

# Multiobserver Reliability of Ultrasound Assessment of Salivary Glands in Patients with Established Primary Sjögren Syndrome

Nemanja Damjanov, Vera Milic, Juan Carlos Nieto-González, Iustina Janta, Lina Martínez-Estupiñan, Belén Serrano, Carmen Mata, María Montoro, Denisa Stanciu, Jelena Marinković-Erić, Francisco Javier López-Longo, Luis Carreño, and Esperanza Naredo

**ABSTRACT. Objective.** To evaluate the multiobserver reliability of salivary gland ultrasonography (SGUS) for scoring greyscale (GS) parenchymal inhomogeneity and parenchymal color Doppler (CD) signal in patients with established primary Sjögren syndrome (pSS).

**Methods.** The study comprised 2 multiobserver reliability assessments in patients with pSS in 2 European centers. The first reliability exercise was performed on 24 patients with pSS and 8 controls who were independently evaluated with GS and CD US by 5 observers at the Institute of Rheumatology, Belgrade, Serbia. The second reliability exercise was carried out on 10 patients with pSS who were independently assessed with GS and CD US by 8 observers at the Hospital G.U. Gregorio Marañón, Madrid, Spain. SGUS parenchymal inhomogeneity and parenchymal CD signal were semiquantitatively scored using a 4-grade scoring system. The multiobserver agreement was calculated by the overall agreement and Light's  $\kappa$  statistics.

**Results.** A total of 640 SGUS examinations were performed in the first reliability exercise and a total of 320 examinations in the second reliability exercise. Multiobserver reliability was good ( $\kappa = 0.71$ – $0.79$ ) to excellent ( $\kappa = 0.81$ – $0.82$ ) for GS parenchymal inhomogeneity in both exercises. There was a moderate ( $\kappa = 0.53$ – $0.58$ ) to good ( $\kappa = 0.70$ ) multiobserver reliability for parenchymal CD signal in the first exercise. However, there was no agreement or only a fair agreement ( $\kappa = 0.03$ – $0.29$ ) for parenchymal CD signal in the second exercise.

**Conclusion.** US may be a reliable technique in the multiobserver scoring of GS parenchymal inhomogeneity of major SG in patients with established pSS. CD scoring of SG needs further standardization to be used in multicenter studies. (First Release August 15 2016; *J Rheumatol* 2016;43:1858–63; doi:10.3899/jrheum.151220)

## Key Indexing Terms:

ULTRASONOGRAPHY    DOPPLER    SALIVARY GLANDS    SJÖGREN SYNDROME

Recent advances in medical imaging technology have led to more frequent use of diagnostic tools [ultrasound (US), magnetic resonance imaging (MRI), and computed tomography] in the investigation of the major salivary glands (SG), in patients with primary Sjögren syndrome (pSS). US may represent the imaging modality of choice because of its noninvasiveness, wide availability, high resolution, and low cost.

Previous clinical studies confirmed the usefulness of

SGUS in the evaluation of patients with pSS<sup>1,2,3</sup>. SGUS improved the diagnostic performance of both the American College of Rheumatology (ACR) and the American-European collaborative group (AECG) classification criteria for pSS<sup>4,5</sup>. Inclusion of SGUS information to the AECG criteria resulted in sensitivity increasing from 77.9% to 87% with specificity remaining almost unchanged<sup>4</sup>. Moreover, the addition of SGUS data to the 2012 ACR criteria increased their sensi-

*From the Rheumatology Department, Institute of Rheumatology and the Institute of Social Medicine and Informatics, University of Belgrade, Medical School, Belgrade, Serbia; Rheumatology Department, Hospital General Universitario Gregorio Marañón, Madrid, Spain.*

*N. Damjanov, MD, PhD, Professor, Rheumatology Department, Institute of Rheumatology, University of Belgrade, Medical School; V. Milic, MD, PhD, Rheumatology Department, Institute of Rheumatology, University of Belgrade, Medical School; J.C. Nieto-González, MD, Rheumatology Department, Hospital General Universitario Gregorio Marañón; I. Janta, MD, Rheumatology Department, Hospital General Universitario Gregorio Marañón; L. Martínez-Estupiñan, MD, Rheumatology Department, Hospital General Universitario Gregorio Marañón; B. Serrano, MD, Rheumatology Department, Hospital General Universitario Gregorio Marañón; C. Mata, MD, Rheumatology Department, Hospital General*

*Universitario Gregorio Marañón; M. Montoro, MD, Rheumatology Department, Hospital General Universitario Gregorio Marañón; D. Stanciu, MD, Rheumatology Department, Hospital General Universitario Gregorio Marañón; J. Marinković-Erić, MD, PhD, Professor, Institute of Social Medicine and Informatics, University of Belgrade, Medical School; F.J. López-Longo, MD, Rheumatology Department, Hospital General Universitario Gregorio Marañón; L. Carreño, MD, Rheumatology Department, Hospital General Universitario Gregorio Marañón; E. Naredo, MD, PhD, Rheumatology Department, Hospital General Universitario Gregorio Marañón.*

*Address correspondence to Dr. V. Milic, Institute of Rheumatology, Resavska 69, Belgrade 11000, Serbia. E-mail: veramilic@ptt.rs*  
*Accepted for publication June 2, 2016.*

tivity from 64.4% to 84.4%, with specificity decreasing slightly from 91.1% to 89.3%<sup>5</sup>. Thus, studies concluded that SGUS may be a component in the future of pSS classification criteria. However, further studies are needed to support the implementation of SGUS in routine diagnosis of pSS, such as a feasible and reliable assessment system<sup>6,7</sup>.

It should be noted that of all US evaluated variables, only parenchymal inhomogeneity with multiple focal hypo/anechoic rounded areas was shown to have diagnostic value for the disease<sup>1,8</sup>. This observation has been confirmed in comparison studies with MR<sup>9,10</sup>. Moreover, parenchymal inhomogeneity has been proven to be much more specific than sensitive in pSS<sup>3,6,8,11</sup>. In addition, color Doppler (CD) has proven to be useful in determining the degree of vascularization within the salivary glands that is considered a surrogate marker of glandular inflammation<sup>12,13</sup>. There are a few studies, however, reporting data on SGUS reliability in pSS<sup>1,4,11,14</sup>. Accordingly, reaching an optimal interobserver agreement could represent a step forward in improving the applicability of SGUS data to routine clinical practice.

The objective of our present study was to assess the multi-observer reliability of a feasible greyscale (GS) scoring system focused on SG parenchymal inhomogeneity and a CD scoring system for SG parenchymal vascularization among ultrasonographers with different levels of experience in the procedure in patients with established pSS at 2 European centers.

## MATERIALS AND METHODS

**Study design.** This study comprised 3 sections: (1) development and consensus among 5 rheumatologist ultrasonographers from 2 European centers of a novel GS scoring system for SG parenchymal inhomogeneity and a novel CD scoring system for SG parenchymal vascularization in patients with pSS; (2) a first multiobserver reliability exercise in patients with pSS and controls, held in Belgrade, Serbia, and (3) a second multiobserver reliability exercise in patients with pSS, done in Madrid, Spain.

Two Spanish experienced observers (EN and JCNG) took part in both reliability exercises. They were part of the international team of 3 experienced and 2 inexperienced ultrasonographers in the first exercise in Belgrade. In the second part of our study, organized in Madrid, they joined with 6 inexperienced ultrasonographers.

**Patients.** Included in the study were patients with an established diagnosis of pSS according to AECG criteria<sup>15</sup>. Twenty-four patients (all female; mean age  $50.2 \pm 11.2$  yrs, range 22–74; mean disease duration  $9.5 \pm 3.6$  yrs, range 1–21) with pSS who consecutively attended the Outpatient Rheumatology Clinic of the Institute of Rheumatology, Medical School, University of Belgrade, and 8 controls, recruited consecutively from the outpatient clinic and diagnosed with osteoarthritis (all female; mean age  $48.6 \pm 10.4$  yrs, range 23–62) were included for the first multiobserver reliability exercise. All controls responded negatively to sicca questions on dryness, past and current diseases, and treatments. For the second multiobserver reliability exercise, 10 patients with pSS were recruited (all female, mean age  $61.6 \pm 15.5$  yrs, range 37–90; mean disease duration  $10 \pm 5.8$  yrs, range 1.9–16.9) who consecutively attended the Outpatient Rheumatology Clinic of the Rheumatology Department, Hospital GU Gregorio Marañón.

The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committees of the Institute of Rheumatology, Belgrade, Serbia, and the Hospital GU Gregorio Marañón, Madrid, Spain.

Written informed consent was obtained from each subject before SGUS assessment.

**SGUS scoring systems.** A meeting of the investigators who participated in the first multiobserver reliability exercise was held before this exercise. These experts discussed and agreed on a standardized scoring method for 3 key areas. First, the appearance of normal SG parenchyma needed to be determined. This was defined as an oval structure with a punctiform echostucture pattern similar to normal thyroid parenchyma, with well-defined borders. Second, a semiquantitative scoring of SG parenchymal inhomogeneity needed to be outlined. This was achieved through a 4-grade semiquantitative scoring system, as follows: grade 0 = homogeneous punctiform pattern isoechogenic to normal thyroid gland; grade 1 = mild inhomogeneity without focal hypo/anechoic rounded areas; grade 2 = moderate inhomogeneity with focal hypo/anechoic rounded areas; grade 3 = severe inhomogeneity with focal hypo/anechoic rounded areas occupying the whole cross section of a gland. US grade  $\geq 2$  was considered abnormal. Third, intraglandular CD signal was interpreted through a 4-grade semiquantitative scoring system as follows: grade 0 = no parenchymal flow; grade 1 = flow signals in  $< 25\%$  of the cross section of a gland; grade 2 = flow signals in  $25\%–50\%$  of the cross section of a gland; grade 3 = flow signals in  $> 50\%$  of the cross section of a gland. It should be noted that the normal large vessels visible within the SG (i.e., external carotid artery and retro-mandibular vein in the parotid SG and facial artery and vein in the submandibular SG) were excluded from the CD score. Representative US images of the GS and CD scoring systems in patients with pSS are shown in Figure 1 and Figure 2, respectively.

**SGUS reliability assessment.** The subjects underwent GS and CD US examination of the 2 parotid and the 2 submandibular SG. SGUS scanning was performed first on GS (i.e., B-mode) and then on CD at each SG by sweeping from submandibular SG in longitudinal view to parotid SG in transverse view. The findings were evaluated in 2 perpendicular planes (i.e., longitudinal and transverse planes to the SG).

**First reliability assessment.** Two Serbian and 3 Spanish rheumatologists blindly, independently, and consecutively carried out an SGUS assessment for all 32 subjects. Three observers had 12–17 years of experience in SGUS, while the other 2 each had 1 year of experience. Observers worked in Belgrade over 2 days (total 16 h with breaks). The examination of the 4 SG of each patient or control took about 5 min.

The SGUS evaluation was performed with a commercially available real-time scanner (My Lab 70 XVG, Esaote Biomedica) with a 4–13 MHz linear array transducer, operating a B-mode frequency of 13 MHz. The CD assessment was performed by setting the machine at pulse repetition frequency (PRF) 750 Hz, gain at 50%, and frequency at 12.5 MHz. These settings were used in the examination of each subject by all observers.

**Second reliability assessment.** Eight Spanish rheumatologists blindly, independently, and consecutively performed an SGUS assessment for all 10 patients. Of the 8 observers, 6 had no experience in SGUS, and 2 (EN and JCNG) were experienced observers. The inexperienced observers were experts in musculoskeletal US but were not experienced in assessment of the SG. They learned to recognize SG pathology related with pSS, but they were not familiar with other SG pathologies because this was outside the scope of our study. One week before the exercise, the 6 inexperienced observers learned how to perform and to interpret SGUS in pSS in a 2-h training session. The examination of each patient took around 5 min.

The SGUS evaluation was carried out with a commercially available real-time scanner (LOGIQ E9, GE Medical Systems Ultrasound and Primary Care Diagnostics) with an 8–15 MHz linear arrays transducer, operating a B-mode frequency of 15 MHz. The CD assessment was performed by setting the machine at PRF 1200 Hz, gain at 21.0 db, and frequency at 7.5 MHz. These settings were used in the examination of each patient by all observers.

**Statistical analysis.** The data were analyzed using the Statistical Package for Social Sciences 16.0. The reliability of pathological US score for diagnosis pSS was assessed by the area under the receiver-operation characteristic curve (AUC-ROC)<sup>16</sup>. Interobserver agreement was calculated by the overall

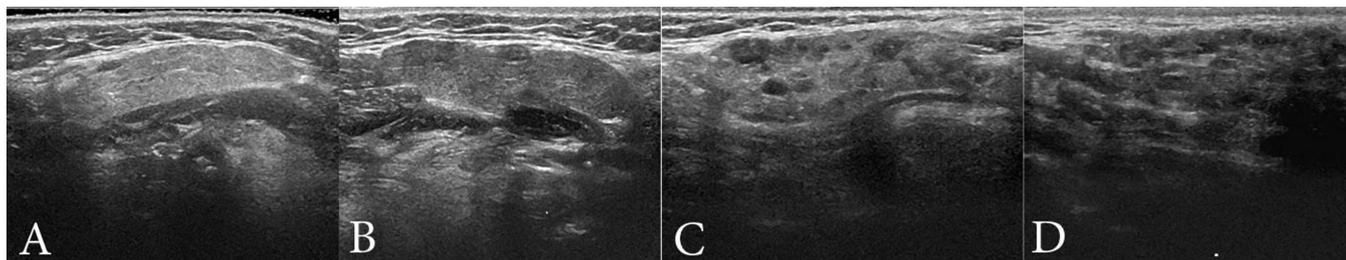


Figure 1. GS US images of submandibular gland. A: grade 0; B: grade 1; C: grade 2; D: grade 3. GS: greyscale; US: ultrasound.

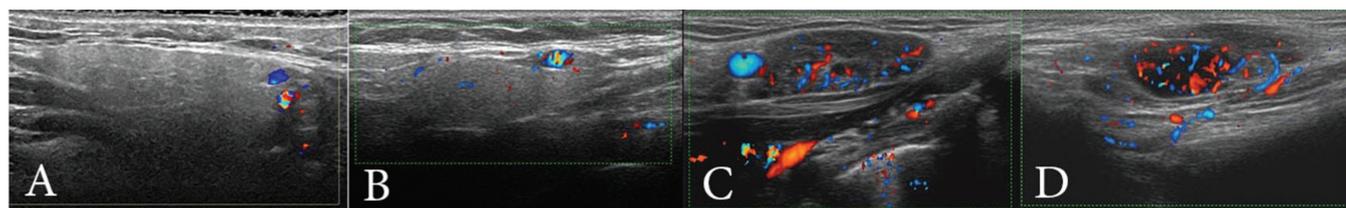


Figure 2. CD US images of submandibular gland. A: grade 0; B: grade 1; C: grade 2; D: grade 3. CD: color Doppler; US: ultrasound.

agreement (defined as the mean percentage of observed exact agreements), and Light's  $\kappa$  statistics (i.e., mean  $\kappa$  for all pairs of observations) for evaluating agreement among multiple observers. Weighted  $\kappa$  coefficients with absolute weighting were computed for GS and CD scores. A 2-tailed  $p \leq 0.05$  was considered significant. Reliability was interpreted according to Landis and Koch ( $\kappa < 0$  absence of agreement, 0.10–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 good, and 0.81–1 excellent)<sup>17</sup>.

## RESULTS

A total of 640 US examinations of SG were performed in the first reliability assessment. In the second exercise, a total of 320 US examinations of SG were performed. Features of study subjects and all assigned US scores of SG in both reliability exercises are shown in Table 1. The majority of patients with pSS (58.27%) in the first exercise and 46.9% of patients in the second exercise had pathological US scores ( $\geq 2$ ). The diagnostic reliability of pathological US score  $\geq 2$  for both parotid glands was similar and good (AUC-ROC right parotid gland 0.75; AUC-ROC left parotid gland 0.77). For submandibular glands, the diagnostic reliability US score  $\geq 2$  was slightly lower (AUC-ROC right submandibular gland 0.62; AUC-ROC left submandibular gland 0.63).

**Multiobserver agreement for GS parenchymal inhomogeneity of SG.** For the first reliability assessment, the overall agreement and the  $\kappa$  values for GS parenchymal inhomogeneity of each SG are shown in Table 2. Multiobserver agreement was good for both submandibular glands and the right parotid gland, and excellent for the left parotid gland. In addition, the agreement for GS parenchymal inhomogeneity between experienced observers ( $\kappa$  range 0.76–0.89) was higher than agreement between less experienced ones ( $\kappa$  range 0.60–0.73). Further, the overall interobserver agreement for GS parenchymal inhomogeneity in the control group was higher than in the pSS group (78.12%/68.02%,  $p < 0.01$ ). The percentage interobserver agreement for GS parenchymal inhomogeneity ranged from 78.5% to 87.5% in the control

Table 1. Features of study subjects and assigned US scores of SG in both reliability exercises.

Features	Study Subjects		
	First Exercise Patients with pSS, n = 24	Controls, n = 8	Second Exercise Patients with pSS, n = 10
Female, n (%)	24 (100)	8 (100)	10 (100)
Age, yrs, mean (range)	50.2 (22–74)	48.6 (23–62)	61.6 (37–90)
Disease duration, yrs, mean $\pm$ SD	9.5 $\pm$ 3.6	—	10 $\pm$ 5.8
Xerophthalmia, n (%)	21 (87)	—	10 (100)
Xerostomia, n (%)	24 (100)	—	10 (100)
Glandular swelling, n (%)	16 (66)	—	3 (30)
Biopsy LSG*	15/18	—	10/10
RF, n (%)	20 (83)	—	7 (70)
ANA, n (%)	21 (87)	—	8 (80)
Anti-SSA/SSB antibodies, n (%)	21 (87)	—	6 (60)
US scores of salivary glands B mode**, %			
Grade 0	20.25	83.33	33.1
Grade 1	21.49	4.16	20.0
Grade 2	37.40	8.33	25.6
Grade 3	20.87	4.16	21.3
Color Doppler***, %			
Grade 0	0	0	23.1
Grade 1	71.07	80.21	76.9
Grade 2	21.90	7.30	0
Grade 3	7.03	12.50	0

\*positive/total. \*\* mean prevalence GS scores parenchymal inhomogeneity (B mode) of SG. \*\*\* mean prevalence CD scores of SG. SG: salivary gland; US: ultrasonography; pSS: primary Sjögren syndrome; LSG: lip salivary gland; RF: rheumatoid factor; ANA: antinuclear antibodies.

group and from 54.17% to 67.92% in the pSS group. The  $\kappa$  values for GS parenchymal inhomogeneity ranged from 0.94 to 0.98 in the control group and from 0.81 to 0.84 in the pSS group (data not shown).

The overall agreement and the  $\kappa$  values for GS

Table 2. Multiobserver agreement for GS parenchymal inhomogeneity and CD signal of each SG in the first exercise (5 observers).

SG	Mean Overall Agreement, %		Weighted $\kappa$ (95% CI)*		p
	B mode	Doppler	B mode	Doppler	
Right parotid gland	68.7	84.3	0.79 (0.68–0.87)	0.70 (0.57–0.82)	< 0.0001
Experienced observers	76.2	87.4	0.88 (0.80–0.93)	0.88 (0.79–0.93)	< 0.0001
Inexperienced observers	61.2	81.2	0.73 (0.52–0.86)	0.32 (–0.01 to 0.60)	0.031
Left parotid gland	71.5	84.3	0.81 (0.71–0.89)	0.58 (0.43–0.73)	< 0.0001
Experienced observers	79.3	88.0	0.89 (0.82–0.94)	0.88 (0.79–0.93)	< 0.0001
Inexperienced observers	62.4	80.6	0.70 (0.47–0.84)	0.29 (–0.05 to 0.58)	0.047
Right submandibular gland	50.9	72.1	0.71 (0.57–0.82)	0.58 (0.43–0.73)	< 0.0001
Experienced observers	61.2	78.7	0.76 (0.62–0.86)	0.79 (0.66–0.88)	< 0.0001
Inexperienced observers	51.8	65.6	0.60 (0.33–0.78)	0.27 (–0.08 to 0.56)	0.062
Left submandibular gland	60.1	67.1	0.73 (0.60–0.84)	0.53 (0.40–0.69)	< 0.0001
Experienced observers	65.6	73.1	0.79 (0.66–0.88)	0.70 (0.54–0.83)	< 0.0001
Inexperienced observers	55.6	61.2	0.62 (0.36–0.79)	0.12 (–0.23 to 0.44)	0.254

\* p < 0.0001. GS: greyscale; SG: salivary gland.

parenchymal inhomogeneity of each SG in the second reliability exercise are shown in Table 3. There was also a good multiobserver agreement for both submandibular glands and the left parotid gland, and an excellent multiobserver agreement for the right parotid gland. The agreement difference between right and left parotid glands was small (0.82 and 0.76, respectively). In addition, the agreement for GS parenchymal inhomogeneity of SG between experienced observers ( $\kappa$  range 0.82–0.87) was higher than agreement between less experienced ones ( $\kappa$  range 0.63–0.77).

**Multiobserver agreement for parenchymal CD signal in SG.** The overall agreement and the  $\kappa$  values for parenchymal CD signal in each SG in the first reliability exercise are shown in Table 2. Multiobserver agreement was moderate for both submandibular glands and the left parotid gland, and good for the right parotid gland. The agreement for CD signal in each SG between experienced observers was significantly higher ( $\kappa$  range 0.70–0.88) than agreement between less experienced ones ( $\kappa$  range 0.12–0.32). In the control healthy group, the overall interobserver agreement for CD signal was

higher than in the pSS group (78.12%/68.02%, respectively). The percentage interobserver agreement for CD signal were slightly higher in the pSS group (72.08% to 80.83%) than in the control group (52.5% to 95.0%). The  $\kappa$  values for CD signal were 0.91–0.98 in the control group and 0.74–0.79 in the pSS group (data not shown).

The overall agreement and the  $\kappa$  values for parenchymal CD signal in each SG in the second reliability exercise are shown in Table 3. There was fair multiobserver agreement for both submandibular glands and there was no agreement for both parotid glands. Although the mean percentage agreement between inexperienced observers for CD signal of SG was significantly higher (range 75.7%–92.9%) than between experienced observers (range 50%–60%), the values of  $\kappa$  coefficients for inexperienced observers were poor.

## DISCUSSION

US is a noninvasive, available, high-resolution, and low-cost imaging modality with proven usefulness for the evaluation of the major SG in patients with pSS. Parenchymal inhomogeneity

Table 3. Multiobserver agreement for GS parenchymal inhomogeneity and CD signal of each SG in second exercise (8 observers).

SG	Mean Overall Agreement, %		Weighted $\kappa$ (95% CI)			
	B mode	Doppler	B mode	p	Doppler	
Right parotid gland	90.5	76.0	0.82 (0.78–0.85)	< 0.0001	0.04 (–0.05 to 0.13)	0.361
Experienced observers	90.0	50.0	0.87 (0.28–1.00)	0.001	0.13 (–0.19 to 0.46)	0.195
Inexperienced observers	78.6	92.9	0.77 (0.72–0.81)	< 0.001	–0.03 (–0.05 to 0.002)	0.016
Left parotid gland	87.4	77.8	0.76 (0.71–0.81)	< 0.0001	0.03 (–0.07 to 0.13)	0.544
Experienced observers	93.3	60.0	0.85 (0.33–1.00)	< 0.001	0.20 (–0.17 to 0.57)	0.14
Inexperienced observers	75.2	90.5	0.73 (0.66–0.80)	< 0.001	–0.04 (–0.07 to 0.02)	< 0.001
Right submandibular gland	86.3	67.5	0.75 (0.67–0.82)	< 0.001	0.29 (0.19–0.43)	< 0.0001
Experienced observers	93.3	60.0	0.82 (0.40–1.00)	< 0.001	0.84 (0.40–1.00)	< 0.0001
Inexperienced observers	75.7	75.7	0.63 (0.55–0.71)	< 0.001	0.25 (0.08–0.41)	< 0.001
Left submandibular gland	83.7	63.2	0.71 (0.61–0.79)	< 0.001	0.23 (0.11–0.34)	< 0.001
Experienced observers	93.3	60.0	0.84 (0.40–1.00)	< 0.001	0.20 (–0.17 to 0.57)	0.14
Inexperienced observers	74.8	71.9	0.63 (0.49–0.76)	< 0.001	0.19 (0.06–0.32)	< 0.001

SG: salivary gland; GS: greyscale; CD: color Doppler.

geneity with multiple focal hypo/anechoic rounded areas was shown to have high diagnostic value in combination with other AECG criteria for the diagnosis of pSS<sup>1,8</sup>. Longer duration of pSS is associated with increased structural damage of SG and more severe parenchymal inhomogeneity observed by US. Patients with established pSS are suitable for the evaluation of multiobserver reliability parenchymal inhomogeneity scoring. Therefore, we decided to evaluate the multiobserver reliability of SGUS for scoring GS parenchymal inhomogeneity and parenchymal CD signal in patients with established pSS in comparison to healthy subjects.

In our study, a simple 4-grade SGUS scoring system for assessing GS parenchymal inhomogeneity in patients with pSS proved good to excellent reliability within 2 different international groups comprising experienced and inexperienced observers. However, SGUS is more reliable for low scores (i.e., in identifying normal glands) than for high scores (the quantification of the abnormalities is less precise). This is supported by our finding in the first exercise. The overall interobserver agreement for GS parenchymal inhomogeneity in the control group (with low US scores) was significantly higher than in the pSS group (with high US scores). Most published studies showed (even if the scoring systems were not exactly the same) that the simplified definition of an abnormal gland for pSS diagnosis is a score  $\geq 2$  (i.e., when evident hypoechoic areas are present)<sup>1,4,5,18,19</sup>. In our study, diagnostic reliability of this simplified definition of pathological US score  $\geq 2$  was good for parotid glands and moderate for submandibular glands, evaluated by the AUC-ROC curve. This shows high reliability of a simplified approach to the recognition of normal versus abnormal SG by US in patients with pSS. A simplified SGUS scoring system for SG parenchymal inhomogeneity published by Theander and Mandl<sup>14</sup> resulted in excellent reliability between 2 observers. Some previous studies, which had used different GS scoring systems, have also reported a high interobserver reliability<sup>9,11,20</sup>. However, to the best of our knowledge, only 1 study previously assessed SGUS inter-reliability among more than 3 observers<sup>21</sup>. In this study the multiobserver agreement between 5 observers who used a 5-point rating scale was lower ( $\kappa = 0.2-0.6$ ) than in our study.

We found that the agreement for GS parenchymal inhomogeneity of salivary glands between experienced observers was higher than agreement between less experienced ones in both exercises. Of particular note was that in our study the interreliability for GS among a group of mostly inexperienced observers in second exercise was good ( $\kappa = 0.63-0.77$ ). When considering that relatively little time was needed for both the training of operators and the assessment of patients, SGUS seems to be a feasible tool for the assessment of patients with pSS. We showed that US is useful for the followup of these patients. However, it is not proven that US is reliable for the diagnostic examination of patients with pSS

by inexperienced observers. Our study does not attempt to assess the diagnostic capability of SGUS for pSS. Therefore, our results do not allow us to evaluate the reliability of SGUS finding for the early diagnosis of pSS.

Very few studies have incorporated evidence arising from Doppler mode in the evaluation of SG for patients with pSS<sup>12,13,22</sup>. We used a 4-grade semiquantitative scoring system on small vessel signals within the gland, working on the hypothesis that greater vascularization correlates to increased SG inflammation. CD multiobserver reliability proved to be high, ranging from moderate to good in the first reliability assessment. However, although the overall agreement was acceptable in the second reliability assessment, the  $\kappa$  values were fair or nonsignificant. The absence of grades 2 and 3 in the CD second reliability assessment may have led to distorted  $\kappa$  results, owing to specific limitations of the statistic tool. Interestingly, agreement for CD signals of SG between inexperienced observers was better than agreement between experienced observers in the second exercise. We concluded that many Doppler signals were generated from the movements of the SG tissue generated by pulsations of facial arteries (which should be considered as artefacts). Equipment used for the second exercise was more sensitive regarding detection of these movements in comparison to equipment used for the first exercise. Higher sensitivity of the equipment made differentiation of true CD signals and artefacts more challenging. Experienced observers did not count these artefacts, while inexperienced ones counted them as positive Doppler signals. Most probably that was the reason for the greater variability of findings between experienced observers in comparison to inexperienced ones.

There were several limitations that could influence the results of our study: small number of patients; 2 different groups of patients observed in 2 centers; use of different equipment in 2 centers; and inclusion of inexperienced ultrasonographers not familiar with the whole spectrum of SG pathology, which increased the chances of missing other pathology of the SG. Participation of inexperienced ultrasonographers increased the chance of both missing other pathologies of SG and misinterpretation of SG pathology.

Our results showed that SGUS was a reliable technique for assessing parenchymal inhomogeneity by multiple assessors in patients with established pSS in 2 European centers. SGUS is feasible to perform in routine clinical practice and in multicenter studies. Further standardization on SG CD finding is necessary to include this mode in SGUS scoring of patients with pSS. Future validation of SGUS is needed for the assessment of its usefulness in patients with pSS<sup>23</sup>.

## REFERENCES

1. Salaffi F, Carotti M, Iagnocco A, Luccioli F, Ramonda R, Sabatini E, et al. Ultrasonography of salivary glands in primary Sjogren's syndrome: a comparison with contrast sialography and scintigraphy. *Rheumatology* 2008;47:1244-9.

2. Milic V, Petrovic R, Boricic I, Radunovic G, Marinkovic-Eric J, Jeremic P, et al. Ultrasonography of major salivary glands could be an alternative tool to sialoscintigraphy in the American-European classification criteria for primary Sjögren's syndrome. *Rheumatology* 2012;51:1081-5.
3. Hocevar A, Ambrozic A, Rozman B, Kveder T, Tomsic M. Ultrasonographic changes of major salivary glands in primary Sjögren's syndrome. Diagnostic value of a novel scoring system. *Rheumatology* 2005;44:768-72.
4. Cornec D, Jousse-Joulin S, Pers JO, Marhadour T, Cochener B, Boisramé-Gastrin S, et al. Contribution of salivary gland ultrasonography to the diagnosis of Sjögren's syndrome: toward new diagnostic criteria? *Arthritis Rheum* 2013;65:216-25.
5. Cornec D, Jousse-Joulin S, Marhadour T, Pers JO, Boisramé-Gastrin S, Renaudineau Y, et al. Salivary gland ultrasonography improves the diagnostic performance of the 2012 American College of Rheumatology classification criteria for Sjögren's syndrome. *Rheumatology* 2014;53:1604-7.
6. Delli K, Dijkstra PU, Stel AJ. Diagnostic properties of ultrasound of major salivary glands in Sjögren's syndrome: a meta-analysis. *Oral Dis* 2015;21:792-800.
7. Cornec D, Jousse-Joulin S, Saroux A, Devauchelle-Pensec V. Salivary gland ultrasound to diagnose Sjögren's syndrome: a claim to standardize the procedure. *Rheumatology* 2015;54:199-200.
8. Milic VD, Petrovic RR, Boricic IV, Radunovic GL, Pejnovic NN, Soldatovic I, et al. Major salivary gland sonography in Sjögren's syndrome: diagnostic value of a novel ultrasonography score (0-12) for parenchymal inhomogeneity. *Scand J Rheumatol* 2010;39:160-6.
9. Niemelä RK, Takalo R, Pääkkö E, Suramo I, Päivänsalo M, Salo T, et al. Ultrasonography of salivary glands in primary Sjögren's syndrome. A comparison with magnetic resonance imaging and magnetic resonance sialography of parotid glands. *Rheumatology* 2004;43:875-9.
10. Makula E, Pokorny G, Kiss M, Vörös E, Kovács A, Csernay L, et al. The place of magnetic resonance and ultrasonographic examinations of the parotid gland in the diagnosis and follow-up of primary Sjögren's syndrome. *Rheumatology* 2000;39:97-104.
11. Wernicke D, Hess H, Gromnica-Ihle E, Krause A, Schmidt WA. Ultrasonography of salivary glands: a highly specific imaging procedure for diagnosis of Sjögren's syndrome. *J Rheumatol* 2008;35:285-93.
12. Carotti M, Salaffi F, Manganelli P, Argalia G. Ultrasonography and colour Doppler sonography of salivary glands in primary Sjögren's syndrome. *Clin Rheumatol* 2001;20:213-9.
13. Jousse-Joulin S, Devauchelle-Pensec V, Morvan J, Guias B, Penneç Y, Pers JO, et al. Ultrasound assessment of salivary glands in patients with primary Sjögren's syndrome treated with rituximab: Quantitative and Doppler waveform analysis. *Biologics* 2007; 1:311-9.
14. Theander E, Mandl T. Primary Sjögren's syndrome diagnostic and prognostic value of salivary gland ultrasonography using a simplified scoring system. *Arthritis Care Res* 2014;66:1102-7.
15. Vitali C, Bombardieri S, Jonsson R, Moutsopoulos HM, Alexander EL, Carsons SE, et al. Classification criteria for Sjögren's syndrome: a revised version of the European criteria proposed by the American-European Consensus Group. *Ann Rheum Dis* 2002;61:554-8.
16. Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988;240:1285-93.
17. Landis JR, Koch GG. A one way components of variance model for categorical data. *Biometrics* 1977;33:671-9.
18. Makula E, Pokorny G, Rajtar M, Kiss I, Kovacs A, Kovacs L. Parotid gland ultrasonography as a diagnostic tool in primary Sjögren's syndrome. *Br J Rheumatol* 1996;35:972-7.
19. De Vita S, Lorenzon G, Rossi G, Sabella M, Fossaluzza V. Salivary gland echography in primary and secondary Sjögren syndrome. *Clin Exp Rheumatol* 1992;10:351-6.
20. Hocevar A, Rainer S, Rozman B, Zor P, Tomsic M. Ultrasonographic changes of major salivary glands in primary Sjögren's syndrome. Evaluation of a novel scoring system. *Eur J Radiol* 2007;63:379-83.
21. Yoshiura K, Yuasa K, Tabata O, Araki K, Yonetsu K, Nakayama E, et al. Reliability of ultrasonography in sialography in the diagnosis of Sjögren's syndrome. *Oral Surg Oral Med Oral Pathol Oral Radiol Endod* 1997;83:400-7.
22. Carotti M, Ciapetti A, Jousse-Joulin S, Salaffi F. Ultrasonography of the salivary glands. *Clin Exp Rheumatol* 2014;32 Suppl 80:S61-70.
23. Jousse-Joulin S, Milic V, Jonsson MV, Plagou A, Theander E, Luciano N, et al; US-pSS Study Group. Is salivary gland ultrasonography a useful tool in Sjögren's syndrome? A systematic review. *Rheumatology* 2016;55:789-800.