Ultrasound scoring systems affect the distribution of sialadenitis scores in Sjögren's syndrome: an inter-system reproducibility study

G. Cafaro[1], C. Perricone[1], R. Bursi[1], I. Riccucci[1], S. Calvacchi[1],
G. Lomurno[2], R. Gerli[1], E. Bartoloni[1]

[1]Rheumatology Unit, Department of Medicine and Surgery, University of Perugia;
[2]Oral Surgery Unit, Department of Medicine and Surgery, University of Perugia, Italy.

Abstract

Objective

Salivary gland ultrasonography (SGUS) is commonly employed in the diagnosis and follow-up of patients with Sjögren's syndrome (SS) and multiple scoring systems have been developed to quantify the grade of sialadenitis of major salivary glands (SG). Their diagnostic performance seems overall comparable, however, the parameters evaluated by the various systems are different. The objective of this study was to compare how four different scoring systems affect the distribution of sialadenitis grades.

Methods

One hundred and three SGUS images from 26 SS patients were blindly scored by two investigators according to the De Vita, Salaffi, Milic and OMERACT scoring systems in independent sessions.

Results

The distribution of SGUS images according to De Vita, Salaffi, Milic and OMERACT systems was significantly different. At post-hoc analysis, Milic system performed differently compared to the De Vita ($p<0.0001$), OMERACT ($p<0.0001$) and Salaffi ($p<0.0001$) systems, showing a relative overestimation of sialadenitis grade.

Conclusion

Milic scoring system showed to relatively overestimate the grade of sialadenitis compared to De Vita, Salaffi and OMERACT systems. Although all scoring systems seem to be comparable in terms of diagnostic accuracy, in the prospect of selecting one system to be potentially included in future versions of SS classification criteria, it is important to compare their ability to classify SGUS images among the various degrees of sialadenitis.

Key words

Sjögren's syndrome, salivary glands, ultrasonography, sialadenitis

*Giacomo Cafaro, MD**
*Carlo Perricone, MD, PhD**
Roberto Bursi, MD
Ilenia, Riccucci, MD
Santina, Calvacchi MD
Giuseppe, Lomurno MD
Roberto Gerli, MD
Elena Bartoloni, MD

**These authors contributed equally.*

Please address correspondence to:
Roberto Gerli,
Sezione di Reumatologia,
Dipartimento di Medicina e Chirurgia,
Università di Perugia,
Piazzale Giorgio Menghini 1,
06129 Perugia, Italy.
E-mail: roberto.gerli@unipg.it
ORCID iD: 0000-0002-4684-575X

Received on May 25, 2022; accepted on June 27, 2022.

Competing interests: none declared.

Introduction

Sjögren's syndrome (SS) is a systemic autoimmune disease characterised by a protean clinical picture depending on the organs involved. Most patients display a predominant affection of the exocrine glands, mainly lacrimal and salivary glands (SG)s, and dry eye and dry mouth are the most common symptoms (1). The diagnosis of the disease is based on the clinical picture, on the presence of laboratory features, such as anti-Ro/SSA, anti-La/SSB autoantibodies, and on the detection of inflammatory infiltrates in a pathological specimen of SGs. Although not included in any set of classification criteria, ultrasonography (US) of the major SGs has been used for decades to non-invasively detect the inflammation and damage caused by the disease. Because some US aspects can be very unspecific, numerous methods have been developed to quantify the severity of sialadenitis. The first score was published in 1992 by De Vita *et al*. (2) and lately the initiative outcome measures in rheumatology (OMERACT) have developed the most recent of the scoring systems available (3). SGUS grading is based on US parameters, such as echogenicity, glandular inhomogeneity, regularity of glandular borders and the presence of hypoechoic areas and hyperechoic bands. Nonetheless, each scoring system employs and considers a different combination of these parameters and with different cut-offs in order to assign the score to a gland.

Despite SGUS is still not included in SS classification criteria, its usefulness in the diagnosis and follow-up is beyond any doubt. Multiple studies have found that a positive SGUS examination is very reliable in predicting classification of the patient as SS according to American-European consensus group (AECG) (4), American College of Rheumatology (ACR) (5) and ACR/European League against Rheumatism (EULAR) (6) classification criteria with an overall agreement ranging between 80% and 86% (7). Additionally, the weight of SGUS, if included in classification criteria, would be similar to that of a positive Schirmer's test or of dry mouth symptoms (8). A meta-analysis confirmed that SGUS may even be diagnostically superior to sialography, a test included in the AECG classification criteria (4). Even more interestingly, numerous studies have shown significant associations of SGUS features with clinical, laboratory and pathological features, such as disease activity scores, serological status, salivary flow, so that some Authors even suggest employing SGUS as a surrogate score for glandular domains in the classification criteria (9-12).

There is extensive and solid evidence on the performance of the various scoring systems in terms of sensitivity, specificity, positive and negative predictive values. The majority of studies report a sensitivity between 70% and 94% and a specificity over 85% for all the available scoring systems which seem to perform similarly (13). However, no studies have assessed the concordance among different methods in attributing a certain damage score to the same gland. Nonetheless, in the prospect of selecting a reference scoring system that may potentially be included in future versions of SS classification criteria, a careful selection and comparison among those available is essential. Therefore, in this study we evaluated the inter-method reliability of different scoring systems for the grading of sialadenitis.

Materials and methods

The study was carried out according to the Declaration of Helsinki. Ethical approval and informed consent waiver for retrospective data analysis was granted by the institutional ethical committee Comitato Etico Regionale Umbria (3994/19).

Imaging

Consecutive images of SGUS examinations performed with Esaote MyLabSeven (Esaote, Genoa, Italy) on patients classified as affected by primary SS according to the 2016 ACR/EULAR criteria (6) were screened and one representative longitudinal scan was selected for each major salivary gland. US images were then coded, anonymized and randomised, in order

Table I. Patients' characteristics. Data are provided as median (range) or percentage.

Age, years	53 (30-75)
Female sex, %	96.2
Disease duration, months	100 (0-360)
Antinuclear antibodies, %	91.3
Xerostomia, %	59.1
Xerophthalmia, %	81.8
Anti-SSA/Ro, %	95.7
Anti-SSB/La, %	43.5
Rheumatoid factor, %	65.2
Positive MSG biopsy, %	88.9
Focus score	1.88 (1.0-4.5)

Table II. Results of *post-hoc* analysis. Comparison of the distribution of glandular scores among De Vita, OMERACT, Milic and Salaffi scoring systems. Data are considered significant for $p \leq 0.0083$.

	Z	*p*
Salaffi *vs*. De Vita	-0.015	1.000
Salaffi *vs*. OMERACT	-0.340	0.353
Salaffi *vs*. Milic	-1.238	**< 0.0001**
De Vita *vs*. OMERACT	-0.325	0.424
De Vita *vs*. Milic	-1.223	**< 0.0001**
OMERACT *vs*. Milic	-0.898	**< 0.0001**

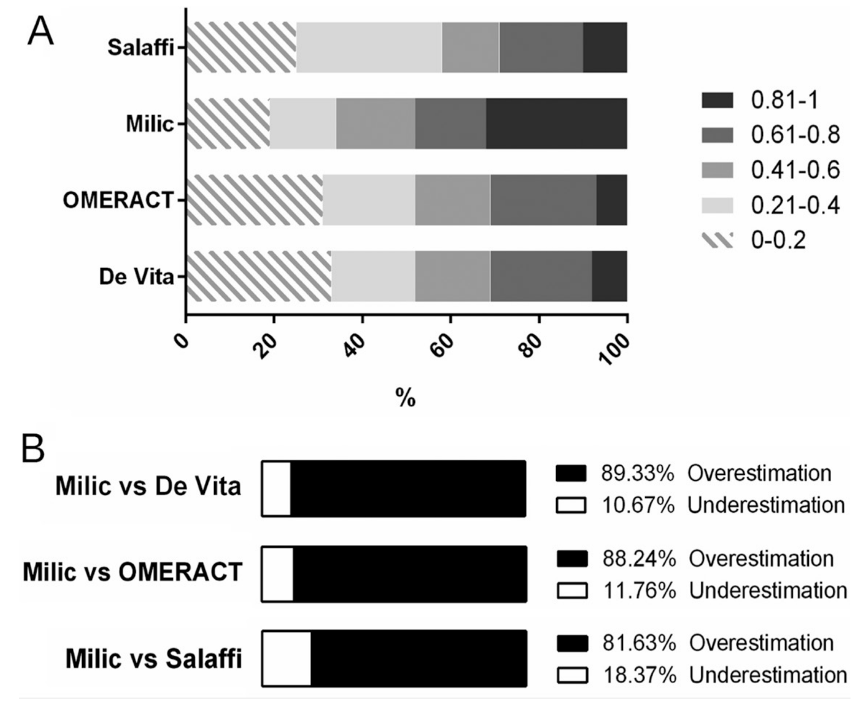

Fig. 1. Distribution of SGUS images according to sialadenitis score. Distribution of SGUS images according to mean normalised sialadenitis score **(A)**. Distribution of SGUS images with a discordant score according to Milic system overestimating or underestimating the score compared to each of the other scoring systems evaluated **(B)**.

to avoid consecutive sequential images from the same patient.

Four scoring systems were selected, namely De Vita (2), Salaffi (14), Milic (15) and OMERACT (3). The parameters considered and their definitions are shown in Supplementary Table S1. Two rheumatologists with experience in SGUS, blinded to patient identity, evaluated all images and provided a score according to each of the selected scoring systems. The scoring was performed in four independent sessions, at least 10 days apart, shuffling the images in between by random number generation, in order to prevent any recalling of specific images by the investigators.

Data analysis

The scores were normalised on a 0-1 scale in order to be able to compare the 4 scoring systems with each other and a mean score between the two raters was calculated for each gland. The distribution of the glands among the various scores were compared by Friedman test followed by *post-hoc* pairwise analysis by Wilcoxon signed-rank test. Significance was set at $\alpha = 0.05$. Bonferroni correction was applied for post-hoc analysis, resulting in a significance level set at $p \leq 0.0083$.

Results

One hundred and three US images were collected from 26 SS patients and scored as described above. Demographic and disease-related features of SS cohort are reported in Table I. A statistically significant difference in the score attribution depending on the system employed was found (χ^2 <0.0001). As shown in Table II, the *post-hoc* analysis depicted a significant difference in score attribution according to Milic compared to De Vita (Z=-1.223, $p \leq 0.0001$), OMERACT (Z=-0.898, $p \leq 0.0001$) and Salaffi (Z=-1.238, $p \leq 0.0001$). No other statistically significant differences were found among the scoring systems analysed.

To explore the differences underlying these results, we subsequently analysed the distribution of the severity grades among the scores and found that gland distribution according to Milic system was skewed towards more severe grades, *i.e.* their mode is higher, compared to De Vita, Salaffi and OMERACT. In fact, a normalised mean score >0.8 was assigned to over 25% of US images according to Milic system and to less than 10% according to the other three systems (Fig. 1A).

Finally, in order to understand how and why the Milic scoring system showed such discrepancies, we selected the US images for which a discordant grade of severity was attributed by each pair of scores and evaluated the rate of over- and under-estimation. The Milic system showed an overestimation rate of 89%, 88% and 82% compared to De Vita, OMERACT and Salaffi, respectively (Fig. 1B).

Discussion

The results of the present study demonstrated that SGUS 0–3 Milic scoring system overestimates the degree of US-detected sialadenitis compared to the De Vita, Salaffi and OMERACT scoring systems, although this difference does not seem to have a significant impact on diagnostic sensitivity and specificity.

SGUS is currently routinely performed in the work-up of patients with sicca symptoms or suspected SS in most Rheumatology Clinics and numerous scoring systems have been developed in order to quantify the grade of sialadenitis of major SGs. To our knowledge, this is the first study to compare the performance

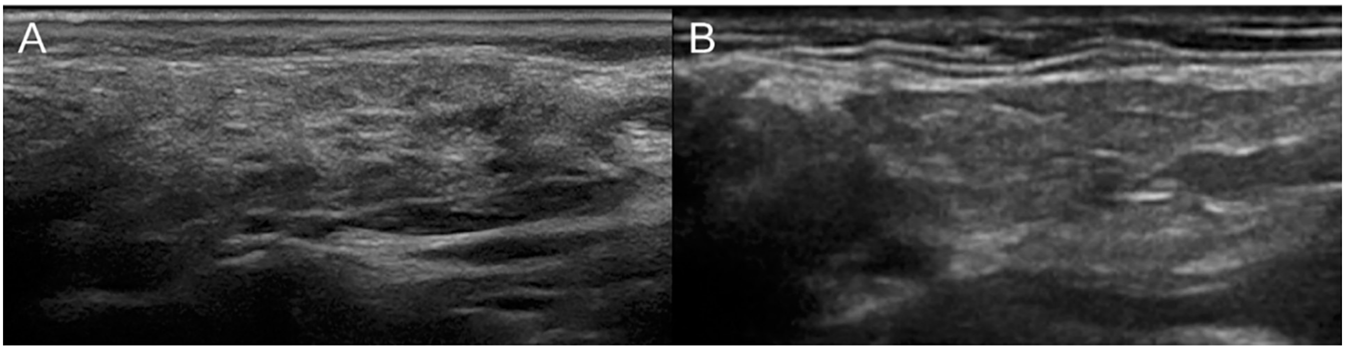

Fig. 2. Sample images with discordant scores. **A)** Right parotid gland showing diffuse inhomogeneity due to the presence of multiple small hypoechoic areas and few hyperechoic bands. The margins are not clearly distinguishable but posterior margin is visible. This image was scored as grade 3 Milic (diffuse inhomogeneity due to both hypoechoic areas and hyperchoic bands), grade 2 De Vita (hypoechoic areas are not large and confluent and hyperechoic bands are not diffuse), grade 2 OMERACT (hypoechoic areas are not occupying the entire gland) and grade 3 Salaffi (hypoechoic areas < 6mm, regular glandular volume). **B)** Left submandibular gland showing diffuse inhomogeneity due to the presence of not well-defined hypoechoic areas, irregular contours and partially visible posterior border. This image was scored as grade 3 Milic, grade 2 De Vita (hypoechoic areas are not large and confluent and hyperechoic bands are not diffuse), grade 2 OMERACT (hypoechoic areas are not occupying the entire gland) and grade 3 Salaffi (hypoechoic areas < 6mm, regular glandular volume).

of different SGUS scoring systems in terms of distribution of the glands at various scores. Each of the scoring systems considers different US parameters, such as homogeneity, borders regularity, presence of hypoechoic areas and hyperechoic bands, although the pathological counterpart of each of these aspects is still mostly unknown. In this study, we considered four of the most commonly used scoring systems, *i.e.* De Vita (2), Salaffi (14), Milic (15) and OMERACT (3). By strictly applying the parameters of each scoring system, we found that Milic system significantly differs from the other three, which, on the contrary, were not different among each other. In order to understand the reason underlying these findings, we focused on Milic scoring system and, interestingly, found that it tends to distribute the glands towards higher degrees of damage. It is interesting to underline that, unlike the other scoring systems, Milic system is more simple, only accounting for glandular inhomogeneity, while the other scoring systems consider multiple parameters, including the presence of hypoechoic areas and hyperechoic bands (2, 3, 14, 15).

Despite the absence of a strong evidence of pathologic correlations, the presence of hyperechoic bands is often considered a marker of fibro-fatty degeneration of SGs, thus a sign of long-standing sialadenitis (16), while hypoechoic areas are considered an early sign of sialadenitis and a marker of ongoing inflammation and disease activity. For this reason, hypoechoic areas are included in the lower grades of sialadenitis while hyperechoic bands are an indicator of a higher degree of damage by both De Vita and Salaffi systems (Table II). However, both hypoechoic areas and hyperechoic bands produce an alteration of US homogeneity. Consequently, when Milic system is applied, a SG with only hypoechoic areas may be scored equally to one with hyperechoic bands leading to the abovementioned tendency to relatively overestimate the degree of sialadenitis. Two sample images with discordant scores are shown and described in Figure 2.

The studies evaluating the various scoring systems in terms of diagnostic accuracy seem to show comparable performances, with no system being clearly superior to the others. It is likely that when the overall score of the four glands is computed and an appropriate cut-off score is selected, their diagnostic performance may actually be equivalent. Notably, Milic scoring system seems to require higher cut-off values in order to obtain sensitivity and specificity parameters comparable to other systems, thus reinforcing the results of our study (13).

Nonetheless, we strongly believe that studies comparing the performance of different scoring systems are essential in the prospect of including SGUS findings in future SS classification criteria. In fact, one of the main hurdles towards this objective is the absence of a reference standard scoring method, which is unlikely to emerge in the near future. The selection of one scoring method, its cut-off and its relative weight cannot take place aside from a careful consideration of these aspects.

The strength of our study relies on the methodology employed to score the US images. Although the inter-rater reliability of SGUS scoring systems is known to be high or very high (17), the evaluation was performed by two investigators, blinded to patient identity and following a strict randomisation of the images in order to avoid spill-over bias. The main weaknesses of the study are due to the fact that only four scoring methods were evaluated, thus no information can be provided on other available systems.

In the prospect of acquiring additional evidence on the pathologic significance of hypoechoic areas and hyperechoic bands and of including SGUS in the classification criteria of SS (18), differences among SGUS scoring systems should be carefully considered and more comparative studies should be performed.

References

1. CAFARO G, BURSI R, CHATZIS LG *et al.*: One year in review 2021: Sjögren's syndrome. *Clin Exp Rheumatol* 2021; 39 (Suppl. 133): S3-13. https://doi.org/10.55563/clinexprheumatol/eojaol
2. DE VITA S, LORENZON G, ROSSI G, SABELLA M, FOSSALUZZA V: Salivary gland echogra-

phy in primary and secondary Sjögren's syndrome. *Clin Exp Rheumatol* 1992; 10: 351-6.

3. 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: 967-73. https://doi.org/10.1136/annrheumdis-2019-215024
4. VITALI C, BOMBARDIERI S, JONSSON R *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. https://doi.org/10.1136/ard.61.6.554
5. SHIBOSKI SC, SHIBOSKI CH, CRISWELL LA *et al.*: American College of Rheumatology classification criteria for Sjögren's syndrome: a data-driven, expert consensus approach in the Sjögren's International Collaborative Clinical Alliance cohort. *Arthritis Care Res* (Hoboken) 2012; 64: 475-87. https://doi.org/10.1002/acr.21591
6. 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. *Arthritis Rheumatol* 2017; 69: 35-45. https://doi.org/10.1002/art.39859
7. MOSSEL E, DELLI K, VAN NIMWEGEN JF *et al.*: Ultrasonography of major salivary glands compared with parotid and labial gland biopsy and classification criteria in patients with clinically suspected primary Sjögren's syndrome. *Ann Rheum Dis* 2017; 76(11): 1883-89. https://doi.org/10.1136/annrheumdis-2017-211250
8. 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: 180-8. https://doi.org/10.1111/joim.12992
9. FIDELIX T, CZAPKOWSKI A, AZJEN S, ANDRIOLO A, TREVISANI VFM: Salivary gland ultrasonography as a predictor of clinical activity in Sjögren's syndrome. *PLoS One* 2017; 12: e0182287. https://doi.org/10.1371/journal.pone.0182287
10. INANC N, ŞAHINKAYA Y, MUMCU G *et al.*: Evaluation of salivary gland ultrasonography in primary Sjögren's syndrome: does it reflect clinical activity and outcome of the disease? *Clin Exp Rheumatol* 2019; 37 (Suppl. 118): S140-5.
11. KIM J-W, LEE H, PARK S-H, KIM S-K, CHOE J-Y, KIM JK: Salivary gland ultrasonography findings are associated with clinical, histological, and serologic features of Sjögren's syndrome. *Scand J Rheumatol* 2018; 47: 303-10. https://doi.org/10.1080/03009742.2017.1374451
12. MILIC V, COLIC J, CIRKOVIC A, STANOJLOVIC S, DAMJANOV N: Disease activity and damage in patients with primary Sjogren's syndrome: Prognostic value of salivary gland ultrasonography. *PLoS One* 2019; 14: e0226498. https://doi.org/10.1371/journal.pone.0226498
13. ZHOU M, SONG S, WU S *et al.*: Diagnostic accuracy of salivary gland ultrasonography with different scoring systems in Sjögren's syndrome: a systematic review and meta-analysis. *Sci Rep* 2018; 8: 17128. https://doi.org/10.1038/s41598-018-35288-5
14. SALAFFI F, ARGALIA G, CAROTTI M, GIANNINI FB, PALOMBI C: Salivary gland ultrasonography in the evaluation of primary Sjögren's syndrome. Comparison with minor salivary gland biopsy. *J Rheumatol* 2000; 27: 1229-36.
15. MILIC VD, PETROVIC RR, BORICIC IV *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. https://doi.org/10.3109/03009740903270623
16. ZABOTTI A, ZANDONELLA CALLEGHER S, GANDOLFO S *et al.*: Hyperechoic bands detected by salivary gland ultrasonography are related to salivary impairment in established Sjögren's syndrome. *Clin Exp Rheumatol* 2019; 37 (Suppl. 118): S146-52.
17. 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* 2020; 7: 581248. https://doi.org/10.3389/fmed.2020.581248
18. GENG Y, LI B, DENG X, JI L, ZHANG X, ZHANG Z: Salivary gland ultrasound integrated with 2016 ACR/EULAR classification criteria improves the diagnosis of primary Sjögren's syndrome. *Clin Exp Rheumatol* 2020; 38: 322-28. https://doi.org/10.55563/clinexprheumatol/13u0rt