Primary Sjögren’s Syndrome
Delli, Konstantina

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2017

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 12-05-2020
Chapter 3C

Ultrasound of the major salivary glands is a reliable imaging technique in patients with clinically suspected primary Sjögren’s syndrome

Konstantina Delli, Suzanne Arends, Jolien F van Nimwegen, Pieter U Dijkstra, Alja J Stel, Frederik KL Spijkervet, Hendrika Bootsma, Arjan Vissink, on behalf of the EULAR US-pSS Study Group

1. Department of Oral and Maxillofacial Surgery
2. Department of Rheumatology and Clinical Immunology
3. Department of Rehabilitation, Center for Rehabilitation

University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

Accepted for publication in European Journal of Ultrasound/ Ultraschall in der Medizin
Introduction

Primary Sjögren’s syndrome (pSS) is a chronic systemic auto-immune disease, with an estimated prevalence of 61 cases per 100,000 inhabitants in the general population [1]. pSS affects the exocrine glands, the salivary and lacrimal glands in particular, resulting in a sensation of dry eyes (keratoconjunctivitis sicca) and dry mouth (xerostomia) [2].

In the assessment of salivary gland component in pSS, ultrasonography of the major salivary glands merits currently special attention as a non-invasive, inexpensive, widely available, easily accessible and non-irradiating imaging modality [3]. Ultrasound is generally considered to be an operator-dependent technology with mediocre repeatability [4]. To date, numerous studies have been published investigating the diagnostic properties of ultrasound in pSS. In a recently published meta-analysis, it was estimated that ultrasound has pooled sensitivity of 0.69 and specificity of 0.92 for the diagnosis of SS [5]. However, the reliability of ultrasound and individual ultrasonographic variables as well as sources of variation in outcomes have not been thoroughly analyzed with regard to the salivary gland involvement in pSS [6].

The aim of this study was to assess the intra- and inter-observer reliability of B-mode ultrasonography in the diagnostic work-up of pSS with regard to the involvement of the major salivary glands in a representative study population of patients with clinically suspected pSS and to analyze sources of variation in outcomes of the ultrasonographic evaluation.

Materials and methods

Patients

Eighty consecutive patients with clinically suspected pSS were included in the study. All patients visited the outpatient clinic of the department of Rheumatology and Clinical Immunology, Multidisciplinary Sjögren's Expertise Center, University Medical Center Groningen, University of Groningen, between January 2015 and September 2015 and were subjected to ultrasonographic evaluation and routine diagnostic work-up, according to American European Consensus Group (AECG) criteria [7].

Procedures

All patients were examined with the same ultrasonographic scanner (Esaote MyLabSeven, Genova, Italy), equipped with a high resolution linear scanner (4-
Each patient was scanned in a supine position with the neck slightly extended and the head turned a little to the opposite side. The parotid glands were examined in both axial and coronal planes, while the submandibular glands only in the coronal plane.

The following images were stored from each patient and used for this reliability study: one showing the thyroid gland, one showing the right submandibular salivary gland, one showing the left submandibular salivary gland, two providing an overview of the right parotid gland and two providing an overview of the left parotid gland. Images were processed in such a way that no patient data were visible, i.e. name, age, gender and date of examination (Supplementary Figure 1). The images were allocated to a random number.

All images were scored independently by three observers (KD, JFN and AJJS for scoring system see below) on the same monitor (MultiSync E231, 23 inches, NEC, Illinois, USA). The observers were blinded for the diagnostic work up, i.e. salivary gland biopsy, circulating auto-antibodies, salivary function tests, tear gland function tests and subjective oral and ocular symptoms [7]. The observers scored all patients in random order in two sessions with a 2-week interval. Prior to the study, exact instructions were given to all observers regarding the scoring and all observers were trained. Additionally, five pSS patients, who were not included in the reliability assessment, were scored in a consensus meeting for calibration, i.e. to train the observers to score the ultrasonographic images consistently.

Ultrasonographic assessments of B-mode images

The following ultrasonographic variables were assessed in each major salivary gland: echogenicity, parenchymal homogeneity, the presence of hypoechogenic areas, the presence of hyperechogenic reflections, and the clearness of posterior glandular border, according to the Hocevar scoring system [3]:

i. Parenchymal echogenicity was evaluated in comparison with the thyroid gland or when there was coincident thyroid gland disease by surrounding anatomical structures (muscular structures, sub-cutaneous fat). Echogenicity was graded 0 if echogenicity was comparable to the thyroid, and 1 if it was decreased.

ii. Homogeneity was graded 0 for a homogeneous gland, 1 for mild inhomogeneity, 2 for evident inhomogeneity, and 3 for a grossly inhomogeneous gland.

iii. Presence of hypoechogenic areas was graded 0 for no hypoechogenic areas, 1 for a few scattered areas, 2 for several areas, and 3 for numerous hypoechogenic areas.

iv. Presence of hyperechogenic reflections in the parotid glands was graded 0 for no hyperechogenic reflections, 1 for a few, scattered, 2 for several, and 3 for numerous hyperechogenic reflections, and in submandibular glands 0 for absent and 1 for present.

v. Clearness of salivary gland borders was graded 0 for clear, regularly defined borders, 1 for partly defined borders, 2 for ill-defined borders, and 3 for borders not visible.

Finally, ultrasound total score was calculated as the sum of the grades for the five variables described above for all four glands (range 0-48). According to the literature, the cut-off point to define positive or negative ultrasound for pSS was set at 15 [8] and 17 [3].

Data analysis

Patient characteristics were presented as number of patients (%) or median (inter-quartile range; IQR: Q1-Q3). Intra-observer reliability was assessed by comparing the ultrasound scores from the first and second session for each rater. Inter-observer reliability was assessed by comparing the ultrasound scores from both sessions between the observers. Intra-observer and inter-observer reliability was calculated using Cohen's Kappa and Fleiss' Kappa, respectively, in combination with the percentage of absolute agreement (calculated as: n-agreement/n-total) for nominal variables, Intraclass Correlation Coefficient (ICC; two-way mixed effects model, single measures, absolute agreement) was assessed for continuous variables. Kappa agreement was calculated with online statistical tools available at: http://www.statstodo.com/CohenKappa_Pgm.php and http://dfreelon.org/recal/recal3.php. Kappa and ICC values were interpreted as follows: 0.00-0.20, poor; 0.20-0.40, fair; 0.40-0.60, moderate; 0.60-0.80, good; and 0.80-1.00, excellent [9].

Variance components (type III ANOVA) were computed to determine the impact of factors influencing variation in ultrasonographic scores. The factors ‘patient’, ‘session’ and ‘observer’, were considered as random factors, and their 2-way interactions were analyzed. Error variation was calculated as the sum of all sources of variation minus patient variation. The contribution of factors to the total variation and the error variation was expressed as percentage. For each patient, the mean of the 6 observations (3 observers, 2 sessions) and the difference of the 6 observations with the mean were calculated and plotted against each other. Statistical analysis was performed with IBM SPSS Statistics 22 (SPSS, Chicago, IL, USA).

Results

Of the 80 included clinically suspected patients with pSS, the median age was 51 years (IQR=52-62), 69 (86%) were female, and median ultrasound total score was 15 (IQR=14-18).
Reliability of B-mode ultrasound

Intra-observer reliability of the US total score was excellent, with ICCs ranging from 0.89 to 0.96 for the three observers (Table 1).

Inter-observer reliability was good to excellent, with ICCs of 0.84 and 0.76 for the ultrasound total score in session one and two, respectively. The kappa ranged from 0.60 to 0.83 and the percentage of absolute agreement from 80 to 92 depending on the cut-offs applied to define positive or negative ultrasound for pSS (Table 2). Inter-observer reliability was higher for cut-off point of ≥17 compared to ≥15.

Regarding the individual ultrasonographic variables of the Hocevar’s scoring system, hypoechoic areas and homogeneity of the parotid glands showed the highest inter-observer reliability with median ICCs of 0.74 and 0.71, respectively, whereas median kappa was for echogenicity of the parotid glands (0.22) was low (Table 3).

Systematic differences and sources of variation

Differences between the three observers were larger for higher ultrasound total scores (Figure 1). The contribution of error variance to the total variance was 21.3%. The interaction of patient and observer made the largest contribution to error variance followed by the main effect of observer. The effects of session, interaction of session and observer, and interaction of session and patient contributed marginally to the error variance (Figure 2 and Supplementary Table 1).

Discussion

In the classification of pSS, involvement of the salivary glands is currently assessed with sialography, scintigraphy, sialometry and histopathology. Recent discussion has focused on the diagnostic accuracy of B-mode ultrasound to evaluate the involvement of the major salivary glands [10]. However, along with assessing the validity of ultrasound, evaluating the reliability of this imaging technique, i.e. the consistency or repeatability of the measurements, is of equal importance.

Our study, including patients with clinically suspected pSS, showed that the intra- and inter-observer reliability of the ultrasound total score, when applying the Hocevar scoring system [3] was excellent. When different cut-off points reported in the literature for ultrasound positivity were applied [3,8], the agreement ranged from moderate to excellent. These results are in accordance with previously published studies using various scoring systems and different study populations [11-19].

### Table 1: Intra-observer reliability of ultrasound of the major salivary glands in patients with clinically suspected pSS.

<table>
<thead>
<tr>
<th></th>
<th>ICC</th>
<th>95% CI lower limit</th>
<th>95% CI upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>ultrasound total score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observer 1</td>
<td>0.96</td>
<td>0.90</td>
<td>0.98</td>
</tr>
<tr>
<td>Observer 2</td>
<td>0.92</td>
<td>0.95</td>
<td>0.99</td>
</tr>
<tr>
<td>Observer 3</td>
<td>0.89</td>
<td>0.83</td>
<td>0.96</td>
</tr>
<tr>
<td>Diagnosis if:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cut-off≥15 [8]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observer 1</td>
<td>0.74</td>
<td>0.58</td>
<td>0.89</td>
</tr>
<tr>
<td>Observer 2</td>
<td>0.77</td>
<td>0.64</td>
<td>0.91</td>
</tr>
<tr>
<td>Observer 3</td>
<td>0.73</td>
<td>0.57</td>
<td>0.88</td>
</tr>
<tr>
<td>Cut-off≥17 [3]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observer 1</td>
<td>0.85</td>
<td>0.73</td>
<td>0.97</td>
</tr>
<tr>
<td>Observer 2</td>
<td>0.89</td>
<td>0.80</td>
<td>0.99</td>
</tr>
<tr>
<td>Observer 3</td>
<td>0.70</td>
<td>0.54</td>
<td>0.85</td>
</tr>
</tbody>
</table>

The cut-off point to define positive or negative ultrasound for pSS was set at 15 [8] and 17 [3].

### Table 2: Inter-observer reliability of ultrasound of the major salivary glands in patients with clinically suspected pSS.

<table>
<thead>
<tr>
<th></th>
<th>ICC</th>
<th>95% CI lower limit</th>
<th>95% CI upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>ultrasound total score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session 1</td>
<td>0.84</td>
<td>0.71</td>
<td>0.91</td>
</tr>
<tr>
<td>Session 2</td>
<td>0.76</td>
<td>0.52</td>
<td>0.87</td>
</tr>
<tr>
<td>Diagnosis if:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cut-off≥15 [8]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session 1</td>
<td>0.75</td>
<td>0.62</td>
<td>0.88</td>
</tr>
<tr>
<td>Session 2</td>
<td>0.60</td>
<td>0.47</td>
<td>0.72</td>
</tr>
<tr>
<td>Cut-off≥17 [3]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Session 1</td>
<td>0.83</td>
<td>0.71</td>
<td>0.96</td>
</tr>
<tr>
<td>Session 2</td>
<td>0.70</td>
<td>0.57</td>
<td>0.83</td>
</tr>
</tbody>
</table>

The cut-off point to define positive or negative ultrasound for pSS was set at 15 [8] and 17 [3].
This suggests that different observers may rather consistently identify in which patients ultrasound of the major salivary glands supports the diagnosis of pSS, but scoring the severity of the ultrasonographic findings is more inconsistent between the observers. The reason for this variance could be mainly attributed to the observers and to the interaction between observer and patient (Figure 2). A possible consequence of this phenomenon is that when monitoring patients over time, the observed change might not be only attributed to the progression of the disease or to the effect of medication, but might be partly the result of the discrepancy in scoring between the different observers.

To the best of our knowledge, this is the first study where an analysis of the sources of variation in outcomes has been investigated in such detail. A next step is to identify the minimal clinically important difference, which is defined as the smallest difference in score in the domain of interest which patients perceive as important change and which would mandate a change in the patient’s management [23].

**Strengths and limitations of the study**

Strength of the current study is that we included a representative study population reflecting the daily practice. Moreover, we focused on the Hocevar scoring system [3]. We have chosen to use this extensive scoring system as it is one of the most detailed ultrasound scoring systems used today [5]. The Hocevar scoring system can be easily ‘transformed’ to practically any of the existing ones by combining certain ultrasound variables. So, it is rather generally applicable. Additionally, the fact that we analyzed the reliability on the individual ultrasonographic characteristics allowed us a more detailed approach. The fact that we analyzed static images instead of live ones might be considered as a limitation of this study, although

When assessing the individual ultrasonographic parameters, the inter-observer reliability was good for homogeneity and the presence for hypoechogenic areas; moderate for hyperechogetic reflections and salivary gland border; and fair for echogenicity. The parotid glands scored lower compared to the submandibular glands for hyperechogetic reflections and salivary gland border. A possible explanation is that there is more room for interpretation when scoring the hyperechogetic reflections in the parotid glands, since these findings could be attributed to pSS as well as to normal ageing [20,21]. Furthermore, when it comes to echogenicity, it might be difficult for the observers to determine the overall echogenicity of a very inhomogenous gland. In general, in the absence of a well-defined consensual scoring system and standardized procedures, ultrasonography has an acceptable reliability for homogenicity of the parenchyma, but not for echogenicity [18]. The development of a validated automatic software may resolve these issues [22].

Additionally, we observed that results between observers were more widespread on patients with higher ultrasonographic scores (approximately scores≥20; Figure 1). This suggests that different observers may rather consistently identify in which patients ultrasound of the major salivary glands supports the diagnosis of pSS, but scoring the severity of the ultrasonographic findings is more inconsistent between the observers. The reason for this variance could be mainly attributed to the observers and to the interaction between observer and patient (Figure 2). A possible consequence of this phenomenon is that when monitoring patients over time, the observed change might not be only attributed to the progression of the disease or to the effect of medication, but might be partly the result of the discrepancy in scoring between the different observers.

To the best of our knowledge, this is the first study where an analysis of the sources of variation in outcomes has been investigated in such detail. A next step is to identify the minimal clinically important difference, which is defined as the smallest difference in score in the domain of interest which patients perceive as important change and which would mandate a change in the patient’s management [23].
this is common approach in similar studies. The development of a consensus and widely accepted ultrasonographic scoring system by the EULAR US-pSS Study Group for evaluating the major salivary glands of patients with pSS will allow better comparison between studies.

Conclusion

Our study showed that B-mode ultrasound of the major salivary glands is a reliable imaging technique for patients with clinically suspected pSS. The study also pointed that results between observers were more widespread on patients with higher ultrasonographic scores and that there are some discrepancies between observers in assessing the severity of ultrasonographic findings. Thus, when assessing ‘true’ changes over time, i.e. when monitoring the activity or progression of pSS, it is advised that each particular patient is scored by the same ultrasonographer at every time point.

References

19. Hofauer B, Mansour N, Heiser C, et al. So-

Figure 2: Variance components of ultrasonographic examination of the major salivary glands. Total variance (left circle) comprised patient variance (main effect) and error variance. Several sources contributed to error variance. These sources (right circle) comprised main effects (session and observer), interaction effects (patient × observer and session × observer) and residual variance (all expressed as percentages of error variance).
Supplementary Figure 1: Representative ultrasonographic images of the major salivary glands: a. parotid gland with normal echostructure; b. parotid gland with echostructure corresponding to pSS; c. submandibular gland with normal echostructure; d. submandibular gland with echostructure corresponding to pSS. Images were processed in such a way that no patient data were visible.

Supplementary Table 1: Impact of components influencing variation in ultrasonographic scores.

<table>
<thead>
<tr>
<th>Sources of variation</th>
<th>Estimate</th>
<th>% of total variation</th>
<th>% of error variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
<td>59.05</td>
<td>78.7</td>
<td>-</td>
</tr>
<tr>
<td>Sessions</td>
<td>0.00</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Observer</td>
<td>4.81</td>
<td>6.4</td>
<td>30.0</td>
</tr>
<tr>
<td>Patient x observer</td>
<td>5.90</td>
<td>7.9</td>
<td>37.0</td>
</tr>
<tr>
<td>Patients x session</td>
<td>1.06</td>
<td>1.4</td>
<td>6.6</td>
</tr>
<tr>
<td>Session x observer</td>
<td>0.43</td>
<td>0.6</td>
<td>2.8</td>
</tr>
<tr>
<td>Variance Error</td>
<td>3.78</td>
<td>5.0</td>
<td>23.7</td>
</tr>
</tbody>
</table>

Dependent Variable: ultrasound total score.
Method: ANOVA (Type III Sum of Squares).


23. Jaeschke R, Singer J, Guyatt GH. Ascertain-