Original papers

Med Ultrason 2014, Vol. 16, no. 1, 32-36
DOI: 10.11152/mu.2014.2066.161.vv1fb2

Inter & Intra-Observer Reliability Of Grading Ultrasound Videoclips With Hand Pathology In Rheumatoid Arthritis By Using Non- Sophisticated Internet Tools (LUMINA Study)

VioletaVlad[1], Florin Berghea[2], Annamaria Iagnocco[3], Mihaela Micu[4], Nemanja Damjanov[5], Vlado Skakic[6], Slavica Prodanovic[5], Goran Radunovic[5], Marcin Szkudlarek[7], Rodina Nestorova[8], Tzvetanka Petranova[9], Jasna Kakavouli[10], Francesco Porta[11], Carlo Perricone[3], Anna Ciechomska[12], Ingrid Moller[13], Luminita Varzaru[14], Porin Peric[15], Christian Dejaco[16], Mihai Bojinca[17], Daniela Fodor[18], Mihaela Milicescu[17], Esperanza Naredo[19]

[1]Clinical Hospital Sf. Maria, Bucharest, Romania, [2]UMF „Carol Davila" Bucharest, Romania, [3]Sapienza Universita di Roma, Italy, [4]Rehabilitation Clinical Hospital, Cluj Napoca, Romania, [5]Institute of Rheumatology Belgrade, Serbia, [6]Institute for Treatment and Rehabilitation Niskabanja, Serbia, [7]University of Copenhagen, Hospital at Koge, Denmark, [8]Rheumatology Center „St. Irina" Sofia, Bulgaria, [9]Clinic of Rheumatology, Sofia, Bulgaria, [10]Private Rheumatology Practice, Katerini, Greece, [11]University of Florence, Italy, [12]Interclyde Royal Hospital, Glasgow, UK, [13]Instituto Poal de Reumatologia, Barcelona, Spain, [14]Center of Rheumatic Diseases „Ion Stoia" Bucharest, Romania, [15]Clinical Hospital Center Zagreb, Croatia, [16]Medical University Graz, Austria, [17]„Dr. Ion Cantacuzino" Hospital, Bucharest, Romania, [18]UMF „Iuliu Haţieganu" Cluj Napoca, Romania, [19]Hospital General Universitario Gregorio Maranon, Madrid, Spain

Abstract

Aim: To evaluate the inter- and intraobserver agreement of a group of European rheumatologist ultrasonographers in grading musculoskeletal ultrasound videoclips posted on the Internet by using a non-sophisticated electronic environment. **Methods:** Forty short movie clips (less than 30 secs) were made available over the Internet to all participants. Normal and pathological RA hand joints and tendons were included in the movie clips. In the first phase 30 investigators from European countries were invited to evaluate the clips and to interpret/grade them. No instruction session was held prior to the initiation of the study. For synovitis the requested scoring system included 0 to3 grades and for tenosynovitis a binary variable 0/1; separate evaluations were performed for gray scale (GS) and Power Doppler (PD) examinations. In the second phase the responders were asked to grade the same clips in a different order without having access to their first grading scale. Light's k and Cohen's k were used to analyse inter- and intraobserver reliability. **Results:** Twenty two European rheumatologists agreed to finalise both study phases. Mean Cohen's κ for intraobserver reliability was 0.614/0.689 for tenosynovitis GS/PD and 0.523/0.621 for synovitis GS/PD. Light's k for interobserver reliability was 0.503 for tenosynovitis evaluation and 0.455 for global (synovitis and tenosynovitis) evaluation. Mean global overall agreement was 84.95% (90.2% for global synovitis). **Conclusions:** An over-the-net US evaluation and grading has shown moderate to good reliability. The results could be improved if a training session is added at the beginning of the study.

Keywords: ultrasonography, rheumatoid arthritis, reliability, internet

Introduction

Ultrasonographic (US) assessment of hand joints in rheumatoid arthritis (RA) is widespread nowadays; rheumatologists are finding serious advantages in their work by using images of synovitis and its vascularization to accurately diagnose and treat patients as the technique is safe, non-invasive, relatively inexpensive and patient-friendly. The method is still considered as „operator-dependent", although standardization of both acquisition and interpretation of US images [1,2] as well as validated semi quantitative scales for synovitis quantification [3] have been published.

To complete the standardization of this procedure a large amount of information regarding US findings in RA patients is still required. A recent systematic review of the reported

Received 27.12.2013 Accepted 15.01.2014
Med Ultrason
2014, Vol. 16, No 1, 32-36
Corresponding author: Violeta Vlad,
Clinical Hospital Sf. Maria
37-39 Ion Mihalache Bd,
Bucharest, Romania;
Fax: +040212223555
Email:vladvioleta1@gmail.com

work in the area of US joint count and synovitis scoring in RA identified 3004 reports; however, only few attempts met the final criteria for becoming an „*US scoring system*" [4]. One of the limitations in studies involving more investigators consists of a variable degree of interobserver agreement. Agreement testing is usually very time consuming and uncomfortable for the patients involved in the exercise. In order to increase the reliability of US, grading exercises of US images are added to life sessions with patients [5].

Using videoclips instead of US static images seems to offer more advantages as both gray scale (GS) and Power Doppler (PD) examinations are more realistic when evaluated dynamically. A previous study [6] tested the reliability of dynamic images assessment using videoclips posted on CD-ROMs delivered to investigators in two rounds, with good results.

The purpose of this study was to test the reliability of US grading scales on videoclips by using existing non-sophisticated Internet tools. Our protocol included videoclips containing RA hand pathology (synovitis and tenosynovitis in B mode and PD mode) distributed on the Internet for scoring to European rheumatologists highly trained in MSUS.

Materials and methods

Forty US short movie clips (30 secs or less) were recorded by the same ultrasonographer (with 10 years experience in musculoskeletal US) using the same machine (Esaote MyLab 70, with 18 MHz linear probe). The technical characteristics were kept unchanged for all clips. The video clips contained moving images, with appropriate landmarks, according to the literature [1]. Both normal and RA hand joints and tendons were included, with GS and PD pathological elements. The clips were made available over the Internet to all participants being uploaded on *You Tube* (www.youtube.com) in a dedicated channel (LUMINA) and were randomly assigned numbers from 1 to 40 (link to *You Tube* available as supplementary data 1). All video clips were taken from the archive of the US machine, and the patients' names and personal data were erased, so no informed consent of patients was necessary. The study was approved by the Local Ethical Committee.

An on-line questionnaire was created to record each participant's assessment by using the *Survey Monkey* Internet service; one question was created for each videoclip. The questions were designed as follows: „Please score the following items from videoclip no X". In situations where more than one structure was present in the clip (e.g. radiocarpal joint GS, radiocarpal joint PD, intercarpal joint GS, intercarpal joint PD, extensor tendon 4th compartment GS and PD, etc) separate evaluations were requested for the same question (questionnaire available as supplementary data 2). The type of joint/tendon and the transducer position were mentioned; the respondent was asked to mark a grade from 0 to 3 separately for GS and PD synovitis, according to Szkudlarek et al [2] semiquantitative scale, and a grade 0/1 separately in GS and PD for tendons (0 – no tenosynovitis; 1 – tenosynovitis present). The length of video clips was not the same; however, responders had the technical possibility to pause, resume and replay each clip as frequently as they needed.

The pathology included in the videoclips consisted of the following: radiocarpal and intercarpal synovitis – 16 videoclips (dorsal view, at the level of the lunate and capitate bones); metacarpophalangeal (MCP) and proximal interphalangeal (PIP) synovitis (dorsal and volar views) – 11 videoclips; carpal extensors compartments 2, 4 and 6 – 25 videoclips; finger flexors 2-5 tenosynovitis – 7 videoclips: and normal joints and tendons (5 joint clips and 5 tendon clips). The definitions for the US pathological findings used in this study were that of OMERACT [2].

The links to the *You Tube* channel and the *Survey Monkey* questionnaire were sent to 30 US experts or advanced practitioners from Europe. One week after the first grading, each participant entered the second phase of LUMINA exercise and received a second e-mail with a new link to the *You Tube* channel, containing the same video clips in an automatically randomised different order with a different number and a new *Survey Monkey* questionnaire. The participants had no access to their first grading scale. Only the study statistician was unblinded to the correspondence between the two questionnaires.

Statistical analysis

Statistical analysis was carried out using SPSS V.16 software. Values <0.05 were considered significant (table II.). Overall agreement, defined as the percentage of exact agreement observed, was calculated for each pathology type (synovitis/tenosynovitis) and global. Intraobserver concordance was assessed by Cohen's k index. Interobserver concordance was assessed by Light's k (the average of k values obtained for all possible pairs of observation between our observers) [7-11]. Agreement coefficients were classified as follows: 0-0.2 slight, 0.21-0.4 fair, 0.41-0.6 moderate, 0.61-0.8 substantial, 0.81-1.00 almost perfect agreement.

Results

Twenty-two rheumatologists from 18 European centers entered in both study phases. Out of these, 68.1% (15 rheumatologists) had more than 5 years experience in musculoskeletal US and 38% (8 rheumatologists) more

Table I. Overall agreement for ultrasonographic diagnosis in each region and global assessment.

Pathological finding	Rounds	Overall agreement
Global synovitis	R1	90.2%
	R2	89.3%
Global tenosynovitis	R1	76.4%
	R2	79.6%
Global (synovitis and tenosynovitis)	R1	84.6%
	R2	85.3%

Table II. Global and categorical levels of interobsever agreement.

Pathological finding	Rounds	k	p
Tenosynovitis GS	R1	0.370	P<0.001
	R2	0.374	P<0.001
Tenosynovitis PD	R1	0.460	P<0.001
	R2	0.580	P<0.001
Synovitis GS	R1	0.220	P<0.001
	R2	0.221	P<0.001
Synovitis PD	R1	0.399	P<0.001
	R2	0.580	P<0.001
Synovitis volar	R1	0.388	P<0.05
	R2	0.412	P<0.05
Synovitis dorsal	R1	0.364	P<0.05
	R2	0.319	P<0.05
Global synovitis	R1	0.331	P<0.001
	R2	0.333	P<0.001
Global tenosynovitis	R1	0.432	P<0.001
	R2	0.503	P<0.001
Global (synovitis and tenosynovitis)	R1	0.434	P<0.001
	R2	0.455	P<0.001

Table III. Intraobserver agreement values

Pathological finding	Rounds	k	p
Tenosynovitis GS	R1R2	0.614	P<0.001
Tenosynovitis PD	R1R2	0.689	P<0.001
Synovitis GS	R1R2	0.523	P<0.001
Synovitis PD	R1R2	0.621	P<0.001

than 10 years. All of them had spent an average of 13.5 (SD 8.9) hours per week for US examinations, evaluating on average 29.3 (SD 12.7) patients during that time. From all participants, 70% were also performing the clinical management of the patients evaluated with US.

Overall agreement separately calculated for synovitis and tenosynovitis, together with the global calculation were between 76.4-90.2% (table I).

Interobserver reliability

Table II lists the corresponding k values divided for the 2 rounds of the exercise (R1 and R2). Global and categorical interreader agreements were calculated for synovitis and tenosynovitis, for GS and PD evaluations, for volar/dorsal evaluations. The agreement was higher for PD than for GS and for volar compared to the dorsal examinations. In the second round (R2) all the results demonstrated a higher level of agreement.

Intraobserver agreement

Table III lists k values for intraobserver agreement and their statistical significance. For PD scores the agreement was higher than for GS scores for both joints and tendons evaluations.

Discussion

Poor reliability and high complexity are the main barriers in using US in clinical trials [11]. US reliability for various joint type/pathology was repeatedly tested in the last ten years [3,5,6,12-21], as an attempt to overcome its „operator dependency" and to make it usable for multicentre studies development. The results showed a high degree of variability, depending upon the experience of the ultrasonographer, the machine characteristics, the anatomic region studied, and the pathology type.

As far as we know, our study is the first study that used the Internet for the assessment of inter- and intraobserver agreement between experienced ultrasonographers. The main advantages of the method are the short duration and the easy feasibility when compared to patients-based sessions. In a „Teach the teachers" course, 23 musculoskeletal US experts evaluated joints in 24 rheumatic patients during 16 hours and four consecutive sessions [14]. In our long distance Internet evaluation, each participant reported no more than 1 hour for each phase of the exercise. With one exception (synovitis – dorsal approach) the agreement was higher in R2 than in R1; in addition a better agreement was reached for tenosynovitis than for synovitis in both rounds. Global interreader agreement was moderate (k=0.434/0.455 in R1 and R2) and the PD agreement was constantly higher than GS agreement (0.580 in R2 for synovitis and teno-

synovitis). The lowest values are obtained for synovitis GS evaluation. It should be mentioned that k might be underestimating the agreement because of the effect of „too high" prevalence of the lesions in our study [7-11] (video clips were selected in order to contain a specific type of pathology).

A second point of interest for our study is the reliability for testing the hand tendons' pathology in RA. Other researchers recently addressed this problem [13,15,19] focusing on US reliability of wrist and hand examination compared to ankle and foot for tendons' pathology. In all mentioned studies the results for hand and wrist were in the lower range compared to the ankle. Our values, for both inter- and intraobserver agreement, were moderate to good, in the same range with the above mentioned studies. Tenosynovitis evaluation reliability in our study was higher than synovitis evaluation, probably due to the binar 1/0 grading system for tenosynovitis.

A third important point of interest of our study was the evaluation of the results for dorsal and volar hand US evaluations. Although joint recesses evaluated in different US scoring systems are variable, there are recent indices that US volar scores are better correlated to clinical evaluation than dorsal scores [22,23]. Indeed, in daily practice volar evaluation of MCPs and PIPs seems easier to be performed and interpreted, due to the particular position of flexor tendons toward the joints. Volar evaluations in our study, including MCPs 2-5 and PIPs 2-5 were, for all investigators, showing higher degrees of agreement than dorsal sections, suggesting that US examination could be considered more reliable (and repeatable) when volar incidence is preferred for small hand joints examination.

A limitation of our study is represented by the image acquisition process as all the included videoclips were acquired and recorded by the same investigator. However, this impediment could be overcome in future studies by a very strict standardization of US image acquisition process between US experts, prior to the initiation of the study. High k values for intrareader agreement in our study represent consistent proof for the repeatability of internet based US examination of hand joints, when experienced physicians perform the examination.

The addition of a start-up training session, focused on definitions of grades in synovitis evaluation, could further improve the assessments' results. Other Internet-based reliability studies are required in that area; the use of Internet based cooperation between ultrasonographers facilitates telemedicine implementation.

Conflict of interest: none

References:

1. Backhaus M, Burmester GR, Gerber T, et al. Guidelines for musculoskeletal ultrasound in rheumatology. Ann Rheum Dis 2001; 60: 641-649.
2. Wakefield RJ, Balint P, Szkudlarek M, et al. Musculoskeletal ultrasound including definitions for ultrasonographic pathology. J Rheumatol 2005; 32: 2485-2487.
3. Szkudlarek M, Court-Payen M, Jacobsen S, Klarlund M, Thomsen HS, Ostergaard M. Interobserver agreement in ultrasonography of the finger and toe joints in rheumatoid arthritis. Arthritis Rheum 2003; 48: 955-962.
4. Mandl P, Naredo E, Wakefield RJ, Conaghan PG, D'Agostino MA; OMERACT Ultrasound Task Force. A systematic literature review analysis of Ultrasound joint count and scoring systems to assess synovitis in rheumatoid arthritis according to OMERACT filter. J Rheumatol 2011; 38: 2055-2062.
5. Naredo E, Rodriguez M, Campos C, et al. Validity, reproducibility and responsiveness of a twelve joint simplified power doppler ultrasonographic assessment of joint inflammation in rheumatoid arthritis. Arthritis Rheum 2008; 59: 515-522.
6. Koski JM, Saarakkala S, Helle M, et al. Assessing the intra- and inter-reader reliability of dynamic ultrasound images in power Doppler ultrasonography. Ann Rheum Dis 2006; 65: 1658-1660.
7. Kottner J, Audige L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J Clin Epidemiol 2011 ;64: 96-106.
8. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problem of two paradoxes. J Clin Epidemiol 1990; 43: 543-549.
9. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990; 43: 551-558.
10. Landis JR, Koch CG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-174.
11. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol 2012; 8: 23-34.
12. Dougados M, Jousse-Joulin S, Mistretta F, et al. Evaluation of several ultrasonography scoring systems of synovitis and comparison to clinical examination: results from a prospective multicentre study of rheumatoid arthritis. Ann Rheum Dis 2010; 69: 828-833.
13. Naredo E, D'Agostino MA, Wakefield RJ, et al. Reliability of a consensus-based ultrasound score for tenosynovitis in rheumatoid arthritis. Ann Rheum Dis 2013; 72: 1328-1334.
14. Naredo E, Moller I, Moragues C, et al. Interobserver reliability in musculoskeletal ultrasonography: results from a „Teach the Teachers" rheumatologist course. Ann Rheum Dis 2006; 65: 14-19.
15. Micu M, Serra S, Fodor D, Crespo M, Naredo E. Inter-observer reliability of ultrasound detection of tendon abnormalities at the wrist and ankle in patients with rheumatoid arthritis. Rheumatology (Oxford) 2011; 50: 1120-1124.

16. Scheel AK, Schmidt WA, Hermann KG, et al. Interobserver reliability of rheumatologists performing musculoskeletal ultrasonography: results from a EULAR „Train the trainers" course. Ann Rheum Dis 2005; 64: 1043-1049.
17. Wakefield RJ, D'Agostino MA, Iagnocco A, et al. The OMERACT ultrasound group: status of current activities and research directions. J Rheumatol 2007;34:848-851.
18. Naredo E, Wakefield RJ, Iagnocco A, et al. The OMERACT ultrasound task force- status and perspectives. J Rheumatol 2011; 38: 2063-2067.
19. Bruyn GA, Moller I, Garrido J, et al. Reliability testing of tendon disease using two different scanning methods in patients with rheumatoid arthritis. Rheumatology (Oxford) 2012; 51: 1655-1661.
20. Damjanov N, Radunovic G, Prodanovic S, et al. Construct validity and reliability of ultrasound disease activity score in assessing joint inflammation in RA: comparison with DAS-28. Rheumatology (Oxford) 2012; 51: 120-128.
21. Iagnocco A, Perricone C, Scirocco C, et al. The interobserver reliability of ultrasound in knee osteoarthritis. Rheumatology (Oxford) 2012; 51: 2013-2019.
22. Scheel AK, Hermann KG, Kahler E, et al. A Novel ultrasonographic synovitis scoring system suitable for analyzing finger joint inflammation in rheumatoid arthritis. Arthritis Rheum 2005; 52: 733–743.
23. Vlad V, Berghea F, Libianu S, et al. Ultrasound in rheumatoid arthritis: volar versus dorsal synovitis evaluation and scoring. BMC Musculoskelet Disord 2011; 12: 124.