The ability of clinical tests to diagnose stress fractures: a systematic review and meta-analysis
Schneiders AG, Sullivan SJ, Hendrick PA, Hones BD, McMaster AR, Sugden BA, Tomlinson C

CRD summary
This review concluded that the results did not support the specific use of ultrasound or tuning forks as stand-alone diagnostic tests for lower-limb stress fractures. This was a well-conducted review and the authors' conclusions are likely to be reliable.

Authors' objectives
To evaluate the diagnostic accuracy of clinical tests to identify stress fractures in the lower limb.

Searching
PubMed, EMBASE, AMED, CINAHL, PEDro, Scopus and SPORTDiscus were searched from inception to May/June 2011 with no language restrictions; search terms were reported. Reference lists of included studies were screened for additional relevant studies.

Study selection
Diagnostic accuracy studies of one or more clinical tests, compared against a radiological reference test, for the examination of suspected stress fractures were eligible for inclusion. Studies had to report or allow computation of sensitivity, specificity and likelihood ratios. Studies had to include only lower-limb stress fractures, and not impose an age restriction for participants. Studies that specifically assessed pathological stress fractures were excluded. Studies had to be published as full reports before June 2011 to be eligible for inclusion.

Most studies included both men and women with average age ranging from 19 to 31 years, where reported. Participants included soldiers or military basic trainees, athletes, and runners with symptoms of stress fracture of the tibia or fibula or knee pain. Index tests assessed were ultrasound (most used a 3cm head and intensity up to 2.0W/cm², 30 second application, where reported) or tuning fork (128Hz to 512Hz). Assessors included sonography technicians, physiotherapists, a nuclear medicine physician, a radiologist and other clinicians not otherwise specified. The reference standard was scintigraphy, roentgenogram, MRI or a combination of these tests. Precise definitions of positive index tests varied between studies (full details given in review).

Two reviewers independently assessed studies for inclusion, with disagreements resolved by consultation with a third reviewer.

Assessment of study quality
The methodological quality of the included studies was assessed independently by two reviewers using the QUADAS tool; disagreements were resolved by a third reviewer. Items were scored between 1 and 4 points for 'yes' responses, with a maximum possible score of 26. The reviewers were blinded to author, date of publication and journal of publication.

Data extraction
One reviewer produced two by two tables of true positive, false positive, true negative and false negative results, which were used to calculate sensitivity, specificity, and positive and negative likelihood ratios for each included test. If raw data were not reported, sensitivity and specificity values reported by the study authors were used. Data were independently checked by a second reviewer.

Sensitivity and specificity were defined as low if they were 50% or less, low to moderate if from 51% to 64%, moderate if from 65% to 74%, moderate to high if from 75% to 84%, and high if they were 85% or higher.

Methods of synthesis
When appropriate, data were pooled using MetaAnalyst software. When values of zero were present in two by two
tables, a value of 0.5 was added to each cell. Forest plots were generated, when appropriate. Heterogeneity was not formally assessed.

**Results of the review**

Nine diagnostic accuracy studies were included in the review including 420 participants. Seven studies investigated therapeutic ultrasound; two studies investigated tuning fork tests to diagnose stress fractures of the lower limb.

Many of the studies did not clearly specify selection criteria for participants. The reference standard varied between studies. Two studies had a significantly long time between the index and reference tests (therefore the pathology may have changed between tests), the index test was not clearly described in all studies, and it was unclear in most studies whether the index test and reference test were undertaken without the examiner’s knowledge of the other test results. However, in all studies, the reference standard was independent of the index test and described in sufficient detail to permit its replication; most studies clearly reported uninterpretable and/or intermediate test results.

The sensitivity and specificity of therapeutic ultrasound to diagnose stress fractures of the lower limb were classed as low to moderate (seven studies; 333 participants); pooled sensitivity was 64% (95% CI 55 to 73), pooled specificity was 63% (95% CI 54 to 71). The pooled positive likelihood ratio was 2.09 (1.10 to 3.52) and the pooled negative likelihood ratio was 0.35 (95% CI 0.08 to 0.86) in seven studies with 333 participants. The pooled diagnostic odds ratio was 6.20 (95% CI 0.70 to 22.75; six studies).

As only two studies assessed the tuning fork test, data were not pooled using meta-analysis. One study assessed the tuning fork test at a frequency of 128Hz, with a sensitivity of 75% (moderate to high) and specificity of 67% (moderate). The other study assessed three different frequencies; 128Hz, 256Hz and 512Hz. Sensitivity was the highest (92.3%) and specificity the lowest (19.3%) at a frequency of 256Hz. Full results were reported in the paper.

**Authors’ conclusions**

The results did not support the specific use of ultrasound or tuning forks as stand-alone diagnostic tests for lower-limb stress fractures; radiological imaging should continue to be used for the confirmation and diagnosis of stress fractures of the lower limb.

**CRD commentary**

The review addressed a well-defined question and had clear inclusion criteria. The search for published studies was thorough, although only studies published in full were eligible for inclusion, which increased the potential for publication bias. Appropriate methods to reduce reviewer error and bias were undertaken throughout the review process.

An appropriate quality assessment tool was used and full results were reported. Adequate details of the included studies were presented. The included studies examined athletes and/or military personnel, which the authors acknowledged as a limitation of the review, as results may not be generalisable to other populations with a high incidence of stress fractures, including those with underlying pathologies such as osteoporosis, rheumatoid arthritis and cancer. Pooling of ultrasound studies appeared to be appropriate, while a narrative synthesis was appropriate for studies assessing tuning forks. Heterogeneity was not formally assessed, which the authors acknowledged as a limitation of the review; there was considerable heterogeneity between study results.

This was a well conducted systematic review and the authors’ conclusions are likely to be reliable.

**Implications of the review for practice and research**

**Practice:** The authors stated that the results of the review did not support the specific use of ultrasound as a stand-alone diagnostic test for lower-limb stress fractures, and that since the overall diagnostic accuracy of the tests investigated was not strong, radiological imaging should continue to be used for the confirmation and diagnosis of stress fractures of the lower limb.

**Research:** The authors stated that the use of single-photon emission computed tomography should be investigated to determine its ability to be used as a gold standard measure when diagnosing stress fractures. The authors also suggested that future studies use the Standards for the Reporting of Diagnostic Accuracy checklist to improve the reporting of diagnostic accuracy studies, that standardised test parameters should be used when applying ultrasound and tuning forks.
(to allow test reproducibility and conformity), larger studies should be undertaken with well-defined methods, the
timeframe between index and reference testing should be appropriate, and that a common pain or sensation rating scale
to monitor patient responses to tests should be used.

**Funding**
Not stated.

**Bibliographic details**
Schneiders AG, Sullivan SJ, Hendrick PA, Hones BD, McMaster AR, Sugden BA, Tomlinson C. The ability of clinical
tests to diagnose stress fractures: a systematic review and meta-analysis. Journal of Orthopaedic and Sports Physical
Therapy 2012; 42(9): 760-771

**PubMedID**
22813530

**DOI**
10.2519/jospt.2012.4000

**Original Paper URL**
http://www.jospt.org/doi/abs/10.2519/jospt.2012.4000#.U6LjFZ1wa70

**Indexing Status**
Subject indexing assigned by NLM

**MeSH**
Adolescent; Adult; Diagnostic Imaging /methods; Diagnostic Tests, Routine; Female; Fractures, Stress /diagnosis;
Humans; Male; Young Adult

**AccessionNumber**
12012047766

**Date bibliographic record published**
18/12/2012

**Date abstract record published**
08/05/2014

**Record Status**
This is a critical abstract of a systematic review that meets the criteria for inclusion on DARE. Each critical abstract
contains a brief summary of the review methods, results and conclusions followed by a detailed critical assessment on
the reliability of the review and the conclusions drawn.