	1	14 (3.5%)
RSm	Grade 0	9	2	5	0	16 (4%)
	Grade 1	54	50	49	57	210 (52.5%)
	Grade 2	32	37	36	43	148 (37%)
	Grade 3	3	11	10	0	24 (6%)
RPar	Grade 0	11	4	4	5	24 (6%)
	Grade 1	61	64	53	72	250 (62.5%)
	Grade 2	24	27	39	23	113 (28.3%)
	Grade 3	4	5	4	0	13 (3.3%)
Session 2						
LSm	Grade 0	9	2	6	0	17 (4.3%)
	Grade 1	43	44	40	15	142 (35.5%)
	Grade 2	37	44	43	72	196 (49%)
	Grade 3	10	10	11	13	44 (11%)
LPar	Grade 0	10	3	4	4	21 (5.3%)
	Grade 1	61	71	46	41	219 (54.8%)
	Grade 2	21	23	46	51	141 (35.3%)
	Grade 3	7	3	4	4	18 (4.5%)
RSm	Grade 0	6	0	8	0	14 (3.5%)
	Grade 1	55	62	51	16	184 (46%)
	Grade 2	35	31	36	79	181 (45.3%)
	Grade 3	4	7	5	5	21 (5.3%)
RPar	Grade 0	12	3	2	2	19 (4.8%)
	Grade 1	55	71	54	46	226 (56.5%)
	Grade 2	27	23	41	50	141 (35.3%)
	Grade 3	6	3	3	2	14 (3.5%)

LSm: left submandibular gland; LPar: left parotid gland; RSm: right submandibular gland; RPar: right parotid gland; CD: colour Doppler.

Examination of the parotid glands consisted of both axial and coronal planes, whereas the submandibular glands were exclusively examined in the coronal plane. Both grey-scale (GS) and CD images of the glands were collected. In addition, the glands were immediately scored using the Hocevar, OMERACT GS, and OMERACT CD scoring systems (13, 16, 17).

The high-resolution images from each patient were randomised and anonymously processed in two rounds in a PowerPoint presentation. Before a CD ultrasound image of a gland, a corresponding GS ultrasound image was added as an anatomical reference as well as to take vascular signal overprojection from extra-parenchymal vasculature around the salivary glands into consideration. Per patient a PowerPoint was created including 8 images respectively, *i.e.* one showing the left submandibular (LSm) GS ultrasound image, one showing the LSm CD ultrasound image, one showing the left parotid (LPar) coronal GS ultrasound image, one showing the LPar coronal CD ultrasound image, one showing the right submandibular (RSm) GS ultrasound image, one showing the RSm CD ultrasound image, one showing the right parotid (RPar) coronal GS ultrasound image and one showing the RPar coronal CD ultrasound image.

Images were scored by four observers: two experienced observers (observer 1 and observer 2), one lesser experienced resident (observer 3) and one inexperienced trainee (observer 4). Observers received written instructions on how to score. Prior to scoring, images from 10 patients diagnosed with SjD, not included in this reliability study, were scored in a calibration session to train the observers in consistent scoring of the ultrasonographic images. After that, all images of the 100 consecutive patients suspect of SjD were scored independently by the four observers in 2 sessions (2 weeks apart) to determine intra- and interobserver reliability.

Ultrasonographic assessments of colour Doppler images

For scoring purposes, the CD OMERACT scoring system was used (13). This scoring system ranges from 0 to 3, where: 0; no visible vascular signals in the glandular parenchyma, 1; focal dispersed vascular signals in the glandular parenchyma, 2; diffuse vascular signals detected in less than 50% of the gland, and 3; diffuse vascular signals in more than 50% of the glandular parenchyma. In addition, in session 2, the presence of extra-parenchymal vasculature seen with CD was scored as an additional variable. A score of 0 represented absence, while a score of 1 represented the presence of extra-parenchymal vasculature.

Data analysis

Intraobserver reliability was assessed by comparing the CD ultrasound scores obtained during the first and second session of each individual observer. In addition, the reliability of live scoring, as originally done during the visit of the patient, compared to scoring of static images was examined by comparing the live CD ultrasound scores with the CD ultrasound scores from the same experienced observer obtained in session 1. Interobserver reliability was assessed by comparing the CD ultrasound scores across different observers for both sessions. Interobserver reli-

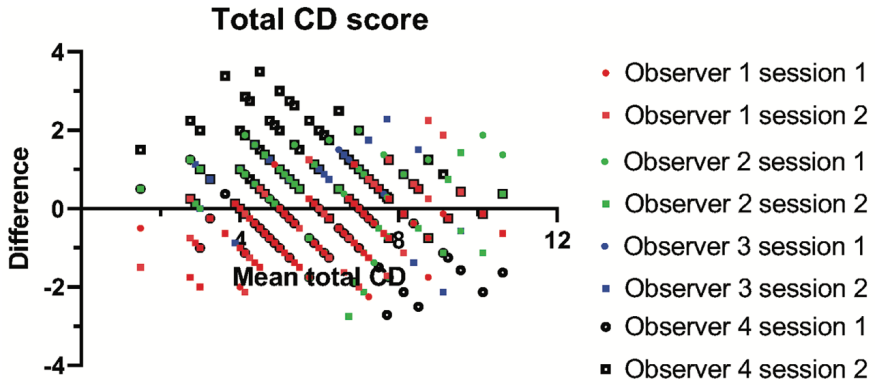

Fig. 1. Differences in total colour Doppler score.
For each patient, the mean of the 8 observations (4 observers, 2 sessions) and the difference of these 8 observations from the mean were calculated and plotted against each other. CD: colour Doppler.

ability was evaluated among all possible observer pairs. Weighted Cohen's Kappa (WCK) was used to evaluate the intraobserver reliability, the reliability of live vs. static scoring, and interobserver reliability for each individual gland. WCK 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 (8, 18). Overall interobserver reliability was determined by comparing the CD ultrasound scores of each gland among all observers. Fleiss kappa (FK) was utilised to quantify overall interobserver reliability when scoring static images. The same interpretation as WCK was applied for interpreting the FK values (8, 18). Lastly, the intra- and interobserver reliability was assessed by comparing the sum scores of the CD assessment for all four glands across all four observers. The sum score was calculated by adding the CD scores of the LSm gland, the LPar, the RSm gland, and the RPar gland (range: 0–12). Intraclass correlation coefficients (ICCs; two-way mixed effects model, single measures, absolute agreement) were used to assess the intra- and interobserver reliability of observers on this sum score of CD assessment (8). ICCs >0.70 were considered as acceptable, with ICCs >0.80 as good agreement between observers (15). Variance components analysis, utilising type III ANOVA, was applied to analyse sources of variation in ultrasonographic scores. In this analysis 'observer', 'patient', and 'session' were considered random factors. In addition, their 2-way interactions were examined. Negative variance components were set to zero. Error variation was calculated by subtracting patient variation from the sum of all sources of variation. The proportion of factors' contributions to both the total variation and error variation was expressed as a percentage. For each patient, the mean of the 8 observations (4 observers, 2 sessions) and the difference of each observation from the mean was calculated and plotted. Statistical analysis will be conducted using IBM SPSS Statistics 28 (SPSS, Chicago, IL, USA).

Results

Patient characteristics

Median age of the 100 included patients clinically suspected of SjD was 56 years (interquartile range (IQR) 44;65) (Supplementary Table S1). In total, 87% of the patients were female. Median total Hocevar score was 20 (IQR 8;28), and median total OMERACT GS score was 7 (IQR 2;10).

Colour Doppler scores

In both sessions grade 1 and grade 2 were scored most frequently (Table I). Submandibular glands were generally assigned higher scores compared to parotid glands. In addition, scores tended to be higher in session 2 compared to session 1. Overall, experienced observer 1 tended to score lower than the mean, where inexperienced observer 4 scored higher than the mean, especially in session 2 (Fig. 1).

Intraobserver reliability

Observer 1 exhibited the highest intraobserver reliability, with good to excellent agreement for all 4 glands, WCK values ranged from 0.72 to 0.81 (Table II). Observer 4 showed the lowest intraobserver reliability, with WCK values ranging from 0.23 to 0.58. The highest WCK values were observed for the parotid glands. Observer 1 and observer 2 had the highest WCK values for RPar, while observers 3 and 4 had the highest WCK values for LPar. ICCs as a measure for total CD score intraobserver reliability ranged from 0.53 to 0.90. Overall experienced observer 1 demonstrated the highest ICC,

Table II. Intraobserver reliability.

Observer	LSm WCK (95%CI)	LPar WCK (95%CI)	RSm WCK (95%CI)	RPar WCK (95%CI)	Total CD score ICC (95% CI)
Observer 1	0.76 (0.66;0.85)	0.72 (0.61;0.84)	0.76 (0.65;0.87)	0.81 (0.71;0.90)	0.90 (0.85;0.93)
Observer 2	0.70 (0.58;0.81)	0.67 (0.54;0.80)	0.59 (0.46;0.71)	0.73 (0.60;0.86)	0.83 (0.72;0.89)
Observer 3	0.67 (0.59;0.79)	0.75 (0.64;0.86)	0.70 (0.58;0.81)	0.74 (0.62;0.86)	0.89 (0.84;0.92)
Observer 4	0.32 (0.19;0.45)	0.58 (0.45;0.71)	0.23 (0.12;0.34)	0.43 (0.30;0.57)	0.53 (-0.08;0.80)
Live observation *vs*. static image session 1	0.43 (0.23;0.63)	0.56 (0.37;0.75)	0.51 (0.31;0.70)	0.60 (0.42;0.78)	0.72 (0.54;0.84)

Scoring session 1 *vs*. scoring session 2.
WCKs: weighted Cohen's kappa's; ICC: intraclass correlation coefficients; CI: confidence intervals; LSm: left submandibular gland; LPar: left parotid gland; RSm: right submandibular gland; RPar: right parotid gland; CD: colour Doppler.

Table III. Interobserver reliability of session 1 and 2.

	LSm WCK (95%CI)	LPar WCK (95%CI)	RSm WCK (95%CI)	RPar WCK (95%CI)	Total CD score ICC (95%CI)	LSm Vasc CK	LPar Vasc CK	RSm Vasc CK	RPar Vasc CK
Session 1									
Overall	*0.46 (0.40;0.53)	*0.66 (0.59;0.72)	*0.52 (0.46;0.59)	*0.63 (0.57;0.69)	0.81 (0.72;0.87)	-	-	-	-
Observer 1 *vs.* observer 2	0.55 (0.43;0.67)	0.62 (0.49;0.74)	0.58 (0.47;0.70)	0.75 (0.64;0.87)	0.81 (0.47;0.91)	-	-	-	-
Observer 1 *vs.* observer 3	0.63 (0.51;0.74)	0.69 (0.56;0.81)	0.72 (0.61;0.83)	0.68 (0.56;0.80)	0.86 (0.40;0.94)	-	-	-	-
Observer 1 *vs.* observer 4	0.42 (0.28;0.55)	0.66 (0.53;0.79)	0.54 (0.42;0.67)	0.63 (0.50;0.76)	0.76 (0.66;0.84)	-	-	-	-
Observer 3 *vs.* observer 2	0.72 (0.61;0.82)	0.73 (0.60;0.85)	0.60 (0.48;0.73)	0.78 (0.67;0.89)	0.89 (0.83;0.92)	-	-	-	-
Observer 2 *vs.* observer 4	0.45 (0.33;0.57)	0.78 (0.69;0.89)	0.46 (0.33;0.59)	0.68 (0.57;0.80)	0.75 (0.59;0.84)	-	-	-	-
Observer 3 *vs.* observer 4	0.51 (0.40;0.62)	0.75 (0.64;0.87)	0.57 (0.45;0.68)	0.54 (0.41;0.67)	0.74 (0.58;0.83)	-	-	-	-
Session 2									
Overall	*0.35 (0.29;0.40)	*0.49 (0.42;0.55)	*0.30 (0.24;0.37)	*0.50 (0.44;0.56)	0.71 (0.53;0.81)	*0.27 (0.19;0.35)	*0.29 (0.21;0.37)	*0.38 (0.30;0.46)	*0.24 (0.16;0.32)
Observer 1 *vs.* observer 2	0.62 (0.51;0.73)	0.59 (0.46;0.72)	0.59 (0.46;0.73)	0.61 (0.48;0.74)	0.82 (0.75;0.88)	0.11	0.16	0.31	0.13
Observer 1 *vs.* observer 3	0.68 (0.58;0.79)	0.59 (0.46;0.71)	0.70 (0.59;0.81)	0.54 (0.41;0.67)	0.81 (0.70;0.88)	0.14	0.12	0.37	0.18
Observer 1 *vs.* observer 4	0.27 (0.16;0.38)	0.51 (0.38;0.63)	0.20 (0.10;0.29)	0.54 (0.42;0.66)	0.58 (-0.003;0.81)	0.14	0.23	0.35	0.26
Observer 2 *vs.* observer 3	0.52 (0.39;0.64)	0.50 (0.36;0.63)	0.51 (0.38;0.64)	0.53 (0.38;0.69)	0.76 (0.66;0.84)	0.47	0.57	0.60	0.39
Observer 2 *vs.* observer 4	0.37 (0.23;0.50)	0.47 (0.34;0.60)	0.22 (0.12;0.33)	0.54 (0.40;0.68)	0.56 (0.01;0.79)	0.34	0.48	0.29	0.36
Observer 3 *vs.* observer 4	0.26 (0.14;0.38)	0.71 (0.59;0.83)	0.23 (0.13;0.33)	0.64 (0.51;0.78)	0.67 (0.23;0.84)	0.33	0.50	0.35	0.47

*Fleiss' kappa.
WCK: weighted Cohen's kappa; CK: Cohen's kappa; ICC: intraclass correlation coefficients; CI: confidence intervals; LSm: left submandibular gland; LPar: left parotid gland; RSm: right submandibular gland; RPar: right parotid gland; CD: colour Doppler; Vasc: extra-parenchymal vasculature.

followed by the lesser experienced observer 3, experienced observer 2 and lastly inexperienced observer 4.

Live vs. static scores

WCKs of the 4 individual glands ranged from 0.43 to 0.60. Intraobserver reliability of live scoring compared to delayed scoring on static images by the same experienced observer expressed in overall ICC was 0.72 (Table II).

Interobserver reliability session 1

LPar showed the highest overall interobserver reliability in the first session, with an FK of 0.66, whereas LSm demonstrated the lowest overall interobserver reliability, with an FK of 0.46 (Table III). In pairwise comparison, RPar demonstrated the highest overall interobserver reliability, with good agreement for comparisons between observer 1 and observer 2 (WCK=0.75), observer 1 and observer 3 (WCK=0.68), and observer 2 and observer 3 (WCK=0.78). The lowest overall interobserver reliability across all possible pairs, apart from the comparison between observer 3 and observer 4 was found for LSm. Interobserver reliability of the total CD score between all the observers was good to excellent with an ICC of 0.81 for the first session. Comparisons between observer 1 and observer 2, observer 1 and observer 3, and observer 2 and observer 3 revealed excellent interobserver reliability for the total CD score, with ICC values of 0.81, 0.86, and 0.89, respectively.

Interobserver reliability session 2

On average, the FK, WCK and ICC values from session 1 were higher than those observed in session 2 (Table III).

Pairwise comparisons including observer 4 tended to show lower WCKs than pairwise comparisons between the other 3 observers. RPar exhibited the highest overall interobserver reliability in the second session, with an FK of 0.50, whereas RSm showed the lowest overall interobserver reliability, with a FK of 0.30. Good to excellent interobserver reliability for the total CD score was observed with ICCs of 0.76 to 0.82 between observer 1, observer 2 and observer 3, respectively.

Overall agreement on the presence of extra-parenchymal vasculature was fair with FKs between 0.24 and 0.38 (Table III). Prevalence of presence and absence of extra-parenchymal vasculature is presented in Supplementary Table S2.

Sources of variation

Patient variance was 74.1%. The variance that could not be attributed to one of the factors or their interaction (residual variance) contributed 15.5% to the total variance. The interaction between observer and session made the largest contribution to error variance, followed by the interaction between patient and session. Main factors observer and session, and the interaction between patient and observer did not contribute to the error variance (Table IV).

Table IV. Impact of components influencing variation in ultrasonographic scores for colour Doppler ultrasonography of the major salivary glands.

Variance estimates Sources of variation	Estimate	% of total variation	% of error variation
Patients	3.183	74.1	-
Observer	0 (-0.035)	0	0
Session	0 (-0.016)	0	0
Patient x observer	0 (-0.047)	0	0
Patient x session	0.112	2.6	10.0
Session x observer	0.336	7.8	30.2
Residual variance	0.666	15.5	59.8

Dependent variable: total colour Doppler score. Method: ANOVA, Type III, Sum of Squares. (Negative variance components) were set to zero.

Discussion

Our study showed that both intra- and interobserver reliability of CD ultrasonography of the major salivary glands using the OMERACT CD scoring system is good, although training is needed to build up experience. This need is reflected in WCKs of 0.23–0.48 for the inexperienced observer *versus* WCKs of 0.72–0.81 for the most experienced observer. Our results are in line with the findings from Hocevar *et al*. that overall intra- and interobserver reliability of CD SGUS is good. (13). Hocevar *et al*. reported slightly higher kappa's, *i.e*. an intraobserver reliability with a Cohen's kappa of 0.84 for all 4 glands combined and an interobserver reliability with a Light's kappa of 0.70 for all 4 glands combined (13). It should be noted that the sample size in their study was 11 times smaller (n=9) than in our study. Furthermore, the 9 observers in that study were all experts in ultrasonographic evaluation of the major salivary glands in SjD patients, which was reflected in higher interobserver reliability.

In our study, we reported lower WCKs and ICCs for both the intra- and interobserver reliability for observer 4. Observer 4 was the least experienced observer and had no experience with ultrasonographic assessment of the major salivary glands (both in B-mode as well as CD) prior to this study. Although, a consensus meeting had taken place, our results show that more experience and training is required to consistently score CD ultrasound signals. Additionally, assessment of the submandibular glands seems more difficult, as reflected by overall lower kappa's for all observers. A similar observation was described in the study by Hocevar *et al*. (13). The lower kappa's for submandibular glands might be due to the presence of the facial artery and vein running through the submandibular gland which can potentially confuse observers (19). Training observers in recognising these vessels could improve CD reliability.

As a matter of fact, agreement on the presence of extra-parenchymal vasculature was only fair with lower kappa's in comparison to kappa's of the CD scores. This demonstrates that it is difficult to reliably recognise normal extra-parenchymal vasculature and additional training should be considered.

When comparing live scoring to static images, lower WCKs and ICCs were observed. Either the interpretation during live scoring is different from static scoring or that live scoring is based on other images, which were not included in the static reliability assessment. The discrepancy in scoring can possibly be attributed to the dynamic nature of performing ultrasonographic evaluations. Furthermore, during examination it is easier to distinguish extra-parenchymal CD signals, which can influence scoring as well. Lastly, blood flow is a dynamic process and is affected by a multitude of factors, *i.e*. movement, swallowing. This can potentially result in a change of CD signal during examination (19, 20).

The largest source of variation was patient variance, indicating that most variation in scoring is explained by actual differences in vascularisation between patients. These differences originate from anatomical variation and variation in signal intensity. While, to our knowledge, no previous data has been published about sources of variation in CD SGUS, this result is similar to the sources of variation analysis of GS SGUS by Delli *et al*. (8). They showed that variance in scoring GS SGUS could be mainly attributed to interaction of observer and patient. Therefore, they hypothesised that when monitoring patients over time, the observed change might not be only attributed to the progression of the disease or effect of medication, but also to the differences in scoring between the different observers. Therefore, they suggested that patients should ideally be scored by the same ultrasonographer. Likewise, the present study showed that variance in scoring CD SGUS could be mainly attributed to interaction of observer and session. Therefore, we

hypothesise here that when monitoring patients over time, the observed change might not be only attributed to the progression of the disease or effect of medication, but also to the differences in scoring in different sessions. As a result, we suggest that for longitudinal assessment of CD images, all images are collected and scored once in the same session by the same trained observer to ensure consistency.

A major strength of our study is the relatively large number of consecutive patients clinically suspected of SjD who visited the Sjögren's expertise centre at the UMCG, reflecting a representative patient population. Another strength is the use of the OMERACT CD scoring system, as this scoring system was specifically developed by a group of experts for assessing vascular signals in the major salivary glands of SjD patients (13). The results of the study should, however, be interpreted with caution since they were primarily based on the analysis of static images, instead of live ones, although this is a common approach in similar studies.

Our study focused on the reliability of CD SGUS. With these data no statements can be made, yet, regarding CD SGUS clinical application. A future study investigating CD signal and its association with clinical, serological and histological parameters in SjD patients is eagerly anticipated.

Conclusion

CD ultrasonography of the major salivary glands is a reliable imaging technique to visualise intraparenchymal vasculature in patients clinically suspected of SjD. In addition, our study showed that the ultrasonographer's experience plays a major role in consistently scoring images. Ultrasonographers should be trained to accurately identify normal extraparenchymal vasculature.

References

1. NEGRINI S, EMMI G, GRECO M *et al*.: Sjögren's syndrome: a systemic autoimmune disease. *Clin Exp Med* 2022; 22(1): 9-25. https://doi.org/10.1007/s10238-021-00728-6
2. MASLINSKA M, PRZYGODZKA M, KWIATKOWSKA B, SIKORSKA-SIUDEK K: Sjögren's syndrome: still not fully understood disease. *Rheumatol Int* 2015; 35(2): 233-41. https://doi.org/10.1007/s00296-014-3072-5
3. 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): 1983-9. https://doi.org/10.1136/annrheumdis-2014-205375
4. MANFRE V, CAFARO G, RICCUCCI I *et al*.: One year in review 2020: comorbidities, diagnosis and treatment of primary Sjögren's syndrome. *Clin Exp Rheumatol* 2020; 38 (Suppl. 126): S10-22.
5. BRITO-ZERON P, THEANDER E, BALDINI C *et al*.: Eular Sjögren syndrome Task Force: Early diagnosis of primary Sjögren's syndrome: EULAR-SS task force clinical recommendations. *Expert Rev Clin Immunol* 2016; 12(2): 137-56. https://doi.org/10.1586/1744666X.2016.1109449. Erratum in: *Expert Rev Clin Immunol* 2017; 13(5): 507. https://doi.org/10.1080/1744666x.2017.1309771
6. 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
7. DELLI K, DIJKSTRA PU, STEL AJ, BOOTSMA H, VISSINK A, SPIJKERVET FK: Diagnostic properties of ultrasound of major salivary glands in Sjögren's syndrome: a meta-analysis. *Oral Dis* 2015; 21(6): 792-800. https://doi.org/10.1111/odi.12349
8. 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
9. JOUSSE-JOULIN S, MILIC V, JONSSON MV *et al*.: US-pSS Study Group: 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. 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
11. 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
12. CAROTTI M, SALAFFI F, DI CARLO M, BARILE A, GIOVAGNONI A: Diagnostic value of major salivary gland ultrasonography in primary Sjögren's syndrome: the role of grey-scale and colour/power Doppler sonography. *Gland Surg* 2019; 8 (Suppl. 3): S159-S167. https://doi.org/10.21037/gs.2019.05.03
13. HOCEVAR 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
14. LEE KA, LEE SH, KIM HR: Diagnostic and predictive evaluation using salivary gland ultrasonography in primary Sjögren's syndrome. *Clin Exp Rheumatol* 2018; 36 (Suppl. 112): S165-72.
15. DE VET HCW, TERWEE CB, MOKKINK LB, KNOL DL: Reliability. In: Measurement in medicine: a practical guide. Practical guides to biostatistics and epidemiology. Cambridge University Press; 2011: 96-149. https://doi.org/10.1017/CBO9780511996214.006
16. FINZEL S, JOUSSE-JOULIN S, COSTANTINO F *et al*.: Patient-based reliability of the Outcome Measures in Rheumatology (OMERACT) ultrasound scoring system for salivary gland assessment in patients with Sjögren's syndrome. *Rheumatology* (Oxford) 2021; 60(5): 2169-76. https://doi.org/10.1093/rheumatology/keaa471
17. 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
18. LANDIS JR, KOCH GG: The measurement of observer agreement for categorical data. *Biometrics* 1977; 33(1): 159-74.
19. MARTINOLI C, DERCHI LE, SOLBIATI L, RIZZATTO G, SILVESTRI E, GIANNONI M: Color Doppler sonography of salivary glands. *AJR Am J Roentgenol* 1994; 163(4): 933-41. https://doi.org/10.2214/ajr.163.4.8092039
20. CAROTTI M, SALAFFI F, MANGANELLI P, ARGALIA G: Ultrasonography and colour doppler sonography of salivary glands in primary Sjögren's syndrome. *Clin Rheumatol* 2001; 20(3): 213-19. https://doi.org/10.1007/s100670170068