Research Paper Review

The ability of clinical tests to diagnose stress fractures: A systematic review and meta-analysis

Schneiders AG, Sullivan SJ, Hendrick PA et al.

ABSTRACT

Study Design
Systematic literature review and meta-analysis.

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

Background
Stress fractures are a bone-related overuse injury primarily occurring in the lower limb and commonly affecting running athletes and military personnel. Physical examination procedures and clinical tests are suggested for diagnosing stress fractures; however, data on the diagnostic accuracy of these tests have not been investigated through a systematic review of the literature.

Methods
A systematic review was conducted in 8 electronic databases to identify diagnostic accuracy studies, published between January 1950 and June 2011, that evaluated clinical tests against a radiological diagnosis of lower-limb stress fracture. Retrieved articles were evaluated using the Quality Assessment of Diagnostic Accuracy Studies tool, and a meta-analysis was performed where appropriate.

Results
Nine articles investigating 2 clinical procedures, therapeutic ultrasound (n = 7) and tuning fork testing (n = 2), met the study inclusion criteria. Meta-analysis was used to statistically analyze the data extracted from the ultrasound articles and demonstrated a pooled sensitivity of 64% (95% confidence interval [CI]: 55%, 73%), specificity of 63% (95% CI: 54%, 71%), positive likelihood ratio of 2.1 (95% CI: 1.1, 3.5), and negative likelihood ratio of 0.3 (95% CI: 0.1, 0.9). Tuning fork test data could not be pooled; however,

sensitivity, specificity, positive likelihood ratio, and negative likelihood ratio ranged from 35% to 92%, 19% to 83%, 0.6 to 3.0, and 0.4 to 1.6, respectively.

Conclusion
The results of this systematic review do not support the specific use of ultrasound or tuning forks as standalone diagnostic tests for lower-limb stress fractures. As the overall diagnostic accuracy of the tests investigated is not strong, based on the calculated likelihood ratios, it is recommended that radiological imaging should continue to be used for the confirmation and diagnosis of stress fractures of the lower limb.

Level Of Evidence
Diagnosis, level 1a-.J Orthop Sports Phys Ther 2012;42(9):760-771, Epub 19 July 2012. doi:10.2519/jospt.2012.4000.

ANALYSIS

Reviewed by Dr. Jeff Muir DC (Research Review Service)

Author's Affiliations

School of Physiotherapy, University of Otago, Dunedin, New Zealand; Division of Physiotherapy Education, University of Nottingham, Nottingham, UK

Background Information

Stress fractures are a bone-related overuse injury commonly seen in athletes and military personnel. Stress fractures have been suggested to account for approximately 10% of all athletic injuries and are most commonly seen in runners, with an incidence of up to 20% in this population. Between 80% and 95% of stress fractures occur in the lower limbs, with the tibia being the most commonly injured bone, accounting for approximately 50% of all cases. Clinically, lower-limb stress fractures can be difficult to diagnose, due to a wide range of potential differential diagnoses that include compartment syndrome, soft tissue injuries, infection, and other overuse conditions such as medial tibial stress syndrome and periostitis.

Radiological imaging for stress fractures has traditionally included plain radiographs, bone scan, magnetic resonance imaging (MRI), or computed tomography. The gold standard for stress fracture diagnosis is either triple-phase technetium-99m bone scan (scintigraphy) or MRI . Other common clinical tests have been suggested to show diagnostic potential, including therapeutic ultrasound and tuning forks. Other methods of detection including biochemical markers of bone turnover have proven to be ineffective at diagnosing stress fractures (1).

Systematic reviews of diagnostic accuracy studies have focused on providing validity measures of a test and, where possible, pooling data through a meta-analysis to offer clinicians a summary of evidence of the test's diagnostic accuracy.

Performance metrics used to measure diagnostic tests include: false positive and false negatives, combined with sensitivity (referring to a test's ability to obtain a positive result when condition is present) and specificity (referring to the ability of a test to obtain a negative result when condition is absent), positive (+LR) and negative (-LR) likelihood ratios and diagnostic odds ratio (DOR).

The current study attempted to systematically review the literature and apply meta-analytic procedures to diagnostic studies, where appropriate, to establish which clinical tests have the best accuracy in musculoskeletal and orthopaedic clinical practice to diagnose stress fractures.

SUMMARY

Search Results
The initial electronic database search yielded a total of 9321 articles. After full-text examination, 9 articles met the inclusion criteria and were assessed using the QUADAS tool (the QUADAS tool identifies criteria considered important to the methodological quality of the retrieved studies). The 9 articles retained after the systematic search investigated either therapeutic ultrasound (n = 7) or tuning fork tests (n = 2) to diagnose stress fractures of the lower limb.

Diagnostic Testing Evaluation

Diagnostic Ultrasound:

- Analysis of the diagnostic ability of ultrasound revealed low to moderate pooled sensitivity (64%; 95% CI: 55-73%) and specificity (63%; 95% CI: 54-71%). Although this suggests low to moderate ability of a positive ultrasound test to identify a stress fracture and a negative test to rule out a stress fracture, it is important to note that the +LR (2.09) and –LR (0.35), which are considered more clinically relevant measures of a test's diagnostic ability, were small;

- The 95% CI (0.7-22.75) associated with the pooled DOR (6.2) was very wide and included values considered to be clinically useless (1.0), as well as those with substantial clinical utility (greater than 20). This indicates that these results are too imprecise to draw any conclusion about the usefulness of ultrasound to accurately diagnose a stress fracture of the lower limb in clinical practice

Tuning Fork Testing:

- Data pertaining to the tuning fork test were not pooled using meta-analysis, due to an insufficient number of studies meeting the inclusion criteria for this review (n = 2). Lesho (2) reported moderate to high sensitivity (75%), moderate specificity (67%), and a +LR of 2.3 for a 128-Hz tuning fork, demonstrating a small but sometimes important ability to identify a stress fracture. The high score on the QUADAS tool indicates that these results are less likely to be subject to bias.
- Wilder et al. (3) scored poorly on the QUADAS tool (12/26), mainly due to unclear reporting of many results, suggesting a high chance of bias. They tested 3 different tuning forks, and found the 256-Hz version to have the highest reported sensitivity (92.3%, 90.0%, and 77.7%) when compared to radiography, MRI, and bone scintigraphy, respectively. However, the specificity values were very low for the 256-Hz tuning fork (19.3%, 20% and 25%, respectively), suggesting overall difficulty diagnosing a stress fracture.

CLINICAL APPLICATION & CONCLUSIONS

The results of this study do not support the specific use of ultrasound as a standalone diagnostic test for lower-limb stress fractures. Additionally, the literature supporting the use of tuning forks needs to be interpreted with caution, considering the limited number of studies investigating this modality, the differing results between different tuning fork frequencies, and the reference standard used. As the overall diagnostic accuracy of the tests investigated is not strong, based on the calculated LRs, it is recommended that radiological imaging should continue to be used for the confirmation and diagnosis of stress fractures of the lower limb. Quantification of the degree of heterogeneity of the studies included in the meta-analysis could not be done; therefore, the pooled results need to be interpreted with caution. More high-quality studies, especially to determine the diagnostic accuracy of tuning forks, are required.

Continued use of imaging remains the most appropriate method to ultimately diagnose stress fractures. While not suitable for exclusive use diagnostically, tuning fork testing remains a simple, non-invasive preliminary test useful in clinical settings.

STUDY METHODS

Literature Search:
The authors searched AMED, CINAHL, Embase, MEDLINE, PEDro, PubMed Scopus, and SPORTDiscus for relevant papers published between 1950 and 2011. In order to be included, studies had to fulfill the following criteria:

1. Investigate a clinical test versus 1 or more index tests for the ability to diagnose a lower limb stress fracture
2. Utilize at least 1 radiological reference test
3. Report or allow computation of diagnostic values (sensitivity, specificity, +LR, and –LR)
4. Not impose an age restriction for participants

Articles specifically investigating pathological stress fractures or studies not conducted on human subjects were excluded.

The methodological quality of the included articles was assessed independently by 2 reviewers using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool developed by Whiting et al. (4), which is frequently used to evaluate the quality of diagnostic studies. The QUADAS tool is composed of 14 items individually scored as either "yes," "no," or "unclear." Nine items relate to bias, 3 items to the quality of the reporting, and 2 items to variability.

STUDY STRENGTHS / WEAKNESSES

The authors acknowledge several limitations with their study:

• The obvious limited number of studies using therapeutic US for diagnostic purposes immediately limits the confidence with which the results of that portion of the study can be applied to practice situations.
• All studies reviewed examined athletes and/or military personnel, as a result, the results cannot be generalized to other populations that also have a high incidence of stress fractures.

- The insertion of dummy values into unpopulated cells to allow computation of data in the meta-analysis might also be considered a limitation of this review.
- Although the use of diminutive values in contingency tables is common practice in logistic regression models, meta-analysis of diagnostic studies is rare, and this approach has not yet been fully validated, despite the meta-analysis computer software that does this by default.
- Two types of heterogeneity also contribute to the limitations of this review: clinical heterogeneity may exist due to the mixed nature of the populations in the contributing studies, whereas statistical heterogeneity may exist due to the inconsistent results reported across studies, which was clearly the case in this review.

Additional References

1. Yanovich R, Evans RK, Friedman E, Moran DS. Bone Turnover Markers Do Not Predict Stress Fracture in Elite Combat Recruits. Clin Orthop Relat Res 2012 Dec 13. [Epub ahead of print]
2. Lesho EP. Can tuning forks replace bone scans for identification of tibial stress fractures? Mil Med 1997; 162: 802-803.
3. Wilder RP, Vincent HK, Stewart J et al. Clinical use of tuning forks to identify running-related stress fractures: a pilot study. Athl Train Sports Health Care 2009; 1:12-18.
4. Whiting P, Rutjes AW, Reitsma JB et al. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003; 3:25.

This review is published with the permission of Research Review Service (www.researchreviewservice.com)