Assessing Elementary Lesions in Gout by Ultrasound: Results of an OMERACT Patient-based Agreement and Reliability Exercise

Lene Terslev, Marwin Gutierrez, Robin Christensen, Peter V. Balint, George A. Bruyn, Andrea Delle Sedie, Emilio Filippucci, Jesus Garrido, Hilde B. Hammer, Annamaria Iagnocco, David Kane, Gurjit S. Kaeley, Helen Keen, Peter Mandl, Esperanza Naredo, Carlos Pineda, Bernd Schicke, Ralf Thiele, Maria Antonietta D'Agostino and Wolfgang A. Schmidt, on behalf of the OMERACT US Gout Task Force

1. Sign up for our monthly e-table of contents
   http://www.jrheum.org/cgi/alerts/etoc

2. Information on Subscriptions
   http://jrheum.com/subscribe.html

3. Have us contact your library about access options
   Refer_your_library@jrheum.com

4. Information on permissions/orders of reprints
   http://jrheum.com/reprints.html

The Journal of Rheumatology is a monthly international serial edited by Earl D. Silverman featuring research articles on clinical subjects from scientists working in rheumatology and related fields.

Assessing Elementary Lesions in Gout by Ultrasound: Results of an OMERACT Patient-based Agreement and Reliability Exercise

Lene Terslev, Marwin Gutierrez, Robin Christensen, Peter V. Balint, George A. Bruyn, Andrea Delle Sedie, Emilio Filippucci, Jesus Garrido, Hilde B. Hammer, Annamaria Iagnocco, David Kane, Gurjit S. Kaeley, Helen Keen, Peter Mandl, Esperanza Naredo, Carlos Pineda, Bernd Schicke, Ralf Thiele, Maria Antonietta D'Agostino, and Wolfgang A. Schmidt, on behalf of the OMERACT US Gout Task Force

ABSTRACT. *Objective.* To test the reliability of the consensus-based ultrasound (US) definitions of elementary gout lesions in patients.

Methods. Eight patients with microscopically proven gout were evaluated by 16 sonographers for signs of double contour (DC), aggregates, erosions, and tophi in the first metatarsophalangeal joint and the knee bilaterally. The patients were examined twice using B-mode US to test agreement and inter- and intraobserver reliability of the elementary components.

Results. The prevalence of the lesions were DC 52.8%, tophus 61.1%, aggregates 29.8%, and erosions 32.4%. The intraobserver reliability was good for all lesions except DC, where it was moderate. The best reliability per lesion was seen for tophus (κ 0.73, 95% CI 0.61–0.85) and lowest for DC (κ 0.53, 95% CI 0.38–0.67). The interobserver reliability was good for tophus and erosions, but fair to moderate for aggregates and DC, respectively. The best reliability was seen for erosions (κ 0.74, 95% CI 0.65–0.81) and lowest for aggregates (κ 0.21, 95% CI 0.04–0.37).

Conclusion. This is the first step to test consensus-based US definitions on elementary lesions in patients with gout. High intraobserver reliability was found when applying the definition in patients on all elementary lesions while interobserver reliability was moderate to low. Further studies are needed to improve the interobserver reliability, particularly for DC and aggregates. (First Release October 15, 2015; J Rheumatol 2015;42:2149–54; doi:10.3899/jrheum.150366)

Key Indexing Words:
| ULTRASONOGRAPHY | GOUT | EXERCISE | OMERACT | RELIABILITY |

Ambroise Paré, INSERM U1173, Laboratoire d'Excellence INFLAMEX, and Université de Versailles Saint-Quentin-en-Yvelines; W.A. Schmidt, MD, Immanuel Krankenhaus Berlin, Medical Center for Rheumatology.

Address correspondence to Dr. L. Terslev, Center for Rheumatology and Spine Diseases, Rigshospitalet-Glostrup, Copenhagen University Hospital, Nordre Ringvej 57, Glostrup, 2650, Denmark.
E-mail: Lene.terslev.01@regionh.dk

Accepted for publication July 23, 2015.

Gout is a common inflammatory joint disease and is caused by the formation and deposition of monosodium urate crystals (MSU) in joints or soft tissues. The diagnosis of gout is conventionally based on the history, clinical examination, uric acid serum levels, and subsequent polarization microscopy of joint or tophus aspirates. Polarization microscopy is the definitive way to diagnose gout. In suspected patients with gout without synovial effusion or clinical tophi, the sampling of relevant material to examine for MSU is challenging, and other diagnostic procedures are warranted.

Uncontrolled hyperuricemia may be associated with chronic kidney disease and cardiovascular disease with subsequent increased morbidity and mortality[1]. This has underlined the importance for an early, accurate diagnosis of gout, and with the development of new therapeutic options, imaging modalities have been investigated to determine whether they may improve disease assessment. In this respect, ultrasound (US) has been shown to be promising because it allows direct visualization of the crystal deposits and is also increasingly available in the clinical setting. Although several published studies have highlighted the involvement of US for the assessment of elementary components in gout, a systematic literature review emphasized the lack of definitions of US elementary lesions[2]. To improve the use of US in the evaluation of gout, an Outcome Measures in Rheumatology Clinical Trials (OMERACT) task force subgroup was formed.

The first step in the standardization process of US as a tool for diagnosis and monitoring of gout was to obtain consensus-based definitions on US elementary lesions, as revealed by the systematic literature review[2] to be double contour (DC), aggregates, tophi, and erosions. The consensus-based definitions were obtained through a Delphi exercise and tested in a subsequent Web exercise[3]. In the latter, a good to excellent reliability was found for all lesions except for aggregates, which was moderate when testing the definitions on static images[3]. The second step in the standardization process and the aim of our present study was to test the reliability of the consensus-based definitions in known patients with gout to ensure the pathologies to be present. This second step is mandatory before it is possible to test the US ability as a diagnostic tool.

MATERIALS AND METHODS

Study design and setting. Our present study was performed according to a prespecified protocol. The reporting of the OMERACT reliability exercise followed the recommendations from the Enhancing the QUAlity and Transparency Of health Research network[4] using the Guidelines for Reporting Reliability and Agreement Studies statement[5].

Ethics committee approval of our study was obtained from the Berlin Medical Association (Berliner Ärztekammer, Eth-17/13). All patients gave informed consent. Following the consensus-based definitions obtained through a Delphi and Web exercise process, a workshop was conducted to evaluate the reliability of detecting the elementary components in patients and assessing agreement between sonographers with experience in US of gout.

Measurements. The following definitions of the US elementary lesion obtained from step 1[3] were tested in the patients with gout: (1) Double contour: "Abnormal hyperechoic band over the superficial margin of the articular hyaline cartilage, independent of the angle of insonation and which may be either irregular or regular, continuous or intermittent and can be distinguished from the cartilage interface sign". (2) Tophus [independent of location (e.g., extra-articular/intra-articular/intra-tendinous)]: "A circumscribed, inhomogeneous, hyperechoic and/or hypoechoic aggregation (which may or may not generate posterior acoustic shadow) which may be surrounded by a small anechoic rim". (3) Aggregates [independent of location (intra-articular/intra-tendinous)]: "Heterogeneous hyperechoic foci that maintain their high degree of reflectivity even when the gain setting is minimized or the insonation angle is changed and which occasionally may generate posterior acoustic shadow". (4) Erosions: "An intra- and/or extra-articular discontinuity of the bone surface (visible in 2 perpendicular planes)"[3]. Examples can be seen in Figure 1.

Patients. To ensure that all elementary lesions may be detectable, 8 patients with polyarticular, tophaceous gout, verified by polarization microscopy from the Medical Center of Rheumatology Berlin-Buch, volunteered for our study. Their clinical data may be seen in Table 1. Each patient underwent a bilateral B-mode US examination (including dynamic examination) of the first metatarsophalangeal (MTP) joints, the cartilage of the intercondylar region of the knees (ICR), and the proximal (PPT) and distal (DPT) parts of the patella tendons. The MTP joints were examined dorsally, from medially to laterally. The tendons were also examined from the medial to the lateral aspect and at the enthesis where signs of erosive changes were also evaluated. The lower extremities were chosen for feasibility reasons. All examinations were performed during a morning and an afternoon session on the same day, in the same room, using 8 Esaote MyLab Twice/Class machines, equipped with 6-18 MHz broadband linear array transducers. The B-mode settings of all 8 US machines were identical.

Ultrasonographers. Sixteen of the rheumatologists previously involved in the development of the consensus-based US definitions for gout participated in the workshop. All rheumatologists were experienced in musculoskeletal US.

Outcomes and rating process. Each patient (numbered 1–8) was assigned to 1 machine (numbered 1–8), and the ultrasonographers (numbered 1–16) then moved from 1 patient to the next in a predefined (randomized) sequence with 13 min allocated for scanning and scoring the findings. All patients were scanned twice by the same examiner to assess the intrareader reliability. The data were collected immediately after the session to ensure no communication between examiners.

Statistical analysis. Intra- and interobserver reliability was estimated based on Cohen κ coefficient. These κ coefficients and the corresponding 95% CI were interpreted according to Landis and Koch[6]: κ values of 0–0.20 were considered poor, 0.20–0.40 fair, 0.40–0.60 moderate, 0.60–0.80 good, and 0.80–1 excellent. Percentage of observed agreement (i.e., percentage of observations that obtained the same score) and prevalence of the observed lesions were also calculated.

Because our study used a hierarchical design, with repeated measures across 8 patients and 16 rheumatologists, Cohen κ for measuring agreement between 2 raters had to be extended according to Light and Fleiss for use with multiple raters, as well as repeated measures within patients[7,8]. Thus, we used a crossed design in which all raters (1–16) evaluated all patients (8)
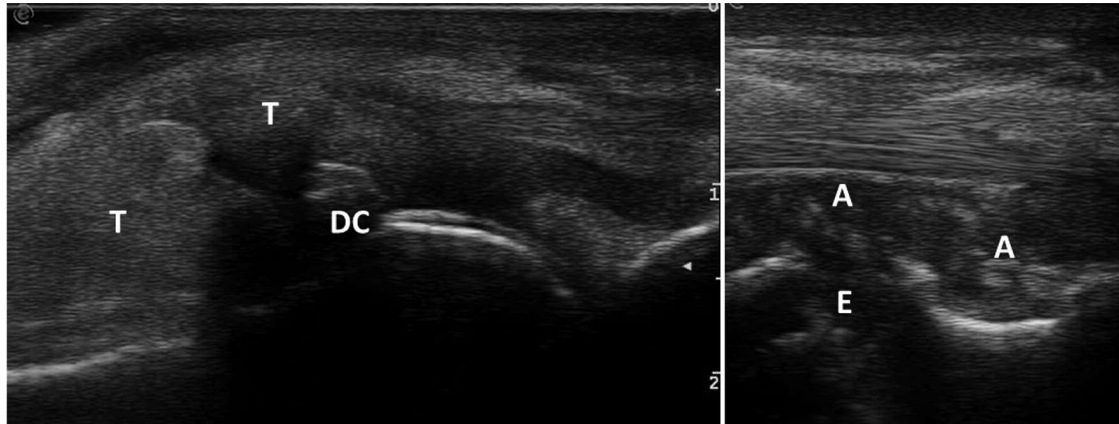
Figure 1. Elementary US lesions of gout. US: ultrasound; T: tophus; DC: double contour; E: erosion; A: aggregate.

Table 1. Clinical data of the examined patients.

Patient No.	Age, Yrs	Disease Onset, Yrs	Rheumatology Care, Mos	Clinically Involved Joint Regions	Maximum Known Urate Level, mg/dl	Current Urate Level, mg/dl	Current Treatment, mg/d
1	67	11	11	T, A, K, D, W, E	10.6	5.9	F 120, Co 1
2	74	5	2	T, K, D	5.8*	5.8	A 300, Co 1
3	65	19	1	T, A, K, D, W, E	8.2	4.9	F 80, Co 0.5
4	70	21	18	T, A, K, D, W, E	6.6*	4.8	A 300
5	72	7	31	T, A, K, E	12.4	3.8	F 120, Co 1
6	69	20	11	T, A, K	9.0	5.4	A 400, Co 0.5
7	68	28	2	T, A, K, D, W, E	8.1	5.5	F 120, Ca**
8	48	6	3	T, A, K, D, W, E	6.4*	3.5	F 120, P 1000, PRED 15

* Uric acid levels only available at time of gout attack and/or with uric acid–lowering therapy. ** The patient has received a subcutaneous injection of 150 mg of Ca 6 weeks before. T: toes; A: ankles; K: knees; D: digits; W: wrists; E: elbows; A: allopurinol; Ca: canakinumab; Co: colchicine; F: febuxostat; P: probenicid; PRED: prednisolone.

in duplicates (2 tests), in all anatomical positions (4 positions: DPT, ICR, MTP1, and PPT), and on both sides of the body (right, left). The outcomes were the facet of differentiation, and they were nested in patient and rheumatologist. For intraobserver reliability, to adjust for the "clustered data" we inflated the standard error (SE) of the κ estimate (tophus, aggregates, and erosions) by multiplying the unadjusted SE with 2.828 (i.e., 8 × 16 independent observations rather than the apparent 1024 clustered observations); for DC, the unadjusted SE was multiplied by 2.000 (i.e., 8 × 16 independent observations rather than the apparent 512 clustered observations). Interobserver reliability was assessed by Light's κ (mean κ across all pairs of interobservers) based on the second test scenario[8].

RESULTS

All patients were men with a mean age of 67 years (range 48–74) and all received urate-lowering therapy (Table 1). The prevalence of lesions are listed in Table 2, showing that DC and tophus are most frequently observed (> 50%), especially in the ICR and MTP1, respectively, while aggregates and erosions were observed in < 30% of the anatomical areas and most frequently in the MTP1 joint for both.

Table 3 summarizes the results of the reliability of US elementary lesions in patients. The intraobserver agreement (absolute terms) was for tophi (87%, 891/1024), aggregates (84%, 857/1024), erosions (87%, 889/1024), and DC (76%, 391/512). Further, the intra- and interobserver κ values are shown. For the intraobserver reliability, the best reliability per lesion was seen for tophus (κ 0.73, 95% CI 0.61–0.85) and lowest for DC (κ 0.53, 95% CI 0.38–0.67). The results for the interobserver reliability per lesion were best for erosions (κ 0.74, 95% CI 0.65–0.81) and lowest for aggregates (κ 0.21, 95% CI 0.04–0.37).

DISCUSSION

Because of the introduction of new treatment options for both acute and chronic gout, research activities have focused on developing validated outcome measures to evaluate treatment effects. This has been the case, in particular, with regard to tophus regression and joint inflammation, including joint

Table 2. Prevalence of observed gout lesions in the first and second round. Both overall prevalence per lesion and prevalence of the lesions per site are given.

US Gout Lesions	Observed Prevalence, %	Positive Findings, n	Total Observed Sites, n
Tophus			
Overall	61.1	1251	2048
PPT	66.4	340	512
DPT	68.6	351	512
MTP1	89.5	458	512
ICR	19.9	102	512
Aggregates			
Overall	29.8	611	2048
PPT	33.6	172	512
DPT	28.9	148	512
MTP1	45.5	233	512
ICR	11.3	58	512
Erosions			
Overall	32.4	663	2048
PPT	21.7	111	512
DPT	16.6	85	512
MTP1	85.0	435	512
ICR	6.3	32	512
DC			
Overall	52.8	541	1024
MTP1	49.4	253	512
ICR	56.3	288	512

US: ultrasound; PPT: proximal patellar tendon; DPT: distal patellar tendon; MTP1: metatarsophalangeal joint 1; ICR: intercondylar region; DC: double contour.

Table 3. Overall intraobserver agreement and the intra- and interobserver reliability estimates for each of the elementary gout lesions.

US Gout Elementary Lesion	Overall Intraobserver Agreement, %	Intraobserver Reliability, κ (95% CI)	Interobserver Reliability, κ (95% CI)
Tophus	87	0.727 (0.605–0.849)	0.691 (0.589–0.771)
Aggregates	84	0.611 (0.460–0.761)	0.209 (0.039–0.368)
Erosions	87	0.699 (0.566–0.832)	0.737 (0.647–0.807)
Double contour	76	0.526 (0.379–0.673)	0.469 (0.323–0.593)

US: ultrasound.

swelling[9]. For both chronic and acute gout, the suggested variables in the OMERACT core domain set mentioned above may be evaluated by US. To validate US as a possible outcome instrument, we set out to standardize US by first defining the US elementary lesions in gout. After the initial step of defining the elementary lesions and test these on static images of well-illustrated lesions[3], the next step was to test the reliability of the definitions in known patients with gout, adding image acquisition to the exercise.

The intraobserver agreement was found to be good for all lesions and moderate for the DC. The interobserver reliability was fair for aggregates (0.21) and moderate for DC (0.47), and good for the other components. These findings are in line with previous studies on reliability for US lesions and regions[10,11], also in gout[12]. Aggregates are heterogeneously

described in the literature[2] and are believed to be deposits of crystals in the soft tissues not large enough to be defined as a tophus[3], while DC is created by the deposits of crystals on the surface of the cartilage[13] and may be detected by US in up to 60% of joints, including asymptomatic joints in patients with gout[14,15,16]. In our study, the overall DC prevalence was 53% (Table 2). The US definition obtained in the Delphi for aggregates is less specific than the other definitions[3] and was proposed to define soft tissue hyperechogenicity often seen in patients with gout[2]. Though the good intraobserver agreement demonstrates that the individual sonographers know what to score, they appear to have different perceptions of the definition because the interobserver agreement is only fair. Even when looking at the individual sites, there is no joint or tendon site that has better reliability for detecting aggregates.

The poor performance of the aggregate definition may also partly be related to statistical factors. The prevalence of aggregates was low overall, and this may explain the fair κ. When the prevalence is very low or very high, low κ values are obtained even though the overall agreements are high (paradox κ)[17]. This is because with high prevalence, the agreement expected by chance is very high and κ is the remaining agreement obtained after discounting agreement by chance. This may also partly explain the lower site agreement for DC.

DC is perceived to be indicative of gout[18]. The reliability for DC was moderate for both intra- and interobserver agreement, which is in contrast to the interobserver reliability of the DC in previous publications where it was found to be excellent in local research groups[14,15,18,19]. This may partly be related to more exercise time on pathological features in small study groups and may indicate that more scanning time together in a group is necessary. Further, the US definition for DC is more detailed and the discrepancy between the interobserver agreement found in the Web exercise and that found in the patient workshop may in part be related to the image acquisition and scanning technique. In static image Web exercises, only clear images of the pathologies are chosen, which might not always resemble the clinical setup with patients. Another possible pitfall might be that the DC may resemble cartilage interface, which can only be seen in an area where the insonation angle is 90° and appears as a white line on the surface of the cartilage. The DC is seen as a white line — punctuate or linear — on the surface of the cartilage, and also in areas where the insonation angle is < 90°, and will move with the cartilage during dynamic exercise as the crystals are deposited on the surface of the cartilage. Though this differentiation between cartilage interface sign and DC is well known, it may have had an effect on the findings during image acquisition in patients — especially in the ICR where the presence of even a minimal effusion makes the cartilage interface sign more frequent. Further studies are needed to evaluate the optimal site for cartilage pathology.

Not surprisingly, the reliability of erosions was good for both intra- and interreader reliability, which can be explained by a longstanding experience in the group scanning the pathology from rheumatoid arthritis (RA). The exercise also demonstrated a good intra- and interreader reliability for tophus, which is an aspect of the urate burden in patients with gout. However, aggregates were the lesions with the lowest reliability both in the static image exercise[3], but even lower in patients, especially between observers. This raises the issue of whether the definition is truly covering the US lesions it is supposed to describe (soft tissue hyperechogenicity). Because a tophus is also a collection of aggregates, there is a risk that these 2 elementary lesions are overlapping and impairing the agreement between observers. Further steps are needed to improve the US definitions for aggregates before

it may be an elementary US lesion in gout as part of future outcome measures.

For DC, the reliability was moderate and further studies are needed to improve the reliability by focusing on MTP joint cartilage before the definitions may be used in multi-center studies and before testing the diagnostic sensitivity of US in gout[20].

Future steps that include collaboration with histologists may be beneficial to improve the definitions followed by Web exercises with focus specifically on these 2 features (aggregates and DC) in images resembling daily clinical situations. Patient workshops are needed to develop optimal image acquisition techniques, and finally the development of an US atlas of pathological lesions may aid in the reliability process because this has been found to highly increase reliability for scoring synovitis in RA between sonographers[21].

Consensus-based US definitions of elementary lesions in gout were tested in patients and showed high intraobserver reliability, but lower interobserver reliability — especially for aggregates. DC had lower interreader reliability than tophus and erosions. Further studies are needed to explore the US definition of aggregates for improving the reliability between sonographers.

APPENDIX 1.

ACKNOWLEDGMENT

REFERENCES

1. Richette P, Perez-Ruiz F, Doherty M, Jansen TL, Nuki G, Pascual E, et al. Improving cardiovascular and renal outcomes in gout: what should we target? Nat Rev Rheumatol 2014;10:654-61.
2. Chowalloor PV, Keen HI. A systematic review of ultrasonography in gout and asymptomatic hyperuricaemia. Ann Rheum Dis 2013;72:638-45.
3. Gutierrez M, Schmidt WA, Thiele RG, Keen HI, Kaeley GS, Naredo E, et al; OMERACT Ultrasound Gout Task Force group. International Consensus for ultrasound lesions in gout: results of Delphi process and web-reliability exercise. Rheumatology 2015; May 13 (E-pub ahead of print).
4. Christensen R, Bliddal H, Henriksen M. Enhancing the reporting and transparency of rheumatology research: a guide to reporting guidelines. Arthritis Res Ther 2013;15:109.
5. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol 2011;64:96-106.
6. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
7. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76:378-82.

8. Light RJ. Measures of response agreement for qualitative data: some generalizations and alternatives. Psychol Bull 1971;76:365-77.

9. Taylor WJ, Schumacher HR Jr, Baraf HS, Chapman P, Stamp L, Doherty M, et al. A modified Delphi exercise to determine the extent of consensus with OMERACT outcome domains for studies of acute and chronic gout. Ann Rheum Dis 2008;67:888-91.

10. Scheel AK, Hermann KG, Kahler E, Pasewaldt D, Fritz J, Hamm B, et al. A novel ultrasonographic synovitis scoring system suitable for analyzing finger joint inflammation in rheumatoid arthritis. Arthritis Rheum 2005;52:733-43.

11. Naredo E, Möller I, Moragues C, de Agustín JJ, Scheel AK, Grassi W, et al; EULAR Working Group for Musculoskeletal Ultrasound. Interobserver reliability in musculoskeletal ultrasonography: results from a "Teach the Teachers" rheumatologist course. Ann Rheum Dis 2006;65:14-9.

12. Lamers-Karnebeek FB, Van Riel PL, Jansen TL. Additive value for ultrasonographic signal in a screening algorithm for patients presenting with acute mono-/oligoarthritis in whom gout is suspected. Clin Rheumatol 2014;33:555-9.

13. Filippucci E, Scirè CA, Delle Sedie A, Iagnocco A, Riente L, Meenagh G, et al. Ultrasound imaging for the rheumatologist. XXV. Sonographic assessment of the knee in patients with gout and calcium pyrophosphate deposition disease. Clin Exp Rheumatol 2010;28:2-5.

14. Howard RG, Pillinger MH, Gyftopoulos S, Thiele RG, Swearingen CJ, Samuels J. Reproducibility of musculoskeletal ultrasound for determining monosodium urate deposition: concordance between readers. Arthritis Care Res 2011;63:1456-62.

15. Ottaviani S, Allard A, Bardin T, Richette P. An exploratory ultrasound study of early gout. Clin Exp Rheumatol 2011; 29:816-21.

16. Huppertz A, Hermann KG, Diekhoff T, Wagner M, Hamm B, Schmidt WA. Systemic staging for urate crystal deposits with dual-energy CT and ultrasound in patients with suspected gout. Rheumatol Int 2014;34:763-71.

17. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43:543-9.

18. Filippucci E, Riveros MG, Georgescu D, Salaffi F, Grassi W. Hyaline cartilage involvement in patients with gout and calcium pyrophosphate deposition disease. An ultrasound study. Osteoarthritis Cartilage 2009;17:178-81.

19. Naredo E, Uson J, Jiménez-Palop M, Martínez A, Vicente E, Brito E, et al. Ultrasound-detected musculoskeletal urate crystal deposition: which joints and what findings should be assessed for diagnosing gout? Ann Rheum Dis 2014;73:1522-8.

20. Ogdie A, Taylor WJ, Weatherall M, Fransen J, Jansen TL, Neogi T, et al. Imaging modalities for the classification of gout: systematic literature review and meta-analysis. Ann Rheum Dis 2014; Jun 10 (E-pub ahead of print).

21. Hammer HB, Bolton-King P, Bakkeheim V, Berg TH, Sundt E, Kongtorp AK, et al. Examination of intra and interrater reliability with a new ultrasonographic reference atlas for scoring of synovitis in patients with rheumatoid arthritis. Ann Rheum Dis 2011; 70:1995-8.