Pain drawing evaluation—the problem with the clinically biased surgeon

Intra- and interobserver agreement in 50 cases related to clinical bias

Tomas Reigo[1], Hans Tropp[2] and Toomas Timpka[3]

To assess whether the clinical knowledge of the treating surgeon had any effect on the reliability of the pain-drawing evaluation, drawings from 50 low-back pain patients were evaluated by the treating surgeon and by three colleagues who had no clinical knowledge of the patient. The evaluation was repeated after 10 days. The treating surgeons were also blinded to clinical data. The kappa value in the evaluation when the surgeon had clinical knowledge of the patient was lower (0.29 (95% CI 0.13–0.45)) than the kappa value in the evaluations made without clinical knowledge (0.60 (CI 0.45–0.75)). The differences observed in interobserver reliability between open and blind evaluations suggest that clinical knowledge of a patient influences the evaluation of the pain drawings.

[1]Linköping Spine Centre, Institution of Orthopaedic Surgery , [2]Department of Orthopaedic Surgery, Eksjö Hospital, Sweden, [3]Department of Community Medicine, Faculty of Health Sciences, University Hospital, SE-581 85 Linköping, Sweden. Tel +46 13-224163. Fax -224050
Submitted 97-11-15. Accepted 98-03-08

The pain drawing was initially instituted as an aid to document the patient's pain in a quantitative and qualitative way. It gradually became apparent that much information could be gleaned from these drawings about the patient's psychological status. Since about 70–80% of all low back-pain patients show no obvious organic pathology and psychological factors are important for the development of chronic pain syndromes (Waddell et al. 1989), it would be of interest to identify their presence. The pain drawing has been reported to predict high scores on evaluation with MMPI (Minnesota Multiphasic Personality Inventory) for hysteria and hypochondria (Ransford et al. 1976), for other behavior changes (Von Bayer et al. 1983, Parker et al. 1995) and for outcome of treatment (McNeill et al. 1986, Takata and Hirotani 1995).

The pain drawing can be assessed at a glance, in contrast to a complicated scoring system. It is said that the pain drawing may allow the physician to identify nine tenths of the patients who are likely to need psychological support before surgery (Ransford et al.1976).

We assessed the degree of interobserver repeatability and intraobserver reproducibility of the pain drawing classification. We also examined whether the evaluation might be biased by the surgeon's clinical knowledge of the patient.

Patients and methods

We reviewed 50 consecutive patients (27 women) referred for surgeon's evaluation by general practitioners because of back pain. The patients had a mean age of 46 (25–78) years. The median duration of symptoms was 27 (3–180) months. No patients were being treated because of alcohol abuse or psychiatric disturbances at the time of evaluation. All patients were Scandinavians.

Two surgeons from the Spine Unit at the University Hospital in Linköping examined 25 patients each. The assessment included a standard questionnaire, a physical and radiological examination and a standard pain-drawing. The surgeon gave his clinical opinion and evaluated the pain drawings. The latter were classified according to Ransford et al. (1976) and Udén et al. (1988) into 4 categories: organic, possibly organic, possibly non-organic and non-organic. For statistical purposes, this 4-category classification was condensed into 2 groups, 1 organic and 1 psychogenic. Both surgeons also evaluated the pain drawings from their colleague's patient, including a psychogenic scoring, according to Sivik et al. (1992).

Thereafter, 2 other surgeons evaluated blindly all pain drawings with the Ransford /Udén method. After 10 days, all 4 surgeons evaluated the pain drawings once again, but in mixed order and blindly, concerning all other information, including identification.

Table 1. Interobserver reliability. First open evaluation versus first blind evaluation. Statistics in pairs by 2x2 contingency tables

First evaluation		Nontreating surgeon IV, blind		
		1	2	Total
Treating surgeons, open	1	24	15	39
	2	3	8	11
Total		27	23	50
Kappa 0.25 (CI 0.01–0.49); P pos 73%, P neg 47%				

First evaluation		Nontreating surgeon III, blind		
		1	2	Total
Treating surgeons, open	1	25	14	39
	2	2	9	11
Total		27	23	50
Kappa 0.33 (CI 0.23–0.56); P pos 76%, P neg 53%				

First evaluation		Nontreating surgeon III, blind		
		1	2	Total
Nontreating surgeon IV, blind	1	22	5	27
	2	5	18	23
Total		27	23	50
Kappa 0.60 (CI 0.38–0.82); P pos 81%, P neg 78%				

Table 2. Interobserver reliability. First evaluation. Blind versus blind observation. Statistics in pairs by 2x2 contingency tables

First blind evaluation		Nontreating surgeon III		
		1	2	Total
Treating surgeons	1	24	8	32
	2	3	15	18
Total		27	23	50
Kappa 0.55 (CI 0.20–0.90); P pos 81%, P neg 73%				

First blind evaluation		Nontreating surgeon IV		
		1	2	Total
Treating surgeons	1	26	6	32
	2	1	17	18
Total		27	23	50
Kappa 0.71 (CI 0.52–0.90); P pos 88%, P neg 83%				

To estimate the intraobserver reproducibility and interobserver repeatability, the kappa coefficient was generated by setting the observed proportion of agreement in relation to the proportion of agreement expected by chance. The kappa coefficient ranges from +1.0 (complete agreement) to 0.0 (chance agreement) to less than 0.0 (less agreement than expected by chance). No clear-cut interpretation of the kappa coefficient can be given, although some authors suggest a grading in which kappa values from 1 to 0.75, from 0.75 to 0.40, and from 0.4 to −1, would indicate respectively an excellent, a good-to-fair and a poor agreement (Koran 1975, Gjörup 1988, Seigel et al. 1992).

The kappa coefficient should be accompanied by separate values for the observed proportions of positive and negative agreements. This is specially important if the number of positive findings is small (Cicchetti and Feinstein 1990). The first observation consisted of 25 patients for whom the surgeon had clinical knowledge of sick history (open investigation) and 25 for whom surgeons I and II had no such knowledge (blind investigation). In the second observation, both groups of patients were blinded.

Surgeons III and IV made 2 sets of blind evaluations of the same 50 patients. The 2 treating surgeons' evaluations of the patients where the surgeon had not been the treating doctor were combined, so that the evaluated groups consisted of 50 patients in each group.

Intraobserver reliability was calculated for all surgeons. For the treating surgeons, the cases with clinical knowledge of the patients' history were combined and the cases for whom the surgeon had no clinical knowledge formed the second group. For the "non-treating" surgeons, a mean kappa value was calculated. The results are given with the weighted kappa value, together with the confidence interval of 95%. The chance of positive or negative agreement is also given according to Fleiss (1981).

Results

The assessments made by the non-treating surgeons on the day of the patient visit showed higher interobserver reliability than those made by both of the treating-non-treating surgeon pairs (Table 1). The interobserver reliability showed a higher kappa value, if both the treating surgeon and the non-treating surgeon had a blind observation at the first visit. (Table 2). At the assessment 10 days after the patient visit, there was no difference in the interobserver reliability between the non-treating and the treating-non-treating surgeon pairs (Table 3). The mean agreement for the treating-non-treating surgeon pairs was 0.29 (CI 0.13–0.45). For the non-treating pairs, the kappa value was 0.60 (CI 0.45–0.75). The intraobserver reliability for the treating surgeons evaluating their own patients showed a kappa value of 0.56 (CI 0.29–0.83) (Table 4).

Table 3. Interobserver reliability. Second evaluation. Blind versus blind observations. Statistics in pairs by 2x2 contingency table

Second blind evaluation		Nontreating surgeon III		
		1	2	Total
Treating surgeons	1	21	11	32
	2	3	15	18
Total		24	26	50
Kappa 0.45 (CI 0.22–0.68); P pos 75%, P neg 68%				
Second blind evaluation		Nontreating surgeon IV		
		1	2	Total
Treating surgeons	1	27	5	32
	2	2	16	18
Total		29	21	50
Kappa 0.71 (CI 0.51–0.91); P pos 89%, P neg 82%				
Second blind evaluation		Nontreating surgeon III		
		1	2	Total
Nontreating surgeon IV	1	22	7	29
	2	2	19	21
Total		24	26	50
Kappa 0.64 (CI 0.43–0.85); P pos 83%, P neg 81%				

Table 4. Intraobserver reliability comparing the biased surgeons' "open" evaluations to the biased surgeons' "blind" evaluations and the nonbiased surgeons' evaluation

Type of evaluation	Kappa	CI
Treating surgeon with knowledge of patients' findings	0.56	(0.29–0.83)
Nontreating surgeon without knowledge of findings	0.84	(0.74–0.94)

Discussion

Our findings suggest that knowledge of a patient's clinical status affects the interpretation of the pain drawing, previously not discussed in studies of pain drawings. In the second evaluation, the interobserver reliability had increased to kappa values indicating good agreement (> 0.40) for the 2 surgeons who had clinical knowledge of the patient. However, it can be argued as to whether the second observation made by the treating surgeon can rightly be considered as an independent observation, and this should be addressed in a future study. In addition, the intraobserver reliability findings suggest that the treating surgeon is affected by knowing the clinical history of the patient.

The reliability and reproducibility of various tests or clinical findings in orthopedic literature have been measured by correlation coefficients, and the use of kappa statistics has mostly been restricted to the reliability measurements of radiographic classifications (Andersen et al. 1990, Thomsen et al. 1991) and in later years to clinical applications (Strender et al. 1997). The instrument can been seen as a stronger statistical tool, diminishing the influence of chance. In the scientific literature, it is a widely used measure for agreement between observers of independent observations. However, it is also important to observe the proportion of positive agreement, which can be said to show the usefulness of the investigated parameter in clinical practice and not simply rely on the kappa value. In later years, the pain drawing has been used as a screening tool (Öhlund et al. 1996) and for prediction of treatment outcome (Takata and Hirotani 1995). It has also been used to classify the severity of pathological changes in disc herniation (Vucetic et al. 1995) and relate pain patterns to clinical findings (Brismar et al 1996). Some of the variations in predicting the outcome of the pain drawing may be explained by the clinical knowledge bias that we found.

Andersen E, Jorgensen L G, Hededam L T. Evan's classification of trochanteric fractures: an assessment of the interobserver and intraobserver reliability. Injury 1990; 21: 377-8.

Von Bayer C L, Bergström K, Brodwin S. Invalid use of pain drawings in psychological screening of back patients. Pain 1983; 16: 103-7.

Brismar H, Vucetic N, Svensson O. Pain patterns in lumbar disc hernia. Drawings compared to surgical findings in 159 patients. Acta Orthop Scand 1996; 67: 470-2.

Cicchetti D V, Feinstein A R. High agreement but low kappa. II. Resolving the paradoxes. J Clin Epidiol 1990; 43: 551-8.

Fleiss J L. Statistical methods for rates and proportions. Library of Congress Cataloguing in Publication Data, New York, USA 1981.

Gjörup T. The kappa coefficient and the prevalence of a diagnosis. Methods Inf Med 1988; 27: 184-6.

Koran L M. The reliability of clinical methods, data and judgments. N Engl J Med 1975; Sept: 642-6.

McCombe P F, Fairbanks J C T, Cockersole B C, Pynsent P B. Reproducibility of physical signs in low-back pain. Spine 1989; 14: 908-18.

McNeill T W, Sinkorra G, Leavitt F. Psychologic classification of low-back pain patients: a prognostic tool. Spine 1986; 11: 955-9.

Parker H, Wood P L, Main C J. The use of pain drawing as a screening measure to predict psychological distress in chronic low-back pain. Spine 1995; 20: 236-43.

Ransford A O, Cairns D, Mooney V. The pain drawing as an aid to the psychologic evaluation of patients with low-back pain. Spine 1976; 2: 127-34.

Seigel D G, Podgor M J, Remaley N A. Acceptable values of Kappa for comparison of two groups. Am J Epidemiol 1992; 135: 571-8.

Sivik T M, Gustafsson E, Klingberg Olsson K. Differential diagnosis of low-back pain patients: A simple quantification of the pain drawing. Nord J Psych 1992; 46/1.

Strender L-E, Sjöblom A, Sundell K, Ludwig R, Taube A. Inter-examiner reliability in physical examination of patients with low back pain. Spine 1997; 22: 814-20.

Takata K, Hirotani H. Pain drawing in the evaluation of low-back pain. Int Orthop 1995; 19: 361-6.

Thomsen N O B, Overgaard S, Olsen L H, Hansen H, Nielsen S T. Observer variation in the radiographic classification of ankle fractures. J Bone Joint Surg (Br) 1991; 73: 4: 676-8.

Udén A, Åström M, Bergenudd H. Pain drawings in chronic back pain. Spine 1988; 13: 389-92.

Vucetic N, Maatanen H, Svensson O. Pain and pathology in lumbar disc hernia. Clin Orthop 1995; 320 : 65-72.

Waddell G, Pilowsky I, Bond M R. Clinical assessment and interpretation of abnormal illness behaviour in low-back pain. Pain 1989; 39: 41-53.

Öhlund C, Eek C, Palmblad S, Areskoug B, Nachemson A. Quantified pain drawing in subacute low-back pain. Validation in a nonselected outpatient industrial sample. Spine 1996; 21: 1021-31.