

BMJ Open

Inter-rater reliability of ultrasound assessment for grading structural tendon changes in supraspinatus tendinopathy

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2016-011746
Article Type:	Research
Date Submitted by the Author:	07-Mar-2016
Complete List of Authors:	Ingwersen, Kim; Hospital Lillebalt, Department of Physiotherapy; University of Southern Denmark, Department of Sports Science and Clinical Biomechanics Hjarbaek, John; Odense University Hospital, Department of Radiology, Musculoskeletal section Eshoej, Henrik; University of Southern Denmark, Department of Sports Science and Clinical Biomechanics Larsen, Camilla; University of Southern Denmark, Department of Sports Science and Clinical Biomechanics; University College Lillebaelt - Campus Odense, Health Sciences Research Centre Vobbe, Jette; Hospital Lillebaelt, Orthopedic Department, Shoulder Unit Juul-Kristensen, Birgit; University of Southern Denmark, Department of Sports Science and Clinical Biomechanics; Hogskolen i Bergen, Institute of Occupational Therapy, Physiotherapy and Radiography, Department of Health Sciences
Primary Subject Heading:	Radiology and imaging
Secondary Subject Heading:	Diagnostics, Rehabilitation medicine, Sports and exercise medicine
Keywords:	Reliability, Tendinopathy, Ultrasound < RADIOLOGY & IMAGING, ULTRASONOGRAPHY, Shoulder < ORTHOPAEDIC & TRAUMA SURGERY

SCHOLARONE™
Manuscripts

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Inter-rater reliability of ultrasound assessment for grading structural tendon changes in supraspinatus tendinopathy

Kim Gordon Ingwersen^{1,2}, John Hjarbaek³, Henrik Eshøj¹, Camilla Marie Larsen^{1,4}, Jette Vobbe⁵,
Birgit Juul-Kristensen^{1,6}

¹Department of Sports Science and Clinical Biomechanics, University of Southern Denmark, Odense, Denmark.

²Department of Rehabilitation, Hospital Lillebaelt - Vejle Hospital, Vejle, Denmark

³Department of Radiology, Musculoskeletal section, Odense University Hospital, Odense, Denmark

⁴Health Sciences Research Centre, University College Lillebaelt, Denmark

⁵Shoulder Unit, Orthopaedic Department, Hospital Lillebaelt, Vejle Hospital, Vejle, Denmark

⁶Institute of Occupational Therapy, Physiotherapy and Radiography, Department of Health Sciences, Bergen University College, Bergen, Norway

Corresponding author: Kim Gordon Ingwersen, Department of Rehabilitation - Hospital Lillebaelt, Kappeltoft 25, DK-7100, Vejle. Tlf.nr: +45 79 40 61 75, E-mail: kim.riis@rsyd.dk.

Keywords: Reliability; Tendinopathy; Ultrasound; Sonography; Shoulder.

Word count: 2826

Inter-rater reliability of ultrasound assessment for grading structural tendon changes in supraspinatus tendinopathy

ABSTRACT

Aim. To evaluate the inter-rater reliability of measuring structural changes in the tendon of patients, clinically diagnosed with supraspinatus tendinopathy (Cases) and healthy participants (Controls), on ultrasound (US) images captured by standardized procedures

Methods. A total of 70 participants (39 patients) were included for assessing inter-rater reliability of measurements of fibrillar disruption, neovascularity, number and total length of calcifications and tendon thickness. Linear weighted kappa, Intra Class Correlation (ICC), Standard Error of Measurement (SEM), Limits Of Agreement (LOA) and Minimal Detectable Change (MDC) were used to evaluate reliability.

Results. “Moderate - Almost perfect” kappa was found for grading fibrillar disruption, neovascularity and number of calcifications ($k: 0.60 - 0.96$). For total length of calcifications and tendon thickness ICC was “Excellent” ($0.85 - 0.90$), with $SEM_{(Agreement)}$ ranging from $0.63 - 2.94\text{mm}$ and $MDC_{(group)}$ ranging from $0.28 - 1.29\text{mm}$. In general, SEM, LOA and MDC showed larger variation for calcifications than for tendon thickness.

Conclusion. Inter-rater reliability was moderate to almost perfect, when a standardized procedure was applied for measuring structural changes on captured US images and movie sequences of relevance for patients with supraspinatus tendinopathy. Future studies should test intra- and inter-rater reliability of the method in vivo for use in clinical practice, in addition to validation against a gold standard, such as MRI.

STRENGTHS AND LIMITATIONS OF THIS STUDY:

- A standardized procedure for performance of US of the supraspinatus is presented
- A specific procedure for grading and measuring tendinopathy related changes is presented
- Grading and measurement is possible to be performed reliable
- Performance of the method in vivo are warranted to validate the method to clinical practice

INTRODUCTION

Rotator Cuff (RC) tendinopathy can be considered a continuum of pathology, and tailored rehabilitation according to the stage in this continuum is recommended.^{1 2} Anamnesis and special orthopaedic tests are often used when diagnosing RC tendinopathy, but these tests often lack high specificity and sensitivity, making diagnosis uncertain,³ thus challenging precise and targeted treatment.

Grey-Scale (GS) ultrasound (US) and Power Doppler (PD) visualization of RC tendons may be helpful to detect signs of tendinopathy, such as hypoechoic areas, fibrillar disruption, neovascularisation, calcifications embedded in the tendon or odema and confirm the “a priori” hypothesis of RC tendinopathy, provided satisfactory clinimetric properties of the US method.⁴ However, US is an operator dependent technique and requires thorough training and experience in performance and assessment before precise diagnoses can be made, especially in relation to more subtle changes as often seen within tendinopathy.⁵ Poor to fair reliability has previously been found when comparing diagnoses made by US novel and experienced clinicians.⁶⁻⁸ Further, when grading subtle structural tendon changes, especially hypoechoic areas, only fair, and thus unsatisfactory reliability has been found, even among experienced clinicians.^{5 7 9-11}

Standardised procedures for capturing and assessing US is known to increase reliability of US based diagnoses.⁵ Previously, assessment of tendinopathy were found reliable, in patients with tendinopathy in the elbow, ankle or knee, when using standardised procedures for measuring GS and PD.^{4 10}

For the shoulder, however, there is lack of clinically relevant, standardized and reliable methods for assessing tendinopathy. Since US is highly influenced by clinician experience and technique, both standardized US procedures for image and movie capturing, and standardized procedures for assessment of structural changes in relation to tendinopathy need to be defined.

Therefore, the aim of this study was from standardized US procedures for image and movie capturing, to evaluate the inter-rater reliability of measuring and grading structural changes in the

tendon of patients clinically diagnosed with supraspinatus tendinopathy (Cases) and healthy participants (Controls).

MATERIAL AND METHODS

Study design

The study followed the protocol for diagnostic procedures in reproducibility studies.¹² This protocol includes a three-phase study design consisting of a 1) training; 2) an overall agreement and 3) a study phase (the actual reliability study) (Figure 1).

The phases constitute a methodological model for optimizing procedures, and aim at eliminating clinician subjectivity as much as possible. The aim of the training phase is to secure that raters have sufficient competence and experience in performing the procedures. The overall agreement phase is an extended training phase and secures that gross systematic bias between raters are minimized, and requires at least 80% agreement between raters, before proceeding to phase 3. The study phase, is the final evaluation of reliability of the developed procedures.¹²

Inter-rater reliability (phase 3) between two raters (rater A and B) was tested on measuring and grading structural changes relevant to tendinopathy upon US captured images and movies. Rater A (KI; Physiotherapist) had one year of clinical musculoskeletal US experience, and rater B (JH; Radiologist) had more than fifteen years of clinical musculoskeletal US experience.

US image capturing and measurement

Based upon the literature,^{4 9 10 13-15} consensus was made upon definitions of relevant potential pathological structural changes related to tendinopathy, including 1) *fibrillar disruption (FD)*, 2) *neovascularisation (NV)*, 3) *calcification (CA)*, and 4) *tendon thickness (TT)*. Hereafter, a standardized protocol for US capturing was developed, consisting of three static images (grey-scale), three dynamic movie sequences (grey-scale), and one Doppler movie sequence (*Table 1*).

Table 1: Description of US procedures for capturing image and movie sequences of Fibrillar disruption (FD), Neovascularity (NV), Calcifications (CA) and Tendon Thickness (TT)

1) Fibrillar disruption (FD):

FD was defined as a clear collagen fascicle discontinuity or irregularity of fibrils in an otherwise regular parallel structuring of fibres in the tendon.

A GS picture in the longitudinal axis of the supraspinatus was taken at the sight where FD was most apparent (FD picture). The FD static image was used for classifying presence of FD. A GS posterior-anterior dynamic movie sequence (PA movie) in the longitudinal plane of the supraspinatus tendon was captured, by moving the transducer slowly in the posterior-anterior direction. Further, a caudal-cranial (CC) transversal GS dynamic movie sequence (CC movie) of the supraspinatus tendon was recorded by moving the transducer slowly in the CC direction. The static image and the movie sequence recordings were used as confirmation and assistance in assessing the grade of structural changes, and to secure identification of potential ambiguous GS features, such as anisotropy (erroneous signal caused when the transducer is angled obliquely to the tendon).

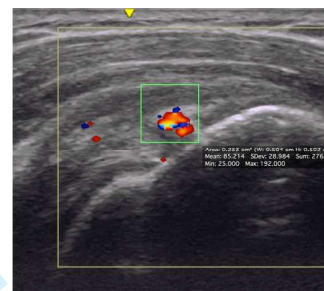
FD was classified in relation to tendon thickness as: 0=normal, 1=mild, 2=moderate, 3=severe, 4=partial rupture (table 2).



Grade 2 FD

2) Neovascularity (NV):

NV was defined as a visualized Power Doppler (PD) signal with minimal artifactual noise. The supraspinatus tendon was evaluated for presence of NV by moving the transducer slowly in the posterior-anterior direction, with the PD feature activated. In case NV was present, a 10 sec dynamic movie sequence was recorded at the point with most NV signal (PD movie). When grading NV from the PD movie sequence, a static image of the location with the most visible NV was captured from the PD movie. A Region of Interest (ROI) (5x5 millimetre (mm)) was placed around the NV and used for grading NV. In participants with no NV a movie sequence was recorded at a random location in the tendon to verify absence of NV. NV was classified in relation to ROI as 0=normal, 1=mild, 2=moderate, 3=severe, 4=Extreme (Table 2).



Grade 2 NV

3) Calcification (CA):

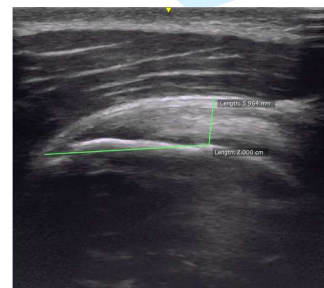
CA was defined as distinct white borders, imbedded in the length of the tendon, often with “shadows” underneath. The PA movie was used to identify the number of CA in the tendon and to measure the length of each CA in the longitudinal axis of the supraspinatus. The length was measured between the most medial and lateral aspect of the distinct white boarder (in mm). The CC movie was used as confirmation and assistance in identifying CA. CA was counted and measured (mm). To obtain the total length of CA, the individual CA lengths were added up to one total per participant.



CA length measure

4) Tendon thickness (TT):

TT was defined as the height, from the humeral head, at a point 20 mm from the supraspinatus tendon-snip (tendon insertion) in the longitudinal axis of the tendon to the most superficial part of the tendon. In practice, an image was captured at a fixed point just laterally from the anterior-lateral corner of the acromion in the longitudinal plane of the supraspinatus. When measuring TT, a mark was placed 20 mm cranial from the supraspinatus tendon-snip (tendon insertion), on the edge of the cartilage of the humeral head. From that mark, the perpendicular thickness of tendon was recorded.¹⁶ The TT picture was recorded bilaterally. TT was measured in mm.



TT measure

Secondly, based upon previous scales used to measure structural changes in tendinopathy at the elbow,^{4 15} two ordinal grading scales for FD and NV were adjusted for use in the shoulder.¹⁷ The scales ranged from 0-4 (FD: 0=Normal tendon; 4=Partial rupture, corresponding to disruption of the fibers in the full thickness of the tendon; NV (0=Normal, including no signal; 4= Extreme, including Doppler activity in more than 50% of the region of interest, ROI) (Table 2).

Table 2. Grading scales with definitions for fibrillar disruption (FD) and neovascularity (NV)

Grade	Fibrillar disruption	Neovascularity
0	Normal	Normal (No signal)
1	Mild (Involving under 25% of the height of the tendon)	Mild (Single small signal in the Region Of Interest, ROI)
2	Moderate (Involving 25-50% of the height of the tendon)	Moderate (Doppler activity in less then 25% of the ROI)
3	Severe (Involving more than 50% of the height of the tendon)	Severe (Doppler activity in 25-50% of the ROI)
4	Partial rupture (Disruption of the fibers in the full thickness of the tendon)	Extreme (Doppler activity in more then 50% of the ROI)

Abbreviations: ROI: Region of interest

CA was analysed as number of calcifications and total length (in mm), while TT was measured in mm.¹⁶

Rater A performed capturing of all US images and movie sequences with the participant seated, the shoulder internally rotated with the dorsal side of the hand placed on the sacrum, and the elbow flexed and directed laterally, to optimize visualisation of the supraspinatus tendon.¹⁸

A GE LOGIQ e B12 (GE Healthcare, Wisconsin, USA) with a 5.0 – 13.0 MHz linear transducer was used for image capturing. All US scannings were standardized and performed for GS imaging at 13.0 MHz and 56% gain, while PD scanning was performed with a pulse repetition frequency of 0.41 kHz and gain at 56%. Manufacturer recommendations for musculoskeletal imaging of the shoulder were pre-set for remaining parameters.

Captured images and movie sequences were stored with unique identifier labels on an external hard disk. Measurement of captured images and movie sequences was performed in “OsiriX v.5.8.2 32-bit” (Rater A) and RadiAnt DICOM viewer 1.9.16 (32 bit) (Rater B).

1 In the overall agreement and study phase, raters were blinded to each other's results and the
2 participant status (Case/Control), and images and movies were stored for at least 21 days before
3 measurements to secure blinding of rater A.
4
5
6
7
8
9

10 **Training and overall agreement phases**

11 In the training phase, rater A and B practiced the US procedures for capturing, measuring and
12 grading the captured images and movies on 10 participants (cases and controls). Overall agreement
13 phase was performed on 20 participants (10 cases and 10 controls), and the overall agreement of at
14 least 80% on each parameter (Present/Not present for Dichotomised variables, CA, NV, FD; no
15 significant ($p>0.05$) rater difference for continuous variables, TT, CA) was obtained before the
16 actual reliability study.
17
18
19
20
21
22
23
24
25
26
27

28 **Study phase 3 (Actual Reliability study)**

29 *Participants*

30 General inclusion criteria were: 18-65 years old; the ability to understand spoken and written
31 Danish; no prior shoulder surgery/dislocation; no sensory or motor deficits in the neck/arm; no
32 suspected competing diagnoses (rheumatoid arthritis, cancer, neurological disorders, fibromyalgia,
33 psychiatric illness).
34
35
36
37
38
39
40

41 Inclusion criteria for cases were: clinical diagnosis of RC tendinopathy with current shoulder
42 complaints lasting for at least three months prior to inclusion; pain located in the proximal lateral
43 aspect of the upper arm (C5 dermatome) aggravated by shoulder abduction; positive 'Full Can test'
44 and/or 'Jobe's test', and/or pain at 'Resisted External Rotation test'; and positive 'Hawkins-
45 Kennedy test' and/or 'Neer's test'; and US verification of at least one of the following
46 characteristics: FD, NV, CA (the involved side), or side difference (increased/decreased) TT of the
47 supraspinatus tendon.¹⁹
48
49
50
51
52
53
54
55
56

57 Exclusion criteria for cases were pain (during rest) rated above 40 mm (Visual Analogue pain Scale,
58
59
60

1 range: 0 to 100mm); bilateral shoulder pain; less than 90 degrees of active elevation of the arm; full
2 thickness rupture in the supraspinatus tendon (verified by US); calcification above 5 mm in the
3 vertical distance (x-ray); corticosteroid injection within the latest six weeks; humerus fracture (x-
4 ray); diagnoses of glenohumeral osteoarthritis; frozen shoulder; clinically suspected labrum lesion;
5 symptomatic osteoarthritis in the acromioclavicular joint; or symptoms from the cervical spine.¹⁹

6 Inclusion criteria for controls were no shoulder discomfort within the latest 3 months and negative
7 clinical shoulder tests.

8 Cases were consecutively recruited from specialised shoulder units at three hospitals in Denmark.¹⁹

9 Controls were recruited by advertisement among staff from The Department of Sports Science and
10 Clinical Biomechanics, University of Southern Denmark, and the Rehabilitation Department,
11 Lillebaelt hospital - Vejle hospital.

12 Informed consent was obtained from participants before inclusion.

13 STATISTICS

14 Linear weighted Cohen's kappa (LWk) was used to calculate inter-rater reliability with 95%
15 Confidence intervals (CI) for the ordinal variables (FD, number of CA and NV). Firstly, a linear
16 weighing (LWk version 1) was applied, corresponding to the formula: $1-|i-j|/(k-1)$, where i and j are
17 the number of rows and columns, and k is the maximum number of possible ratings.²⁰ Secondly, the
18 same weighing was used (LWk version 2), but with the restriction, that disagreement between grade
19 0 and >0, was weighted as zero, to account for the ability to differentiate between healthy and non-
20 healthy.

21 Kappa was interpreted as ≤ 0.00 =Poor; 0.01-0.20=Slight; 0.21-0.40=Fair; 0.41-0.60=Moderate;
22 0.61-0.80=Substantial and 0.81-1.00=Almost perfect.²¹

23 For the continuous variables (TT, total length of CA), Intra Class Correlation (ICC) (3.1) was
24 calculated as a measure of reliability. ICC was interpreted as < 0.40 =Poor, 0.40-0.75=Fair to Good
25 and > 0.75 =Excellent reliability.²² Bland-Altman plots with 95% Limits of Agreement (LOA) were
26

calculated as a measure of absolute agreement for TT (right and left) and total length of CA, and between-rater difference was tested by a paired t-test. Funnel effects and systematic bias were assessed visually and from Pearson's correlation coefficient, r . Standard Error of Measurement (SEM) was calculated as $SEM_{(Agreement)}^{23}$ to extrapolate results to the general population of potential raters, and Minimal Detectable Change (MDC) was calculated at individual ($MDC_{Individual}$) and group level (MDC_{group}).²⁴ Unpaired t-test was calculated, for defining a potential cut-point of TT between cases and controls.

Data was analysed in Stata/IC 14 (2015, Statacorp, College Station, Texas, USA). P-values <0.05 were considered significant.

RESULTS

There were no differences in demographics between cases and controls, except for pain and discomfort, as expected due to the study design (Table 3).

Table 3: Demographics (Study phase; n=40)

	Cases (n=24)	Controls (n=16)	p-values
Sex (woman/men)	10/14	10/6	0.20
Mean age (years) (SD)	47.0(9.3)	39.8(15.4)	0.13
Height (cm) (SD)	176.2(10.75)	171.9(7.8)	0.18
Weight (kg) (SD)	79.7(18.1)	71.6(19.3)	0.10
BMI	25.4(3.6)	24.1(5.7)	0.12
Dominant arm right	21/24	14/16	0.30
Duration of pain (months)(SD)	24.3(34.9)	0(0)	<0.01
VAS rest (0-100)(SD)	6.5(7.4)	0(0)	<0.01
VAS activity (0-100)(SD)	36.8(16.4)	0(0)	<0.01
VAS Sleep (0-100)(SD)	30.0(23.6)	0(0)	<0.01
VAS Max (0-100)(SD)	70.5(14.1)	0(0)	<0.01
DASH (0-100)(SD)	23.6(11.1)	1.0(2.29)	<0.01

BMI: Body Mass Index; VAS: Visual Analog Scale;
DASH: Disability of Arm, Shoulder and Hand (DASH) questionnaire.

Total agreement ranged from 83-99%, linear weighted kappa (LWk version 1) for FD, NV and CA ranged from 0.60 - 0.96, and kappa with constraints (LWk version 2) varied from 0.51 – 0.98, representing reliability of “Moderate - Almost perfect” (Table 4).

Table 4: Inter-rater reliability of grading presence of fibrillar disruption (FD), neovascularization (NV) and number of calcifications (CA) (Study phase; n=40)

Ordinal scale	Total agreement (LWK version 1)	Linear Weighted K (version 1) (95%CI)	Linear weighted K (version 2) (95% CI)
FD	83.3%	0.60 (0.40;0.79)	0.51 (0.30;0.72)
CA	93.8%	0.72 (0.59;0.85)	0.75 (0.56;0.89)
NV	99.2%	0.96 (0.85;1.0)	0.98 (0.93;1.0)

Linear Weighted K (version 1): No cut-point applied in weights schedule; Linear Weighted K (version 2): Cut-point applied in weights when rater A and B disagrees between grade 0 or >0; FD: Fibrillar disruption; CA: Calcification; NV: Neovascularity.

For total length of CA and TT ICC ranged from 0.85 – 0.90 (Excellent), with $SEM_{(Agreement)}$ ranging from 0.63 – 2.94mm, $MDC_{(group)}$ from 0.28 – 1.29mm, and $MDC_{(individual)}$ from 1.75 – 8.15mm (Table 5).

Table 5: Inter-rater reliability of tendon thickness (TT) and total length of calcification (CA) (Study phase; n=40)

Continuous scale	Rater A (mm (SD))	Rater B (mm (SD))	Diff. (mm (SD))	P	LOA (mm)	SEM (mm)	$MDC_{(G)}$ (mm)	$MDC_{(I)}$ (mm)	ICC (95%CI)
TT Right	7.18 (1.08)	7.29 (1.09)	-0.11 (0.56)	0.22	-1.20 ; 0.98	0.63	0.28 (3.87%)	1.75 (24.2%)	0.87 (0.76;0.93)
TT Left	6.96 (1.26)	7.11 (0.98)	-0.15 (0.49)	0.07	-1.11 ; 0.81	0.74	0.33 (4.69%)	2.05 (29.1%)	0.90 (0.82;0.95)
Total length CA	2.81 (4.95)	2.28 (4.16)	0.53 (2.45)	0.18	-4.27 ; 5.34	2.94	1.29 (72.01%)	8.15 (320.2%)	0.85 (0.74;0.92)

LOA: Limits Of Agreement. SEM: Standard Error of Measurement (Agreement); $MDC_{(G)}$: Minimal Detectable Change (Group level); $MDC_{(I)}$: Minimal Detectable Change (Individual level); TT: Tendon Thickness; CA: Calcification

No systematic rater differences were found in measured TT and total length of CA (Table 5).

Bland-Altman plots showed no funnel effects, but a small interaction between difference and increased mean was found for TT in left shoulder ($r=0.35$, $p=0.03$) (Figure 2). In general, LOA showed larger variation for CA than for TT (Table 5; Figure 2).

No significant difference was found between cases and controls in TT.

DISCUSSION

Inter-rater reliability study, showed moderate to perfect reliability for grading fibrillar disruption, neovascularization and number of calcifications, using standardized procedures. Inter-rater

1 reliability for measuring total length of calcification and tendon thickness was excellent, and MDC
2 indicated small detectable changes for group level, especially in TT.
3
4
5
6
7

8 *Fibrillar disruption (FD) and hypoechoic areas*

9
10 Despite merging hypoechoic areas and FD into one scale, reliability was still only moderate (LWk
11 of 0.60 and 0.51). This was, however, in line with previous studies of tendinopathy, where
12 especially agreement on subtle changes (“Mild abnormality” and “Normal”) was considered
13 difficult, presumably due to difficulties in differing structural changes and anisotropy.^{4 5 7 9 10}
14
15 Grading FD, may be more easily interpreted with in vivo US-examinations, as the examiner is more
16 flexible when evaluating presence of anisotropy.
17
18
19
20
21
22
23
24
25

26 *Neovascularization (NV)*

27
28 The current reliability of NV was almost perfect. The reason for the high reliability in the current
29 study may be the grading of NV in relation to a predetermined Region Of Interest (ROI) (fixed box
30 of 5 x 5mm placed over the area with most NV), as opposed to grading NV relative to the tendon
31 thickness or the tendon in general as previously in tendinopathy of the elbow^{4 15} The current
32 modification was performed to increase standardization, but also to account for between and within
33 variations in tendon thickness, of interest in intervention studies.
34
35
36
37
38
39
40

41 Other studies have found prevalence of NV in 30-65% of symptomatic shoulders with only 25% of
42 asymptomatic shoulders.^{25 26} The current study found prevalence of NV in 38% of the cases and 0%
43 in the control group. This large variation in prevalence across previous studies may be due to
44 different populations, PD settings, measurement methods and the position of the participant arm
45 during US image capturing. The current study placed the hand at the sacrum, to maximally stretch
46 the supraspinatus tendon, which may have increased the risk of overlooking NV due to restricted
47 flow in the neo-vessels. Different study designs across studies, makes it difficult to compare
48 prevalence and establish normative levels for use in clinical practice.
49
50
51
52
53
54
55
56
57
58
59
60

Calcification (CA)

The substantial kappa for detecting the total number of CA is in line with previous studies,^{4,7} but LOA, SEM and MDC showed considerably variation on total length of CA. This variation may be due to US-methodological problems, e.g. that shadows underneath CA may falsely be interpreted as FD and/or normal tendon structure may appear hyperechoic, thus resembling CA, which may result in misclassifications. However, reliability of number of CA was high, indicating that measuring individual lengths of CA and/or few undetected/misclassified CA have influenced agreement of total length of CA. One outlier seen in the Bland-Altman plots (figure 2) indicates, that rater A and B disagreed on at least one larger structural change, which, due to the generally small size and low prevalence of CA, have influenced the variation considerably.

Tendon thickness (TT)

Excellent reliability, and MDC of <0.33mm, indicates that the variable is sensitive for detecting changes, in line with a previous study using the same method for measuring TT.¹⁶ This means that it may be a clinically relevant measurement for assessment of changes in tendon properties, such as increased/decreased oedema. Some studies have found significant differences in TT between symptomatic and non-symptomatic participants,^{27,28} which are in contrast to the current and a recent study.⁽³¹⁾ The reason for the discrepancy across studies may be due to different methods for measuring tendon thickness, small sample sizes, different inclusion criteria, or as in the current study the inclusion of more active controls (recruited among health personnel) with potentially thicker tendons than an average population.

One limitation of the study is the transferability to clinical setting, as the present study used captured images and strictly standardized procedures, which are rarely used in clinical settings. In vivo, raters would be more flexible when evaluating presence of anisotropy in the interpretation of

1 potential FD, and also they would be able to perform repeated image capturing and measurements
2 when CA or NV were suspected to be present. Use of a standardised protocol for reliability
3 studies,¹² may be a weakness, since reliability of the current US method and procedures may have
4 been deceptively high compared with a clinical setting. However, if the standardized method has
5 poor reliability in a standardized setting, reliability is assumed also to be poor and the method less
6 relevant for use in a clinical setting. The raters measured and graded the captured images and
7 movies on different DICOM viewers. If this have influenced the reliability is unknown. However,
8 since reliability is found to be high on most variables it is considered not to be of importance, and to
9 mimic clinical practice.
10
11
12
13
14
15
16
17
18
19
20
21
22
23

24 Strengths of this study is the design, incorporating a stepwise and standardized procedure in order
25 to minimize bias and increase reliability.¹² The present standardization of both US image and movie
26 capturing, measuring and grading structural changes is anticipated to increase reliability and
27 sensitivity of the method. Despite one of the raters having relatively few years of US experience
28 reliability was still at a high and satisfactory level, indicating the protocol can be followed by other
29 than very US-experienced clinicians. By using captured images and movie sequences it was secured
30 that both raters had equal underlying basis for interpreting the material in the reliability study.
31
32 Further, use of both linear weighing and weighing with restrictions, of calculated kappa were
33 considered important in ordinal scales, and due to the importance of being able to differ between
34 cases and controls.
35
36
37
38
39
40
41
42
43
44
45
46
47

48 CONCLUSION

49 Inter-rater reliability was moderate to almost perfect, when a standardized procedure was applied
50 for measuring structural changes on captured US images and movie sequences of relevance for
51 patients with supraspinatus tendinopathy. Future studies should test intra- and inter-rater reliability
52
53
54
55
56
57
58
59
60

1 of the method in vivo for use in clinical practice, in addition to validation against a gold standard,
2 such as MRI.
3
4
5
6
7

8 **Competing interests**

9
10 The authors declare that they have no competing interests.
11
12

13 **Funding**

14
15
16
17 Region of Southern Denmark's Research fund, The Danish Rheumatism Association and the
18
19 Ryholts Foundation funded the trial.
20
21
22

23 **Ethical approval**

24
25
26 All procedures performed in studies involving human participants were in accordance with the
27
28 ethical standards of the institutional and/or national research committee and with the 1964 Helsinki
29
30 declaration and its later amendments or comparable ethical standards. The Regional Scientific
31
32 Ethics Committee of Southern Denmark has approved the trial (project ID: S-20130071).
33
34
35
36

37 **Contributions:**

38
39 KGI, BJK, HE and CML conceived and designed the study protocol. KGI and BJK procured the
40
41 project funding. KGI, BJK, JV and JH developed and standardised the ultrasound procedure and
42
43 defined the grading scale. JV and JH secured access and coordinated screening procedures at the
44
45 shoulder units. KGI was project coordinator, performed the inclusion and ultrasound performance.
46
47 KGI and JH was rater. KGI and BJK planed and coordinated the statistical analysis. KGI performed
48
49 the statistical analysis. KGI drafted the manuscript, and BJK, JH, JV, HE and CML contributed to
50
51 the manuscript. All authors read and approved the final manuscript. KGI is the guarantor.
52
53
54
55
56

57 **Data sharing statement:**

No additional data available.

Figure legends

Figure 1 Flowchart of the Training, Overall agreement and Study phase

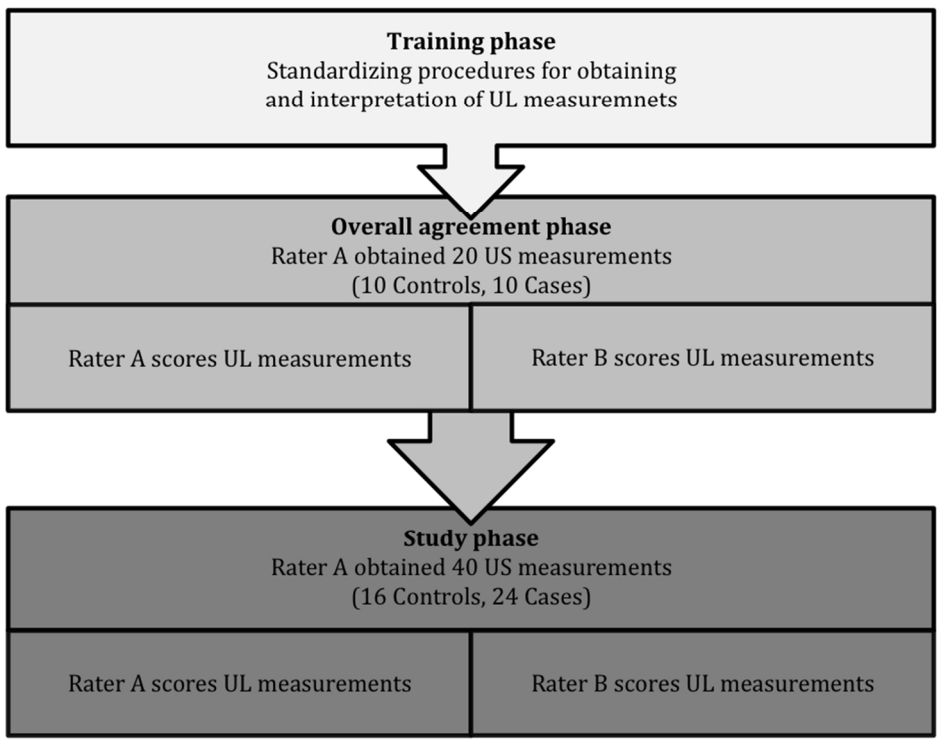
Figure 2: Bland-Altman plots with 95% Limits of Agreement (LOA) for Tendon Thickness (TT) (Right and Left) and total length of Calcifications (CA)

REFERENCE LIST

1. Cook JL, Purdam CR. Is tendon pathology a continuum? A pathology model to explain the clinical presentation of load-induced tendinopathy. *Br J Sports Med* 2009;**43**(6):409-16.
2. Lewis JS. Rotator cuff tendinopathy: a model for the continuum of pathology and related management. *Br J Sports Med* 2010;**44**(13):918-23.
3. Hegedus EJ, Goode A, Campbell S, et al. Physical examination tests of the shoulder: a systematic review with meta-analysis of individual tests. *Br J Sports Med* 2008;**42**(2):80-92; discussion 92.
4. Poltawski L, Ali S, Jayaram V, et al. Reliability of sonographic assessment of tendinopathy in tennis elbow. *Skeletal Radiol* 2012;**41**(1):83-9.
5. 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**(1):14-9.
6. Ottenheijm RP, van't Klooster IG, Starmans LM, et al. Ultrasound-diagnosed disorders in shoulder patients in daily general practice: a retrospective observational study. *BMC Fam Pract* 2014;**15**:115.
7. O'Connor PJ, Rankine J, Gibbon WW, et al. Interobserver variation in sonography of the painful shoulder. *J Clin Ultrasound* 2005;**33**(2):53-6.
8. Thoomes-de Graaf M, Scholten-Peeters GG, Duijn E, et al. Inter-professional agreement of ultrasound-based diagnoses in patients with shoulder pain between physical therapists and radiologists in the Netherlands. *Man Ther* 2014;**19**(5):478-83.
9. O'Connor PJ, Grainger AJ, Morgan SR, et al. Ultrasound assessment of tendons in asymptomatic volunteers: a study of reproducibility. *Eur Radiol* 2004;**14**(11):1968-73.
10. Sunding K, Fahlstrom M, Werner S, et al. Evaluation of Achilles and patellar tendinopathy with greyscale ultrasound and colour Doppler: using a four-grade scale. *Knee Surg Sports Traumatol Arthrosc* 2014.
11. Weinreb JH, Sheth C, Apostolakos J, et al. Tendon structure, disease, and imaging. *Muscles Ligaments Tendons J* 2014;**4**(1):66-73.
12. Patijn J.V, Beek J.V, Blomberg S, et al. Reproducibility and validity studies of diagnostic procedures in manual/musculoskeletal medicine. In: Patijn J, ed.: *International Federation for Manual/Musculoskeletal Medicine*, 2004.
13. Ohberg L, Alfredson H. Ultrasound guided sclerosis of neovessels in painful chronic Achilles tendinosis: pilot study of a new treatment. *Br J Sports Med* 2002;**36**(3).

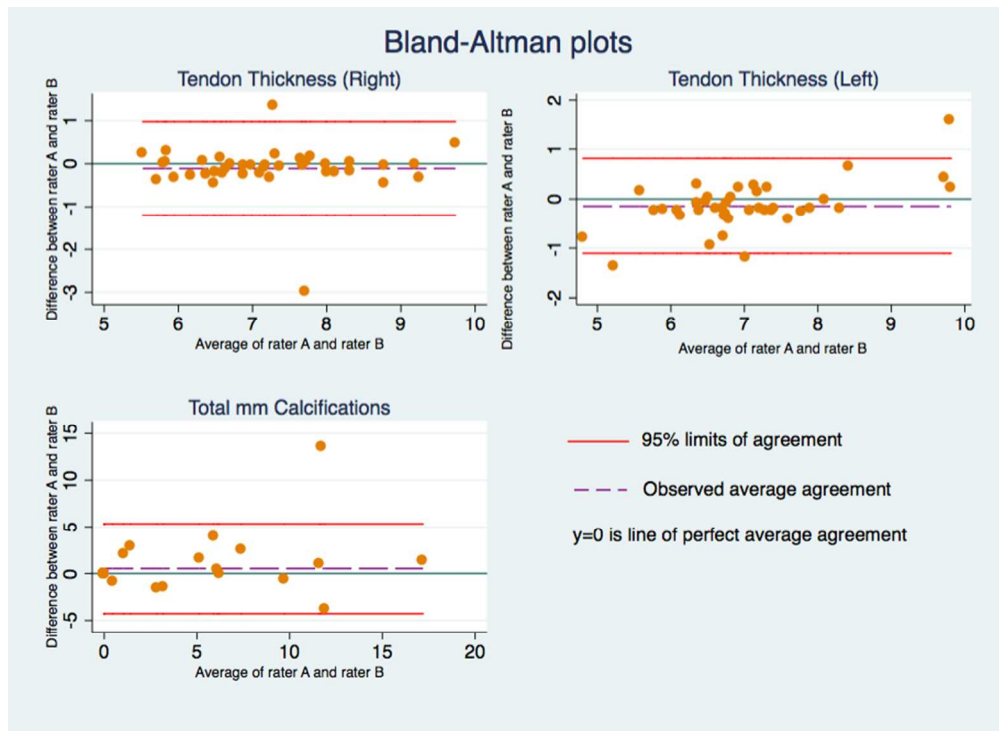
14. Venu KM, Howlett, D.C., Garikipati, R., Anderson, H.J., Bonnici, A.V. Evaluation of the symptomatic supraspinatus tendon - a comparison of ultrasound and arthroscopy. *Radiography* 2002;**8**(4):235-40.
15. Krogh TP, Fredberg U, Stengaard-Pedersen K, et al. Treatment of lateral epicondylitis with platelet-rich plasma, glucocorticoid, or saline: a randomized, double-blind, placebo-controlled trial. *Am J Sports Med* 2013;**41**(3):625-35.
16. Houg Kjaer B. Intra-rater and inter-rater reliability of standardized ultrasound protocol for assessing subacromial structures (Submitted after revision). *Physiotherapy Theory and Practice*, 2015.
17. Ingwersen KG, Hjarbaek J, Eshoej H, et al. Sonographic assessment of supraspinatus tendinopathy as a future diagnostic and effect measure – a reliability study of the assessment methode. XXV Congress of the International Society of Biomechanics. Glasgow, UK: ISB, 2015.
18. Martinoli C. Musculoskeletal ultrasound: technical guidelines. *Insights Imaging* 2010;**1**(3):99-141.
19. Ingwersen KG, Christensen R, Sorensen L, et al. Progressive high-load strength training compared with general low-load exercises in patients with rotator cuff tendinopathy: study protocol for a randomised controlled trial. *Trials* 2015;**16**:27.
20. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 2005;**85**(3):257-68.
21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**(1):159-74.
22. Fleiss JL. Reliability of Measurement. *The Design and Analysis of Clinical Experiments*. Hoboken, NJ, USA.: John Wiley & Sons, Inc., 1999.
23. de Vet HC, Terwee CB, Knol DL, et al. When to use agreement versus reliability measures. *Journal of clinical epidemiology* 2006;**59**(10):1033-9.
24. de Vet HCW TC, Mokkink LB, Knol DL. *Measurement in medicine - A practical guide*. New York: U.S.A: Cambridge University Press, New York, 2011.
25. Kardouni JR, Seitz AL, Walsworth MK, et al. Neovascularization prevalence in the supraspinatus of patients with rotator cuff tendinopathy. *Clin J Sport Med* 2013;**23**(6):444-9.
26. Lewis JS, Raza SA, Pilcher J, et al. The prevalence of neovascularity in patients clinically diagnosed with rotator cuff tendinopathy. *BMC Musculoskelet Disord* 2009;**10**:163.
27. Arend CF, Arend AA, da Silva TR. Diagnostic value of tendon thickness and structure in the sonographic diagnosis of supraspinatus tendinopathy: room for a two-step approach. *Eur J Radiol* 2014;**83**(6):975-9.
28. Michener LA, Subasi Yesilyaprak SS, Seitz AL, et al. Supraspinatus tendon and subacromial space parameters measured on ultrasonographic imaging in subacromial impingement syndrome. *Knee Surg Sports Traumatol Arthrosc* 2015;**23**(2):363-9.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



352x264mm (72 x 72 DPI)

Review only



321x234mm (72 x 72 DPI)

view only

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	1
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	2
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	3
	4	Study objectives and hypotheses	3-4
METHODS			
<i>Study design</i>	5	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	4
<i>Participants</i>	6	Eligibility criteria	7
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	7
	8	Where and when potentially eligible participants were identified (setting, location and dates)	8
	9	Whether participants formed a consecutive, random or convenience series	8
<i>Test methods</i>	10a	Index test, in sufficient detail to allow replication	4-6
	10b	Reference standard, in sufficient detail to allow replication	4-6 (same as index)
	11	Rationale for choosing the reference standard (if alternatives exist)	4
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	6
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	6
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	7
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	7
<i>Analysis</i>	14	Methods for estimating or comparing measures of diagnostic accuracy	8-9
	15	How indeterminate index test or reference standard results were handled	Na.
	16	How missing data on the index test and reference standard were handled	Na.
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	Na.
	18	Intended sample size and how it was determined	4
RESULTS			
<i>Participants</i>	19	Flow of participants, using a diagram	4
	20	Baseline demographic and clinical characteristics of participants	9
	21a	Distribution of severity of disease in those with the target condition	Na.
	21b	Distribution of alternative diagnoses in those without the target condition	Na.
	22	Time interval and any clinical interventions between index test and reference standard	Na.
<i>Test results</i>	23	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	10
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	10
	25	Any adverse events from performing the index test or the reference standard	Na.
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	11-13
	27	Implications for practice, including the intended use and clinical role of the index test	11-13
OTHER INFORMATION			
	28	Registration number and name of registry	14
	29	Where the full study protocol can be accessed	Na.
	30	Sources of funding and other support; role of funders	14

STARD 2015

AIM

STARD stands for “Standards for Reporting Diagnostic accuracy studies”. This list of items was developed to contribute to the completeness and transparency of reporting of diagnostic accuracy studies. Authors can use the list to write informative study reports. Editors and peer-reviewers can use it to evaluate whether the information has been included in manuscripts submitted for publication.

EXPLANATION

A **diagnostic accuracy study** evaluates the ability of one or more medical tests to correctly classify study participants as having a **target condition**. This can be a disease, a disease stage, response or benefit from therapy, or an event or condition in the future. A medical test can be an imaging procedure, a laboratory test, elements from history and physical examination, a combination of these, or any other method for collecting information about the current health status of a patient.

The test whose accuracy is evaluated is called **index test**. A study can evaluate the accuracy of one or more index tests. Evaluating the ability of a medical test to correctly classify patients is typically done by comparing the distribution of the index test results with those of the **reference standard**. The reference standard is the best available method for establishing the presence or absence of the target condition. An accuracy study can rely on one or more reference standards.

If test results are categorized as either positive or negative, the cross tabulation of the index test results against those of the reference standard can be used to estimate the **sensitivity** of the index test (the proportion of participants *with* the target condition who have a positive index test), and its **specificity** (the proportion *without* the target condition who have a negative index test). From this cross tabulation (sometimes referred to as the contingency or “2x2” table), several other accuracy statistics can be estimated, such as the positive and negative **predictive values** of the test. Confidence intervals around estimates of accuracy can then be calculated to quantify the statistical **precision** of the measurements.

If the index test results can take more than two values, categorization of test results as positive or negative requires a **test positivity cut-off**. When multiple such cut-offs can be defined, authors can report a receiver operating characteristic (ROC) curve which graphically represents the combination of sensitivity and specificity for each possible test positivity cut-off. The **area under the ROC curve** informs in a single numerical value about the overall diagnostic accuracy of the index test.

The **intended use** of a medical test can be diagnosis, screening, staging, monitoring, surveillance, prediction or prognosis. The **clinical role** of a test explains its position relative to existing tests in the clinical pathway. A replacement test, for example, replaces an existing test. A triage test is used before an existing test; an add-on test is used after an existing test.

Besides diagnostic accuracy, several other outcomes and statistics may be relevant in the evaluation of medical tests. Medical tests can also be used to classify patients for purposes other than diagnosis, such as staging or prognosis. The STARD list was not explicitly developed for these other outcomes, statistics, and study types, although most STARD items would still apply.

DEVELOPMENT

This STARD list was released in 2015. The 30 items were identified by an international expert group of methodologists, researchers, and editors. The guiding principle in the development of STARD was to select items that, when reported, would help readers to judge the potential for bias in the study, to appraise the applicability of the study findings and the validity of conclusions and recommendations. The list represents an update of the first version, which was published in 2003.

More information can be found on <http://www.equator-network.org/reporting-guidelines/stard>.



BMJ Open

Ultrasound assessment for grading structural tendon changes in supraspinatus tendinopathy - An inter-rater reliability study

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2016-011746.R1
Article Type:	Research
Date Submitted by the Author:	25-Apr-2016
Complete List of Authors:	Ingwersen, Kim; Hospital Lillebalt, Department of Physiotherapy; University of Southern Denmark, Department of Sports Science and Clinical Biomechanics Hjarbaek, John; Odense University Hospital, Department of Radiology, Musculoskeletal section Eshoej, Henrik; University of Southern Denmark, Department of Sports Science and Clinical Biomechanics Larsen, Camilla; University of Southern Denmark, Department of Sports Science and Clinical Biomechanics; University College Lillebaelt - Campus Odense, Health Sciences Research Centre Vobbe, Jette; Hospital Lillebaelt, Orthopedic Department, Shoulder Unit Juil-Kristensen, Birgit; University of Southern Denmark, Department of Sports Science and Clinical Biomechanics; Hogskolen i Bergen, Institute of Occupational Therapy, Physiotherapy and Radiography, Department of Health Sciences
Primary Subject Heading:	Radiology and imaging
Secondary Subject Heading:	Diagnostics, Rehabilitation medicine, Sports and exercise medicine
Keywords:	Reliability, Tendinopathy, Ultrasound < RADIOLOGY & IMAGING, ULTRASONOGRAPHY, Shoulder < ORTHOPAEDIC & TRAUMA SURGERY

SCHOLARONE™
Manuscripts

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Ultrasound assessment for grading structural tendon changes in supraspinatus tendinopathy -
An inter-rater reliability study

Kim Gordon Ingwersen^{1,2}, John Hjarbaek³, Henrik Eshøj¹, Camilla Marie Larsen^{1,4}, Jette Vobbe⁵,
Birgit Juul-Kristensen^{1,6}

¹Department of Sports Science and Clinical Biomechanics, University of Southern Denmark,
Odense, Denmark.

²Department of Rehabilitation, Hospital Lillebaelt - Vejle Hospital, Vejle, Denmark

³Department of Radiology, Musculoskeletal section, Odense University Hospital, Odense, Denmark

⁴Health Sciences Research Centre, University College Lillebaelt, Denmark

⁵Shoulder Unit, Orthopaedic Department, Hospital Lillebaelt, Vejle Hospital, Vejle, Denmark

⁶Institute of Occupational Therapy, Physiotherapy and Radiography, Department of Health
Sciences, Bergen University College, Bergen, Norway

Corresponding author: Kim Gordon Ingwersen, Department of Rehabilitation - Hospital Lillebaelt,
Kabeltoft 25, DK-7100, Vejle. Tlf.nr: +45 79 40 61 75, E-mail: kim.riis@rsyd.dk.

Keywords: Reliability; Tendinopathy; Ultrasound; Sonography; Shoulder.

Word count: 2826

Inter-rater reliability of ultrasound assessment for grading structural tendon changes in supraspinatus tendinopathy

ABSTRACT

Aim. To evaluate the inter-rater reliability of measuring structural changes in the tendon of patients, clinically diagnosed with supraspinatus tendinopathy (Cases) and healthy participants (Controls), on ultrasound (US) images captured by standardized procedures

Methods. A total of 40 participants (24 patients) were included for assessing inter-rater reliability of measurements of fibrillar disruption, neovascularity, number and total length of calcifications and tendon thickness. Linear weighted kappa, Intra Class Correlation (ICC), Standard Error of Measurement (SEM), Limits Of Agreement (LOA) and Minimal Detectable Change (MDC) were used to evaluate reliability.

Results. “Moderate - Almost perfect” kappa was found for grading fibrillar disruption, neovascularity and number of calcifications (k : 0.60 - 0.96). For total length of calcifications and tendon thickness ICC was “Excellent” (0.85 – 0.90), with $SEM_{(Agreement)}$ ranging from 0.63 – 2.94mm and $MDC_{(group)}$ ranging from 0.28 – 1.29mm. In general, SEM, LOA and MDC showed larger variation for calcifications than for tendon thickness.

Conclusion. Inter-rater reliability was moderate to almost perfect, when a standardized procedure was applied for measuring structural changes on captured US images and movie sequences of relevance for patients with supraspinatus tendinopathy. Future studies should test intra- and inter-rater reliability of the method in vivo for use in clinical practice, in addition to validation against a gold standard, such as MRI.

STRENGTHS AND LIMITATIONS OF THIS STUDY:

- A standardized procedure for US capturing and measuring structural changes of the supraspinatus tendon is presented
- A specific procedure for grading and interpreting tendinopathy related changes is presented
- Grading and measurement can be performed reliably

- 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
- Performance of the method in vivo are warranted to validate the method in clinical practice

INTRODUCTION

Rotator Cuff (RC) tendinopathy can be considered a continuum of pathology, and tailored rehabilitation according to the stage in this continuum is recommended.^{1 2} Anamnesis and special orthopaedic tests are often used when diagnosing RC tendinopathy, but these tests often lack high specificity and sensitivity, making diagnosis uncertain,³ thus challenging precise and targeted treatment.

Grey-Scale (GS) ultrasound (US) and Power Doppler (PD) visualization of RC tendons may be helpful to detect signs of tendinopathy, such as hypoechoic areas, fibrillar disruption, neovascularisation, calcifications embedded in the tendon or edema and confirm the “a priori” hypothesis of RC tendinopathy, provided satisfactory clinimetric properties of the US method.^{4 5} However, US is an operator dependent technique and requires thorough training and experience in performance and assessment before precise diagnoses can be made, especially in relation to more subtle changes as often seen within tendinopathy.⁶ Poor to fair reliability has previously been found when comparing diagnoses made by US novel and experienced clinicians.⁷⁻⁹ Further, when grading subtle structural tendon changes, especially hypoechoic areas, only fair, and thus unsatisfactory reliability has been found, even among experienced clinicians.^{6 8 10-12}

Standardised procedures for capturing and assessing US is known to increase reliability of US based diagnoses.⁶ Previously, assessment of tendinopathy were found reliable, in patients with tendinopathy in the elbow, ankle or knee, when using standardised procedures for measuring GS and PD.^{4 11}

For the shoulder, however, there is lack of clinically relevant, standardized and reliable methods for assessing tendinopathy. Since US is highly influenced by clinician experience and technique, both standardized US procedures for image and movie capturing, and standardized procedures for assessment of structural changes in relation to tendinopathy need to be defined.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Therefore, the aim of this study was from standardized US procedures for image and movie capturing, to evaluate the inter-rater reliability of measuring and grading structural changes in the tendon of patients clinically diagnosed with supraspinatus tendinopathy (Cases) and healthy participants (Controls).

MATERIAL AND METHODS

Study design

The study followed the protocol for diagnostic procedures in reproducibility studies.¹³ This protocol includes a three-phase study design consisting of a 1) training; 2) an overall agreement and 3) a study phase (the actual reliability study) (Figure 1).

The phases constitute a methodological model for optimizing procedures, and aim at eliminating clinician subjectivity as much as possible. The aim of the training phase is to secure that raters have sufficient competence and experience in performing the procedures. The overall agreement phase is an extended training phase and secures that gross systematic bias between raters are minimized, and requires at least 80% agreement between raters, before proceeding to phase 3. The study phase, is the final evaluation of reliability of the developed procedures.¹³

Inter-rater reliability (phase 3) between two raters (rater A and B) was tested on measuring and grading structural changes relevant to tendinopathy upon US captured images and movies. Rater A (KI; Physiotherapist) had one year of clinical musculoskeletal US experience, and rater B (JH; Radiologist) had more than fifteen years of clinical musculoskeletal US experience.

US image capturing and measurement

Based upon the literature,^{4 10 11 14-17} consensus was made upon definitions of relevant potential pathological structural changes related to tendinopathy, including 1) *fibrillar disruption (FD)*, 2) *neovascularisation (NV)*, 3) *calcification (CA)*, and 4) *tendon thickness (TT)*. Hereafter, a standardized protocol for US capturing was developed, consisting of three static images (grey-scale),

three dynamic movie sequences (grey-scale), and one Doppler movie sequence (*Table 1*).

Table 1: Description of US procedures for capturing image and movie sequences of Fibrillar disruption (FD), Neovascularity (NV), Calcifications (CA) and Tendon Thickness (TT)

1) Fibrillar disruption (FD):

FD was defined as a clear collagen fascicle discontinuity or irregularity of fibrils in an otherwise regular parallel structuring of fibres in the tendon.

A GS picture in the longitudinal axis of the supraspinatus was taken at the sight where FD was most apparent (FD picture). The FD static image was used for classifying presence of FD. A GS posterior-anterior dynamic movie sequence (PA movie) in the longitudinal plane of the supraspinatus tendon was captured, by moving the transducer slowly in the posterior-anterior direction. Further, a caudal-cranial (CC) transversal GS dynamic movie sequence (CC movie) of the supraspinatus tendon was recorded by moving the transducer slowly in the CC direction. The static image and the movie sequence recordings were used as confirmation and assistance in assessing the grade of structural changes, and to secure identification of potential ambiguous GS features, such as anisotropy (erroneous signal caused when the transducer is angled obliquely to the tendon).

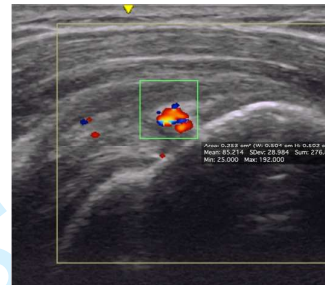
FD was classified in relation to tendon thickness as: 0=normal, 1=mild, 2=moderate, 3=severe, 4=partial rupture (table 2).



Grade 2 FD

2) Neovascularity (NV):

NV was defined as a visualized Power Doppler (PD) signal with minimal artifactual noise. The supraspinatus tendon was evaluated for presence of NV by moving the transducer slowly in the posterior-anterior direction, with the PD feature activated. In case NV was present, a 10 sec dynamic movie sequence was recorded at the point with most NV signal (PD movie). When grading NV from the PD movie sequence, a static image of the location with the most visible NV was captured from the PD movie. A Region of Interest (ROI) (5x5 millimetre (mm)) was placed around the NV and used for grading NV. In participants with no NV a movie sequence was recorded at a random location in the tendon to verify absence of NV. NV was classified in relation to ROI as 0=normal, 1=mild, 2=moderate, 3=severe, 4=Extreme (Table 2).



Grade 2 NV

3) Calcification (CA):

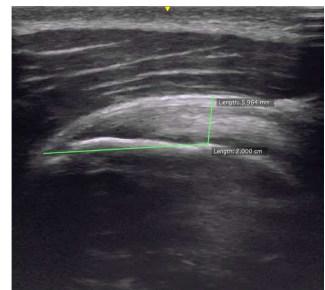
CA was defined as distinct white borders, imbedded in the length of the tendon, often with “shadows” underneath. The PA movie was used to identify the number of CA in the tendon and to measure the length of each CA in the longitudinal axis of the supraspinatus. The length was measured between the most medial and lateral aspect of the distinct white boarder (in mm). The CC movie was used as confirmation and assistance in identifying CA. CA was counted and measured (mm). To obtain the total length of CA, the individual CA lengths were added up to one total per participant.



CA length measure

4) Tendon thickness (TT):

TT was defined as the height, from the humeral head, at a point 20 mm from the supraspinatus tendon-snip (tendon insertion) in the longitudinal axis of the tendon to the most superficial part of the tendon. In practice, an image was captured at a fixed point just laterally from the anterior-lateral corner of the acromion in the longitudinal plane of the supraspinatus. When measuring TT, a mark was placed 20 mm cranial from the supraspinatus tendon-snip (tendon insertion), on the edge of the cartilage of the humeral head. From that mark, the perpendicular thickness of tendon was recorded.¹⁸ The TT picture was recorded bilaterally. TT was measured in mm.



TT measure

Secondly, based upon previous scales used to measure structural changes in tendinopathy at the elbow,^{4 16} two ordinal grading scales for FD and NV were adjusted for use in the shoulder.¹⁹ The scales ranged from 0-4 (FD: 0=Normal tendon; 4=Partial rupture, corresponding to disruption of the fibers in the full thickness of the tendon; NV (0=Normal, including no signal; 4= Extreme, including Doppler activity in more than 50% of the region of interest, ROI) (Table 2) (Appendix: Illustration of grading levels for fibrillar disruption and neovascularity).

Table 2. Grading scales with definitions for fibrillar disruption (FD) and neovascularity (NV)

Grade	Fibrillar disruption	Neovascularity
0	Normal	Normal (No signal)
1	Mild (Involving under 25% of the height of the tendon)	Mild (Single small signal in the Region Of Interest, ROI)
2	Moderate (Involving 25-50% of the height of the tendon)	Moderate (Doppler activity in less than 25% of the ROI)
3	Severe (Involving more than 50% of the height of the tendon)	Severe (Doppler activity in 25-50% of the ROI)
4	Partial rupture (Disruption of the fibers in the full thickness of the tendon)	Extreme (Doppler activity in more than 50% of the ROI)

Abbreviations: ROI: Region of interest

CA was analysed as number of calcifications and total length (in mm), while TT was measured in mm.¹⁸

Rater A performed capturing of all US images and movie sequences with the participant seated, the shoulder internally rotated with the dorsal side of the hand placed on the sacrum, and the elbow flexed and directed laterally, to optimize visualisation of the supraspinatus tendon.²⁰

A GE LOGIQ e B12 (GE Healthcare, Wisconsin, USA) with a 5.0 – 13.0 MHz linear transducer was used for image capturing. All US scanings were standardized and performed for GS imaging at 13.0 MHz and 56% gain, while PD scanning was performed with a pulse repetition frequency of 0.41 kHz and gain at 56%. Manufacturer recommendations for musculoskeletal imaging of the shoulder were pre-set for remaining parameters.

Captured images and movie sequences were stored with unique identifier labels on an external hard

1 disk. Measurement of captured images and movie sequences was performed in “OsiriX v.5.8.2 32-
2 bit” (Rater A) and RadiAnt DICOM viewer 1.9.16 (32 bit) (Rater B).
3

4
5
6 In the overall agreement and study phase, raters were blinded to each other’s results and the
7
8 participant status (Case/Control), and images and movies were stored for at least 21 days before
9
10 measurements to secure blinding of rater A.
11

12 13 14 **Training and overall agreement phases**

15
16
17 In the training phase, rater A and B practiced the US procedures for capturing, measuring and
18
19 grading the captured images and movies on 10 participants (cases and controls). Overall agreement
20
21 phase was performed on 20 participants (10 cases and 10 controls), and the overall agreement of at
22
23 least 80% on each parameter (Present/Not present for Dichotomised variables, CA, NV, FD; no
24
25 significant ($p>0.05$) rater difference for continuous variables, TT, CA) was obtained before the
26
27 actual reliability study.
28
29

30 31 32 **Study phase 3 (Actual Reliability study)**

33 34 *Participants*

35
36
37 General inclusion criteria were: 18-65 years old; the ability to understand spoken and written
38
39 Danish; no prior shoulder surgery/dislocation; no sensory or motor deficits in the neck/arm; no
40
41 suspected competing diagnoses (rheumatoid arthritis, cancer, neurological disorders, fibromyalgia,
42
43 psychiatric illness).
44

45
46 Inclusion criteria for cases were: clinical diagnosis of RC tendinopathy with current shoulder
47
48 complaints lasting for at least three months prior to inclusion; pain located in the proximal lateral
49
50 aspect of the upper arm (C5 dermatome) aggravated by shoulder abduction; positive ‘Full Can test’
51
52 and/or ‘Jobe’s test’, and/or pain at ‘Resisted External Rotation test’; and positive ‘Hawkins-
53
54 Kennedy test’ and/or ‘Neer’s test’; and US verification of at least one of the following
55
56 characteristics: FD, NV, CA (the involved side), or side difference (increased/decreased) TT of the
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

supraspinatus tendon.²¹

Exclusion criteria for cases were pain (during rest) rated above 40 mm (Visual Analogue pain Scale, range: 0 to 100mm); bilateral shoulder pain; less than 90 degrees of active elevation of the arm; full thickness rupture in the supraspinatus tendon (verified by US); calcification above 5 mm in the vertical distance (x-ray); corticosteroid injection within the latest six weeks; humerus fracture (x-ray); diagnoses of glenohumeral osteoarthritis; frozen shoulder; clinically suspected labrum lesion; symptomatic osteoarthritis in the acromioclavicular joint; or symptoms from the cervical spine.²¹

Inclusion criteria for controls were no shoulder discomfort within the latest 3 months and negative clinical shoulder tests.

Cases were consecutively recruited from specialised shoulder units at three hospitals in Denmark as part of a Randomised Controlled Trial (RCT).²¹ Controls were recruited by advertisement among staff from The Department of Sports Science and Clinical Biomechanics, University of Southern Denmark, and the Rehabilitation Department, Lillebaelt hospital - Vejle hospital.

Informed consent was obtained from participants before inclusion.

STATISTICS

Linear weighted Cohen's kappa (LWk) was used to calculate inter-rater reliability with 95% Confidence intervals (CI) for the ordinal variables (FD, number of CA and NV). Firstly, a linear weighing (LWk version 1) was applied, corresponding to the formula: $1 - |i - j| / (k - 1)$, where i and j are the number of rows and columns, and k is the maximum number of possible ratings.²² Secondly, the same weighing was used (LWk version 2), but with the restriction, that disagreement between grade 0 and >0, was weighted as zero, to account for the ability to differentiate between healthy and non-healthy.

Kappa was interpreted as ≤ 0.00 =Poor; 0.01-0.20=Slight; 0.21-0.40=Fair; 0.41-0.60=Moderate; 0.61-0.80=Substantial and 0.81-1.00=Almost perfect.²³

For the continuous variables (TT, total length of CA), Intra Class Correlation (ICC) (3.1) was calculated as a measure of reliability. ICC was interpreted as <0.40 =Poor, $0.40-0.75$ =Fair to Good and >0.75 =Excellent reliability.²⁴ Bland-Altman plots with 95% Limits of Agreement (LOA) were calculated as a measure of absolute agreement for TT (right and left) and total length of CA, and between-rater difference was tested by a paired t-test. Funnel effects and systematic bias were assessed visually and from Pearsons correlation coefficient, r . Standard Error of Measurement (SEM) was calculated as $SEM_{(Agreement)}$ ²⁵ to extrapolate results to the general population of potential raters, and Minimal Detectable Change (MDC) was calculated at individual ($MDC_{Individual}$) and group level (MDC_{group}).²⁶ Unpaired t-test was calculated, for defining a potential cut-point of TT between cases and controls.

For the study phase a sample size of 40 participants was applied, as previously recommended for reliability studies.¹³

Data was analysed in Stata/IC 14 (2015, Statacorp, College Station, Texas, USA). P-values <0.05 were considered significant.

RESULTS

There were no differences in demographics between cases and controls, except for pain and discomfort, as expected due to the study design (Table 3).

Table 3: Demographics (Study phase; n=40)

	Cases (n=24)	Controls (n=16)	p-values
Sex (woman/men)	10/14	10/6	0.20
Mean age (years) (SD)	47.0(9.3)	39.8(15.4)	0.13
Height (cm) (SD)	176.2(10.75)	171.9(7.8)	0.18
Weight (kg) (SD)	79.7(18.1)	71.6(19.3)	0.10
BMI	25.4(3.6)	24.1(5.7)	0.12
Dominant arm right	21/24	14/16	0.30
Duration of pain (months)(SD)	24.3(34.9)	0(0)	<0.01
VAS rest (0-100)(SD)	6.5(7.4)	0(0)	<0.01
VAS activity (0-100)(SD)	36.8(16.4)	0(0)	<0.01
VAS Sleep (0-100)(SD)	30.0(23.6)	0(0)	<0.01

VAS Max (0-100)(SD)	70.5(14.1)	0(0)	<0.01
DASH (0-100)(SD)	23.6(11.1)	1.0(2.29)	<0.01

BMI: Body Mass Index; VAS: Visual Analog Scale;
DASH: Disability of Arm, Shoulder and Hand (DASH) questionnaire.

Total agreement ranged from 83-99%, linear weighted kappa (LWk version 1) for FD, NV and CA ranged from 0.60 - 0.96, and kappa with constraints (LWk version 2) varied from 0.51 – 0.98, representing reliability of “Moderate - Almost perfect” (Table 4).

Table 4: Inter-rater reliability of grading presence of fibrillar disruption (FD), neovascularization (NV) and number of calcifications (CA) (Study phase; n=40)

Ordinal scale	Total agreement (LWK version 1)	Linear Weighted K (version 1) (95%CI)	Linear weighted K (version 2) (95% CI)
FD	83.3%	0.60 (0.40;0.79)	0.51 (0.30;0.72)
CA	93.8%	0.72 (0.59;0.85)	0.75 (0.56;0.89)
NV	99.2%	0.96 (0.85;1.0)	0.98 (0.93;1.0)

Linear Weighted K (version 1): No cut-point applied in weights schedule;
Linear Weighted K (version 2): Cut-point applied in weights when rater A and B disagrees between grade 0 or >0; FD: Fibrillar disruption; CA: Calcification; NV: Neovascularity.

For total length of CA and TT ICC ranged from 0.85 – 0.90 (Excellent), with $SEM_{(Agreement)}$ ranging from 0.63 – 2.94mm, $MDC_{(group)}$ from 0.28 – 1.29mm, and $MDC_{(individual)}$ from 1.75 – 8.15mm (Table 5).

Table 5: Inter-rater reliability of tendon thickness (TT) and total length of calcification (CA) (Study phase; n=40)

Continuous scale	Rater A (mm (SD))	Rater B (mm (SD))	Diff. (mm (SD))	P	LOA (mm)	SEM (mm)	$MDC_{(G)}$ (mm)	$MDC_{(I)}$ (mm)	ICC (95%CI)
TT Right	7.18 (1.08)	7.29 (1.09)	-0.11 (0.56)	0.22	-1.20 ; 0.98	0.63	0.28 (3.87%)	1.75 (24.2%)	0.87 (0.76;0.93)
TT Left	6.96 (1.26)	7.11 (0.98)	-0.15 (0.49)	0.07	-1.11 ; 0.81	0.74	0.33 (4.69%)	2.05 (29.1%)	0.90 (0.82;0.95)
Total length CA	2.81 (4.95)	2.28 (4.16)	0.53 (2.45)	0.18	-4.27 ; 5.34	2.94	1.29 (72.01%)	8.15 (320.2%)	0.85 (0.74;0.92)

LOA: Limits Of Agreement. SEM: Standard Error of Measurement (Agreement); $MDC_{(G)}$: Minimal Detectable Change (Group level); $MDC_{(I)}$: Minimal Detectable Change (Individual level); TT: Tendon Thickness; CA: Calcification

No systematic rater differences were found in measured TT and total length of CA (Table 5).

Bland-Altman plots showed no funnel effects, but a small interaction between difference and increased mean was found for TT in left shoulder ($r=0.35$, $p=0.03$) (Figure 2). In general, LOA showed larger variation for CA than for TT (Table 5; Figure 2).

No significant difference was found between cases and controls in TT.

DISCUSSION

Inter-rater reliability study, showed moderate to perfect reliability for grading fibrillar disruption, neovascularization and number of calcifications, using standardized procedures. Inter-rater reliability for measuring total length of calcification and tendon thickness was excellent, and MDC indicated small detectable changes for group level, especially in TT.

Fibrillar disruption (FD) and hypoechoic areas

Despite merging hypoechoic areas and FD into one scale, reliability was still only moderate (LWk of 0.60 and 0.51). This was, however, in line with previous studies of tendinopathy, where especially agreement on subtle changes (“Mild abnormality” and “Normal”) was considered difficult, presumably due to difficulties in differing structural changes and anisotropy.^{4 6 8 10 11}

Grading FD, may be more easily interpreted with in vivo US-examinations, as the examiner is more flexible when evaluating presence of anisotropy.

Neovascularization (NV)

The current reliability of NV was almost perfect. The reason for the high reliability in the current study may be the grading of NV in relation to a predetermined Region Of Interest (ROI) (fixed box of 5 x 5mm placed over the area with most NV), as opposed to grading NV relative to the tendon thickness or the tendon in general as previously in tendinopathy of the elbow^{4 16} The current modification was performed to increase standardization, but also to account for between and within variations in tendon thickness, of interest in intervention studies.

Other studies have found prevalence of NV in 30-65% of symptomatic shoulders with only 25% of asymptomatic shoulders.^{27 28} The current study found prevalence of NV in 38% of the cases and 0% in the control group. This large variation in prevalence across previous studies may be due to different populations, PD settings, measurement methods and the position of the participant arm during US image capturing. The current study placed the hand at the sacrum, to maximally stretch

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

the supraspinatus tendon, which may have increased the risk of overlooking NV due to restricted flow in the neo-vessels. Different study designs across studies, makes it difficult to compare prevalence and establish normative levels for use in clinical practice.

Calcification (CA)

The substantial kappa for detecting the total number of CA is in line with previous studies,⁴⁸ but LOA, SEM and MDC showed considerably variation on total length of CA. This variation may be due to US-methodological problems, e.g. that shadows underneath CA may falsely be interpreted as FD and/or normal tendon structure may appear hyperechoic, thus resembling CA, which may result in misclassifications. However, reliability of number of CA was high, indicating that measuring individual lengths of CA and/or few undetected/misclassified CA have influenced agreement of total length of CA. One outlier seen in the Bland-Altman plots (figure 2) indicates, that rater A and B disagreed on at least one larger structural change, which, due to the generally small size and low prevalence of CA, have influenced the variation considerably.

Tendon thickness (TT)

Excellent reliability, and MDC of ≤ 0.33 mm, indicates that the variable is sensitive for detecting changes, in line with a previous study using the same method for measuring TT.¹⁸ This means that it may be a clinically relevant measurement for assessment of changes in tendon properties, such as increased/decreased oedema. Some studies have found significant differences in TT between symptomatic and non-symptomatic participants,^{29,30} which are in contrast with the current and a recent study.¹⁸ The reason for the discrepancy across studies may be due to different methods for measuring tendon thickness, small sample sizes, different inclusion criteria, or as in the current study the inclusion of more active controls (recruited among health personnel) with potentially thicker tendons than an average population.

1 One limitation of the study is the transferability to clinical setting, as the present study used
2 captured images and strictly standardized procedures, which are rarely used in clinical settings. In
3 vivo, raters would be more flexible when evaluating presence of anisotropy in the interpretation of
4 potential FD, and also they would be able to perform repeated image capturing and measurements
5 when CA or NV were suspected to be present. Use of a standardised protocol for reliability
6 studies,¹³ may be a weakness, since reliability of the current US method and procedures may have
7 been deceptively high compared with a clinical setting. However, if the standardized method has
8 poor reliability in a standardized setting, reliability is assumed also to be poor and the method less
9 relevant for use in a clinical setting. The raters measured and graded the captured images and
10 movies on different DICOM viewers. Whether this has influenced the reliability is unknown.
11 However, since reliability is found to be high on most variables it is considered not to be of
12 importance, and to mimic clinical practice.
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29

30 Strengths of this study are the design, incorporating a stepwise and standardized procedure in order
31 to minimize bias and increase reliability.¹³ The present standardization of both US image and movie
32 capturing, measuring and grading structural changes are anticipated to increase reliability and
33 sensitivity of the method. Despite one of the raters having relatively few years of US experience
34 reliability was still high and satisfactory, indicating the protocol can be followed by other than very
35 US-experienced clinicians. By using captured images and movie sequences it was secured that both
36 raters had equal underlying bases for interpretation of the reliability study.
37
38
39
40
41
42
43
44
45

46 Further, use of both linear weighing and weighing with restrictions, calculated kappa was
47 considered important in ordinal scales, and due to the importance of being able to differ between
48 cases and controls.
49
50
51
52
53
54

55 CONCLUSION

56
57
58
59
60

1 Inter-rater reliability was moderate to almost perfect, when a standardized procedure was applied
2 for measuring structural changes on captured US images and movie sequences of relevance for
3 patients with supraspinatus tendinopathy. Future studies should test intra- and inter-rater reliability
4 of the method in vivo for use in clinical practice, in addition to validation against a gold standard,
5 such as MRI.
6
7
8
9
10
11
12
13

14 **Competing interests**

15 The authors declare that they have no competing interests.
16
17
18
19
20

21 **Funding**

22 Region of Southern Denmark's Research fund, The Danish Rheumatism Association and the
23 Ryholts Foundation funded the trial.
24
25
26
27
28
29

30 **Ethical approval**

31 All procedures performed in studies involving human participants were in accordance with the
32 ethical standards of the institutional and/or national research committee and with the 1964 Helsinki
33 declaration and its later amendments or comparable ethical standards. The Regional Scientific
34 Ethics Committee of Southern Denmark has approved the trial (project ID: S-20130071).
35
36
37
38
39
40
41
42
43

44 **Contributions:**

45 KGI, BJK, HE and CML conceived and designed the study protocol. KGI and BJK procured the
46 project funding. KGI, BJK, JV and JH developed and standardised the ultrasound procedure and
47 defined the grading scale. JV and JH secured access and coordinated screening procedures at the
48 shoulder units. KGI was project coordinator, performed the inclusion and US image capturing. KGI
49 and JH were raters. KGI and BJK planned and coordinated the statistical analyses. KGI performed
50 the statistical analyses. KGI drafted the manuscript, and BJK, JH, JV, HE and CML contributed to
51
52
53
54
55
56
57
58
59
60

1 the manuscript. All authors read and approved the final manuscript. KGI is the guarantor.
2
3
4
5

6 **Data sharing statement:**
7

8 No additional data available.
9
10

11
12
13 **Figure legends**
14

15 Figure 1 Flowchart of the Training, Overall agreement and Study phase
16

17 Figure 2: Bland-Altman plots with 95% Limits of Agreement (LOA) for Tendon Thickness (TT)
18 (Right and Left) and total length of Calcifications (CA)
19
20
21
22
23
24
25
26
27

28 **REFERENCE LIST**
29

- 30 1. Cook JL, Purdam CR. Is tendon pathology a continuum? A pathology model to explain the
31 clinical presentation of load-induced tendinopathy. *Br J Sports Med* 2009;**43**(6):409-16.
- 32 2. Lewis JS. Rotator cuff tendinopathy: a model for the continuum of pathology and related
33 management. *Br J Sports Med* 2010;**44**(13):918-23.
- 34 3. Hegedus EJ, Goode A, Campbell S, et al. Physical examination tests of the shoulder: a systematic
35 review with meta-analysis of individual tests. *Br J Sports Med* 2008;**42**(2):80-92; discussion
36 92.
- 37 4. Poltawski L, Ali S, Jayaram V, et al. Reliability of sonographic assessment of tendinopathy in
38 tennis elbow. *Skeletal Radiol* 2012;**41**(1):83-9.
- 39 5. Ottenheijm RP, Jansen MJ, Staal JB, et al. Accuracy of diagnostic ultrasound in patients with
40 suspected subacromial disorders: a systematic review and meta-analysis. *Arch Phys Med
41 Rehabil* 2010;**91**(10):1616-25.
- 42 6. Naredo E, Moller I, Moragues C, et al. Interobserver reliability in musculoskeletal
43 ultrasonography: results from a "Teach the Teachers" rheumatologist course. *Ann Rheum
44 Dis* 2006;**65**(1):14-9.
- 45 7. Ottenheijm RP, van't Klooster IG, Starmans LM, et al. Ultrasound-diagnosed disorders in
46 shoulder patients in daily general practice: a retrospective observational study. *BMC Fam
47 Pract* 2014;**15**:115.
- 48 8. O'Connor PJ, Rankine J, Gibbon WW, et al. Interobserver variation in sonography of the painful
49 shoulder. *J Clin Ultrasound* 2005;**33**(2):53-6.
- 50 9. Thoomes-de Graaf M, Scholten-Peeters GG, Duijn E, et al. Inter-professional agreement of
51 ultrasound-based diagnoses in patients with shoulder pain between physical therapists and
52 radiologists in the Netherlands. *Man Ther* 2014;**19**(5):478-83.
- 53 10. O'Connor PJ, Grainger AJ, Morgan SR, et al. Ultrasound assessment of tendons in
54 asymptomatic volunteers: a study of reproducibility. *Eur Radiol* 2004;**14**(11):1968-73.
55
56
57
58
59
60

11. Sunding K, Fahlstrom M, Werner S, et al. Evaluation of Achilles and patellar tendinopathy with greyscale ultrasound and colour Doppler: using a four-grade scale. *Knee Surg Sports Traumatol Arthrosc* 2014.
12. Weinreb JH, Sheth C, Apostolakos J, et al. Tendon structure, disease, and imaging. *Muscles Ligaments Tendons J* 2014;**4**(1):66-73.
13. Patijn J.V, Beek J.V, Blomberg S, et al. Reproducibility and validity studies of diagnostic procedures in manual/musculoskeletal medicine. In: Patijn J, ed.: *International Federation for Manual/Musculoskeletal Medicine*, 2004.
14. Ohberg L, Alfredson H. Ultrasound guided sclerosis of neovessels in painful chronic Achilles tendinosis: pilot study of a new treatment. *Br J Sports Med* 2002;**36**(3).
15. Venu KM, Howlett, D.C., Garikipati, R., Anderson, H.J., Bonnici, A.V. Evaluation of the symptomatic supraspinatus tendon - a comparison of ultrasound and arthroscopy. *Radiography* 2002;**8**(4):235-40.
16. Krogh TP, Fredberg U, Stengaard-Pedersen K, et al. Treatment of lateral epicondylitis with platelet-rich plasma, glucocorticoid, or saline: a randomized, double-blind, placebo-controlled trial. *Am J Sports Med* 2013;**41**(3):625-35.
17. Ottenheijm RP, Joore MA, Walenkamp GH, et al. The Maastricht Ultrasound Shoulder pain trial (MUST): ultrasound imaging as a diagnostic triage tool to improve management of patients with non-chronic shoulder pain in primary care. *BMC Musculoskelet Disord* 2011;**12**:154.
18. Hougs Kjaer B. Intra-rater and inter-rater reliability of standardized ultrasound protocol for assessing subacromial structures (Submitted after revision). *Physiotherapy Theory and Practice*, 2015.
19. Ingwersen KG, Hjarbaek J, Eshoej H, et al. Sonographic assessment of supraspinatus tendinopathy as a future diagnostic and effect measure – a reliability study of the assessment methode. XXV Congress of the International Society of Biomechanics. Glasgow, UK: ISB, 2015.
20. Martinoli C. Musculoskeletal ultrasound: technical guidelines. *Insights Imaging* 2010;**1**(3):99-141.
21. Ingwersen KG, Christensen R, Sorensen L, et al. Progressive high-load strength training compared with general low-load exercises in patients with rotator cuff tendinopathy: study protocol for a randomised controlled trial. *Trials* 2015;**16**:27.
22. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 2005;**85**(3):257-68.
23. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**(1):159-74.
24. Fleiss JL. *Reliability of Measurement. The Design and Analysis of Clinical Experiments*. Hoboken, NJ, USA.: John Wiley & Sons, Inc., 1999.
25. de Vet HC, Terwee CB, Knol DL, et al. When to use agreement versus reliability measures. *Journal of clinical epidemiology* 2006;**59**(10):1033-9.
26. de Vet HCW TC, Mokkink LB, Knol DL. *Measurement in medicine - A practical guide*. New York: U.S.A: Cambridge University Press, New York, 2011.
27. Kardouni JR, Seitz AL, Walsworth MK, et al. Neovascularization prevalence in the supraspinatus of patients with rotator cuff tendinopathy. *Clin J Sport Med* 2013;**23**(6):444-9.
28. Lewis JS, Raza SA, Pilcher J, et al. The prevalence of neovascularity in patients clinically diagnosed with rotator cuff tendinopathy. *BMC Musculoskelet Disord* 2009;**10**:163.
29. Arend CF, Arend AA, da Silva TR. Diagnostic value of tendon thickness and structure in the sonographic diagnosis of supraspinatus tendinopathy: room for a two-step approach. *Eur J Radiol* 2014;**83**(6):975-9.
30. Michener LA, Subasi Yesilyaprak SS, Seitz AL, et al. Supraspinatus tendon and subacromial space parameters measured on ultrasonographic imaging in subacromial impingement syndrome. *Knee Surg Sports Traumatol Arthrosc* 2015;**23**(2):363-9.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For peer review only

BMJ Open: first published as 10.1136/bmjopen-2016-011746 on 24 May 2016. Downloaded from <http://bmjopen.bmj.com/> on April 18, 2024 by guest. Protected by copyright.

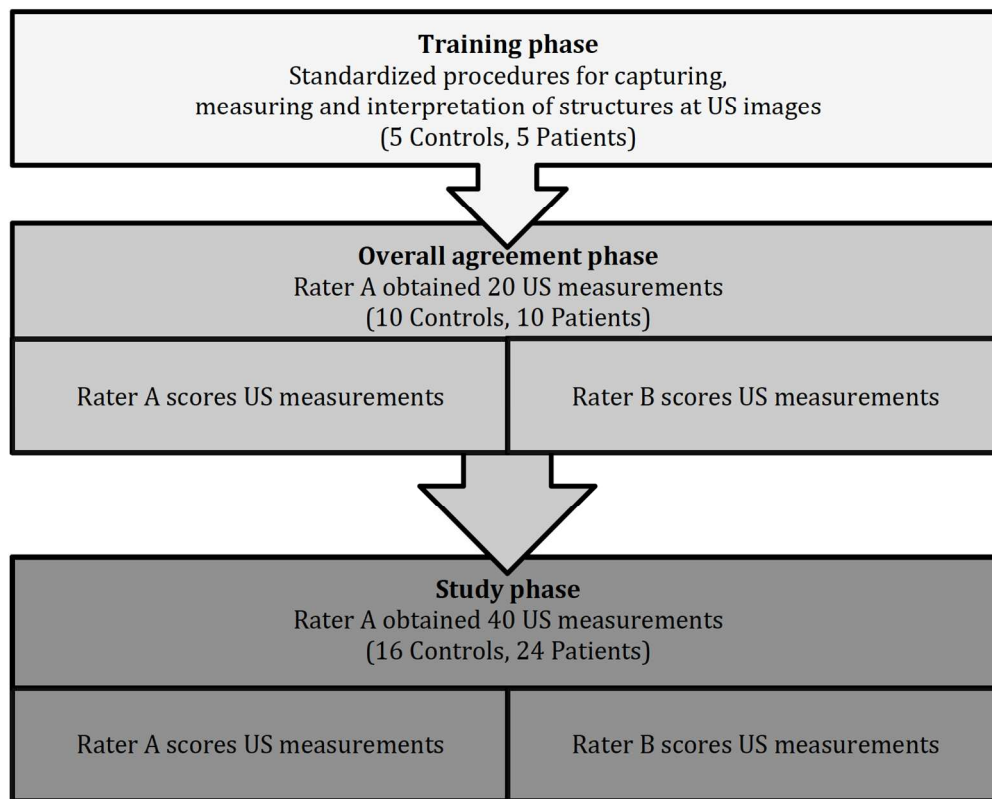


Figure 1 Flowchart of the Training, Overall agreement and Study phase
588x470mm (72 x 72 DPI)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

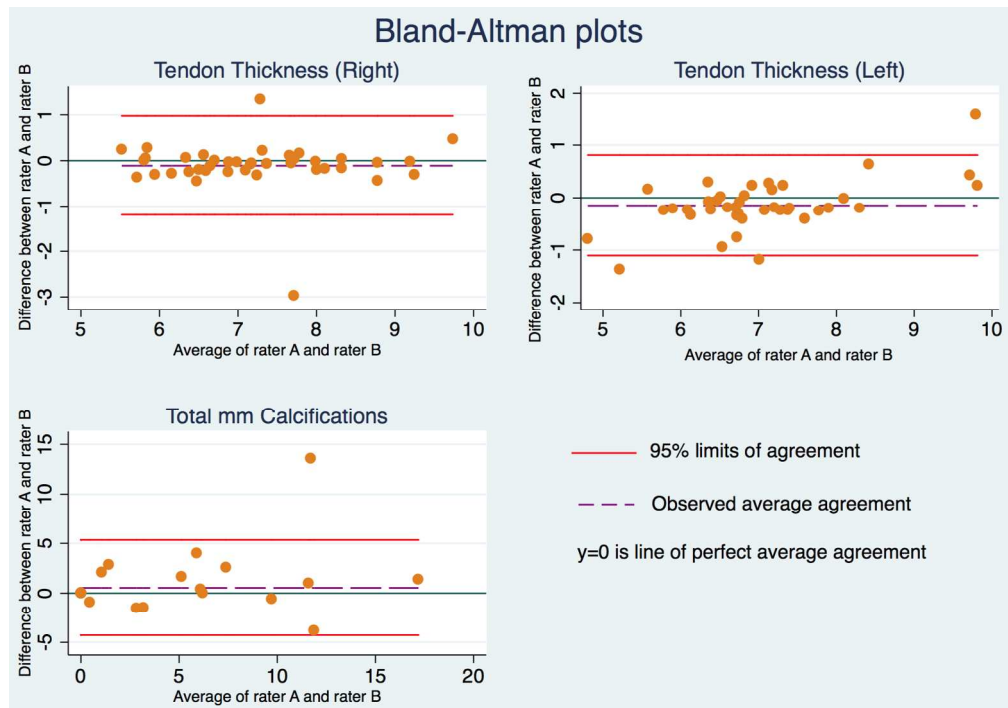
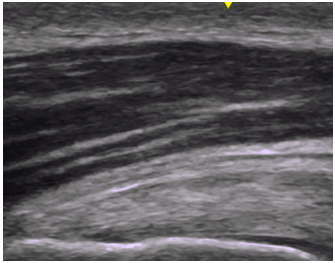
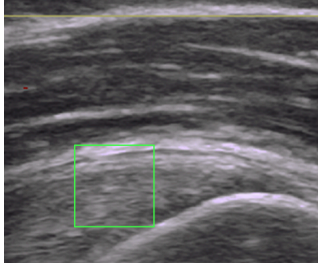
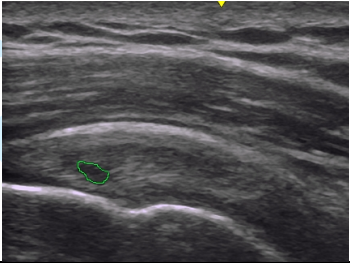
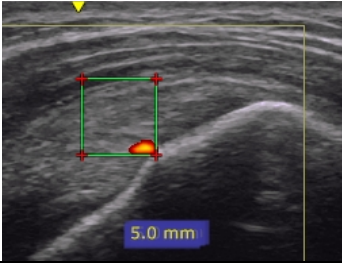
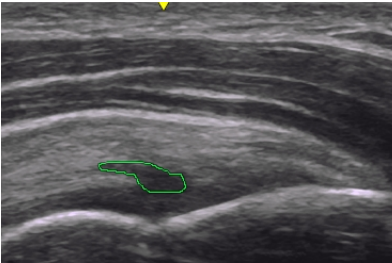
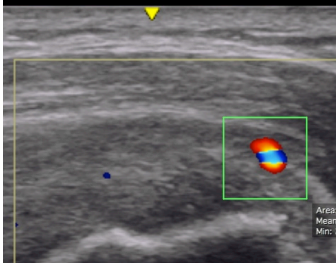
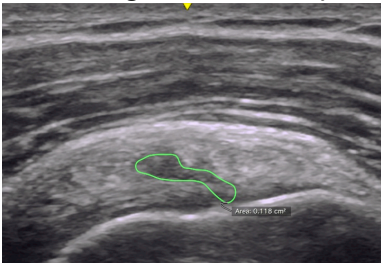
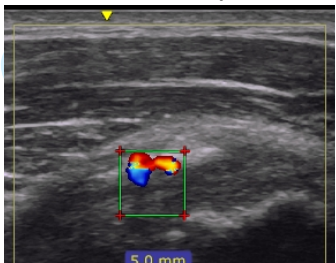
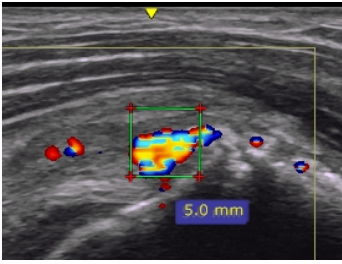


Figure 2: Bland-Altman plots with 95% Limits of Agreement (LOA) for Tendon Thickness (TT) (Right and Left) and total length of Calcifications (CA)
650x459mm (72 x 72 DPI)

Review only

Appendix: Illustration of grading levels for fibrillar disruption and neovascularity

	Fibrillar disruption	Neovascularity
Grade 0	<p>Normal</p> 	<p>Normal (No signal)</p> 
Grade 1	<p>Mild (Involving under 25% of the height of the tendon)</p> 	<p>Mild (Single small signal in the Region Of Interest, ROI)</p> 
Grade 2	<p>Moderate (Involving 25-50% of the height of the tendon)</p> 	<p>Moderate (Doppler activity in less than 25% of the ROI)</p> 
Grade 3	<p>Severe (Involving more than 50% of the height of the tendon)</p> 	<p>Severe (Doppler activity in 25-50% of the ROI)</p> 
Grade 4	<p>Partial rupture (Disruption of the fibers in the full thickness of the tendon)</p> <p>No picture available</p>	<p>Extreme (Doppler activity in more than 50% of the ROI)</p> 

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	1
ABSTRACT			
	2	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	2
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	3
	4	Study objectives and hypotheses	3-4
METHODS			
<i>Study design</i>	5	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	4
<i>Participants</i>	6	Eligibility criteria	7
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	7
	8	Where and when potentially eligible participants were identified (setting, location and dates)	8
	9	Whether participants formed a consecutive, random or convenience series	8
<i>Test methods</i>	10a	Index test, in sufficient detail to allow replication	4-6
	10b	Reference standard, in sufficient detail to allow replication	4-6 (same as index)
	11	Rationale for choosing the reference standard (if alternatives exist)	4
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	6
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	6
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	7
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	7
<i>Analysis</i>	14	Methods for estimating or comparing measures of diagnostic accuracy	8-9
	15	How indeterminate index test or reference standard results were handled	Na.
	16	How missing data on the index test and reference standard were handled	Na.
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	Na.
	18	Intended sample size and how it was determined	4
RESULTS			
<i>Participants</i>	19	Flow of participants, using a diagram	4
	20	Baseline demographic and clinical characteristics of participants	9
	21a	Distribution of severity of disease in those with the target condition	Na.
	21b	Distribution of alternative diagnoses in those without the target condition	Na.
	22	Time interval and any clinical interventions between index test and reference standard	Na.
<i>Test results</i>	23	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	10
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	10
	25	Any adverse events from performing the index test or the reference standard	Na.
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	11-13
	27	Implications for practice, including the intended use and clinical role of the index test	11-13
OTHER INFORMATION			
	28	Registration number and name of registry	14
	29	Where the full study protocol can be accessed	Na.
	30	Sources of funding and other support; role of funders	14

STARD 2015

AIM

STARD stands for “Standards for Reporting Diagnostic accuracy studies”. This list of items was developed to contribute to the completeness and transparency of reporting of diagnostic accuracy studies. Authors can use the list to write informative study reports. Editors and peer-reviewers can use it to evaluate whether the information has been included in manuscripts submitted for publication.

EXPLANATION

A **diagnostic accuracy study** evaluates the ability of one or more medical tests to correctly classify study participants as having a **target condition**. This can be a disease, a disease stage, response or benefit from therapy, or an event or condition in the future. A medical test can be an imaging procedure, a laboratory test, elements from history and physical examination, a combination of these, or any other method for collecting information about the current health status of a patient.

The test whose accuracy is evaluated is called **index test**. A study can evaluate the accuracy of one or more index tests. Evaluating the ability of a medical test to correctly classify patients is typically done by comparing the distribution of the index test results with those of the **reference standard**. The reference standard is the best available method for establishing the presence or absence of the target condition. An accuracy study can rely on one or more reference standards.

If test results are categorized as either positive or negative, the cross tabulation of the index test results against those of the reference standard can be used to estimate the **sensitivity** of the index test (the proportion of participants *with* the target condition who have a positive index test), and its **specificity** (the proportion *without* the target condition who have a negative index test). From this cross tabulation (sometimes referred to as the contingency or “2x2” table), several other accuracy statistics can be estimated, such as the positive and negative **predictive values** of the test. Confidence intervals around estimates of accuracy can then be calculated to quantify the statistical **precision** of the measurements.

If the index test results can take more than two values, categorization of test results as positive or negative requires a **test positivity cut-off**. When multiple such cut-offs can be defined, authors can report a receiver operating characteristic (ROC) curve which graphically represents the combination of sensitivity and specificity for each possible test positivity cut-off. The **area under the ROC curve** informs in a single numerical value about the overall diagnostic accuracy of the index test.

The **intended use** of a medical test can be diagnosis, screening, staging, monitoring, surveillance, prediction or prognosis. The **clinical role** of a test explains its position relative to existing tests in the clinical pathway. A replacement test, for example, replaces an existing test. A triage test is used before an existing test; an add-on test is used after an existing test.

Besides diagnostic accuracy, several other outcomes and statistics may be relevant in the evaluation of medical tests. Medical tests can also be used to classify patients for purposes other than diagnosis, such as staging or prognosis. The STARD list was not explicitly developed for these other outcomes, statistics, and study types, although most STARD items would still apply.

DEVELOPMENT

This STARD list was released in 2015. The 30 items were identified by an international expert group of methodologists, researchers, and editors. The guiding principle in the development of STARD was to select items that, when reported, would help readers to judge the potential for bias in the study, to appraise the applicability of the study findings and the validity of conclusions and recommendations. The list represents an update of the first version, which was published in 2003.

More information can be found on <http://www.equator-network.org/reporting-guidelines/stard>.

