PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Reproducibility of tender point examination in chronic low back pain patients as measured by intra- and interrater reliability and agreement: a validation study
AUTHORS	Jensen, Ole; Callesen, Jacob; Nielsen, Merete; Ellingsen, Torkell

VERSION 1 - REVIEW

REVIEWER	Robert Gerwin, MD, FAAN Associate Professor of Neurology Johns Hopkins Unviersity School of Medicine I have no competeting interests
REVIEW RETURNED	22-Jan-2013

GENERAL COMMENTS	This is a paper that evaluates the reliability and agreement of digital tender point examination in chronic low back pain patients. The authors show acceptable reliability on repeat examination over one week for the same examiner (intrarater reliability) and between two examiners (interrater reliability). Recent studies have shown that the majority of tender points in fibromyalgia are myofascial trigger points that cause local and referred pain. It would have been more useful if the authors could have addressed that issue, rather than looking only at tenderness. The authors looked at the total tender point count. It would be more valuable if the authors could have determined that the same sites that were tender on the first examination were tender on the repeat examination, both over time for each examiner, and between examiners, rather than just looking at the total tender point count.

REVIEWER	Theosore Pincus, Clinical Professor of medicine, NYU School of medicine - no competing interests
REVIEW RETURNED	27-Jan-2013

THE STUDY	Without more than 2 observers, the data collected, while carefully incorrectly analyzed, does not seem to add important information to the literature. A larger number of observers, perhaps more than 2 observations, and/or longitudinal observations could add to the literature, but I am afraid that there is almost nothing new in these findings.
GENERAL COMMENTS	The authors have conducted a careful study of the reproducibility of tender point examination in 39 patients, with careful statistical analyses. The data indicate more that 70% agreement within 3 tender points. The reviewer would raise the following considerations

	for the authors: (1) It would appear of value to emphasize what new information is being presented here. As the authors point out, the initial study of fibromyalgia included 293 patients in 265 controls which established a reliability of tender point examination. They correctly point out that few subsequent studies were done and only of 15 to 25 individuals. However, 39 patients are not substantially greater and results appear essentially similar to what is found in the literature. (2) The authors recognize the limitation of only 2 raters, which effectively excludes the possibility of any “outliers” which are seen in studies of 3 – 10 observers in analysis of physicians as measurers in clinical settings. This is also fewer than the initial report of Wolfe et al. more than 20 years ago. (3) There could be some sources of new information such as comparison of longitudinal data to determine how reproducible a tender point examination is to assess possible changes over time. (4) There is a concern that reliability in general of a physician's examination is less than self-report of patients, which was established in the in the initial report of the Health Assessment Questionnaire (HAQ) more than 30 years ago by Fries et al. One wonders, for example, whether a simple visual analog pain scale might not be as useful in assessing change in status in patients, though obviously would be very limited in explaining the etiology and mechanisms for pain. (5) Perhaps the 2 tables could be consolidated to 1 which actually might be more informative to compare intra- and inter-class correlation. (6) The smallest detectible difference of 5 and 3 tender points could actually classify many people as fibromyalgia or not. This is not the fault of the evaluation but the “state-of-the-art,” but affects the clinical value of tender points in low back pain. (7) One wonders what might be the advantage of performing the study in people who were told it would not affect their care? Patients may be more motivated to report more tender points when they think it will be recorded for their care. (8) While there is little doubt that 2 hours of training, including calibration of thumbs with a dolorimeter, will lead to greater reliability, this appears unrealistic in actual care. An examination requiring 6-8 minutes is not practical for usual rheumatology care—joint counts that would require 2 minutes are not performed by the majority of rheumatologists. These comments no way diminish the appropriate analysis of tender points in centers performing research to understand better mechanisms of how these develop. The reveiwer suggests that nuances of the tender point examination is would appear of interest primarily to sub-specialists within rheumatology. Fewer than 5% of rheumatologists perform tender point examinations although the criteria have been available for more than 20 years. Indeed, the newer versions of criteria emphasize that tender point examination is a continuum and recognize that self-report data are

	more likely to be clinically useful. This is a fine study, but these findings would not appear of importance to a general audience of physicians, for whom tender point examination is primarily of hypothetical interest.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1
Recent studies have shown that the majority of tender points in fibromyalgia are myofascial trigger points that cause local and referred pain. It would have been more useful if the authors could have addressed that issue, rather than looking only at tenderness.

Answer: The purpose of the study was not to compare tender points with trigger points, but to evaluate the reliability and agreement when using tender points as a measure of diffuse hyperalgesia. We are aware of the debate of overlap between tender points and trigger points (Bennett RM et al. Arthritis Research & Therapy 2011), but our focus was the total tender point count as a measure of insufficient pain regulation.
When evaluating questionnaires, Cronbach's alpha may be used as a measure of internal consistency, i.e. that all questions measure the same construct. However, Cronbach's alpha may also be used for measuring internal consistency of other instruments. We did not publish these figures in the first place since this measure has been developed for questionnaires. However, our study showed excellent Cronbach's alpha values which support the hypothesis of the tender point count as a measure of the same construct, e.g. insufficient pain inhibition. We have supplemented the article with the reporting of Cronbach's alpha.

The authors looked at the total tender point count. It would be more valuable if the authors could have determined that the same sites that were tender on the first examination were tender on the repeat examination, both over time for each examiner, and between examiners, rather than just looking at the total tender point count.

Answer: The secretaries also registered the number of painful sites during the examination. The reliability of TP examination at every single point is displayed as an appendix-table.

Reviewer 2

Reviewer: Theodore Pincus, Clinical Professor of medicine, NYU School of medicine - no competing interests

(1) It would appear of value to emphasize what new information is being presented here. As the authors point out, the initial study of fibromyalgia included 293 patients in 265 controls which established a reliability of tender point examination. They correctly point out that few subsequent studies were done and only of 15 to 25 individuals. However, 39 patients are not substantially greater and results appear essentially similar to what is found in the literature.

Answer: We have searched the literature for reliability and agreement studies on tender point examination, but found no study addressing the most important question: what is the reliability and agreement of the total tender point count. So, this is new in our study and the point was mentioned in the Introduction: 'However, only the reliability of testing each TP location as positive was estimated, not the reliability of the total TP counts.'
As pointed out in the article, there are good arguments for considering the reliability of the total count more important than the reliability of every single point. Even in the original article from 1990 (Wolfe et

al, Arthritis and Rheumatism), reliability and agreement data were not reported.

(2) The authors recognize the limitation of only 2 raters, which effectively excludes the possibility of any “outliers” which are seen in studies of 3 – 10 observers in analysis of physicians as measurers in clinical settings. This is also fewer than the initial report of Wolfe et al. more than 20 years ago.

Answer: More raters would require more patients and demand more examination days. Resources were limited in our clinical setting.

(3) There could be some sources of new information such as comparison of longitudinal data to determine how reproducible a tender point examination is to assess possible changes over time.

Answer: We agree that more studies of TPs are needed, including studies of change over time. However, to evaluate such studies is also required more knowledge about reliability and agreement data, and we have now provided a few such data.

(4) There is a concern that reliability in general of a physician's examination is less than self-report of patients, which was established in the in the initial report of the Health Assessment Questionnaire (HAQ) more than 30 years ago by Fries et al. One wonders, for example, whether a simple visual analog pain scale might not be as useful in assessing change in status in patients, though obviously would be very limited in explaining the etiology and mechanisms for pain.

Answer: We have analysed the associations between TPs and the reporting of back and leg pain in LBP patients (Jensen OK et al. Eur J Pain 2010). We certainly found an association, but neither of the two measures could replace the other.

(5) Perhaps the 2 tables could be consolidated to 1 which actually might be more informative to compare intra- and inter-class correlation.

Answer: We started by constructing one table, however, the table was never nice to look at or easy to interpret, and therefore the data were divided in two tables.

6) The smallest detectible difference of 5 and 3 tender points could actually classify many people as fibromyalgia or not. This is not the fault of the evaluation but the “state-of-the-art,” but affects the clinical value of tender points in low back pain.

Answer:
Our data should be interpreted in the following way: all patients with TP counts above 13 (Rater A) could be classified as fibromyalgia, if they had also widespread pain for more than 3 months, but patients with fewer TPs – even if higher than 10 – could not be classified as fibromyalgia with certainty because of the measurement error. Patients with 8 – 13 TPs (Rater A) and widespread pain might have fibromyalgia (probable fibromyalgia), and patients with less TPs would certainly not have fibromyalgia. The figures would be slightly different for Rater B. We would like to welcome similar reproducibility data in patients with widespread pain or arthralgia.

(7) One wonders what might be the advantage of performing the study in people who were told it would not affect their care? Patients may be more motivated to report more tender points when they think it will be recorded for their care.

Answer: All patients had experienced TP examination at their first visit. We do not believe that their expectations – or lack of expectations – had great influence on the result. If so, one would have expected more differences in the mean values between the two test days. It is difficult to explain why

expectations bias should be similar on two different test days.

8) While there is little doubt that 2 hours of training, including calibration of thumbs with a dolorimeter, will lead to greater reliability, this appears unrealistic in actual care. An examination requiring 6-8 minutes is not practical for usual rheumatology care—joint counts that would require 2 minutes are not performed by the majority of rheumatologists.

Answer: The 6-8 minutes included: patient undressing, measuring forward and side-flexion of the back, TP examination and dressing again. So, the TP examination took less than 6-8 minutes.

Indeed, the newer versions of criteria emphasize that tender point examination is a continuum and recognize that self-report data are more likely to be clinically useful.
This is a fine study, but these findings would not appear of importance to a general audience of physicians, for whom tender point examination is primarily of hypothetical interest.

Answer:
The paper presenting the new criteria of fibromyalgia (Wolfe et al 2010) pointed out that these criteria were not made to replace the previous criteria, but rather to provide supplementary criteria that can be used in surveys and used by clinicians who cannot or will not use TP examination.
There are numerous clinical examination techniques with poor reproducibility used in daily praxis, including swollen joint examination in arthritis patients (Chandran V et al, Arthritis and Rheumatism 2009, Tokka & Pincus, Rheum Dis Clin North Am 2009). However, standardised joint examination is still a requisite in the clinical care for arthritis patients.
We hope our validation study can inspire to a new way of using TP examination, e.g. an examination technique that may be used in different pain conditions for a better understanding of the underlying pain mechanisms and to be used in conditions where there is doubt about the clinical evaluation.
And we hope for more validation studies to improve the quality of clinical care.