CONCISE REPORT

Poor accuracy and interobserver reliability of knee arthroscopy measurements are improved by the use of variable angle elongated probes

S P Oakley, I Portek, Z Szomor, A Turnbull, G A C Murrell, B W Kirkham, M N Lassere

Ann Rheum Dis 2002;**61**:540–543

Objectives: (*a*)To determine the accuracy and reliability of arthroscopic measurements of cartilage lesion diameter in an artificial right knee model; (*b*) to determine whether the use of a set of variable angle elongated probes improves performance; and (*c*) to identify other sources of variability.
Methods: Ovoid "lesions" were drawn on the five cartilage surfaces of four plastic knees models. Two observers assessed these 20 lesions arthroscopically, measuring two diameters in orientations parallel and orthogonal to the probe. Observer 1 (orthopaedic surgeon) and observer 2 (arthroscopic rheumatologist) made two sets of measurements, firstly with the conventional probe and five months later with the variable angle elongated (VAE) probes. The knees were disarticulated to determine true lesion diameter.
Results: Observer 1 had negligible bias and good accuracy regardless of orientation or probe type. Observer 2 demonstrated both bias and poor accuracy using the conventional probe. Both improved using VAE probes. Poor interobserver reliability with conventional probes also improved using VAE probes. Major sources of variability could be traced to the probe type, the characteristics of the operator, and the orientation of the lesion in relation to the probe; the lesion location itself did not cause variability.
Conclusions: Variation in accuracy and poor interobserver reliability of measurements with conventional methods of cartilage lesion diameter measurement improved when specially designed measurement probes were used. Arthroscopic measurements performed as well as most clinical and radiographic measures. These findings have important implications for the use of arthroscopy as an outcome in multicentre trials where arthroscopists have different levels of experience.

Arthroscopy is a commonly performed procedure that permits direct visualisation and assessment of articular cartilage in the knee. Many arthroscopic methods have been developed to grade the severity of articular cartilage damage. At times arthroscopy has been used to measure outcome in therapeutic trials.[1–7] Some grading methods estimate lesion size as a percentage of the articular cartilage area occupied.[8] Others require estimation of lesion diameter.[9–11] Less is known about the accuracy and reliability of the latter.

Our three objectives were to determine the accuracy and reliability of arthroscopic measurements of cartilage lesion diameter in an artificial knee model, to determine whether the use of a set of variable angle elongated (VAE) probes improves performance, and to identify other sources of variability (lesion location, orientation of measurement, and factors attributable to the operator).

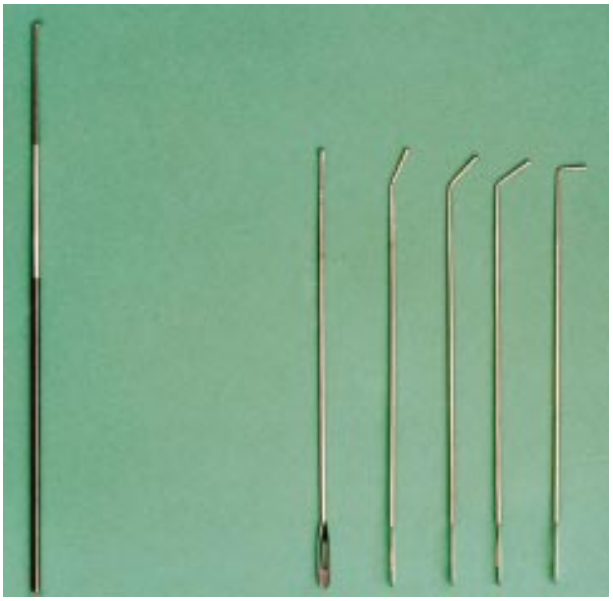

Figure 1 Arthroscopy probes. The conventional arthroscopy probe is shown to the left. The set of five specially designed measurement probes is shown to the right.

METHODS

Lesion diameter may be measured with the aid of a calibrated probe inserted into the joint. Conventional arthroscopic probes have a 3 mm nerve hook 90 degrees to the shaft, which itself has markings at 5 mm intervals. This shaft often cannot be laid flat on cartilage surfaces and parallax is a potential source of error. We (SPO) designed a set of probes with elongated (10–12 mm) foot processes at a range of different angles (VAE probes) allowing the examiner to lay the probe flat against the cartilage surface in the desired orientation (fig 1). These probes have markings at 2 mm intervals.

Two observers performed arthroscopic measurements in the plastic knee models using a 3.5 mm diameter Stortz arthroscope. Observer 1 was an orthopaedic surgeon with 10 years arthroscopic experience while observer 2 was a rheumatologist with three years diagnostic arthroscopic experience. Ovoid shaped lesions were drawn in ink on five cartilage surfaces (patella, medial and lateral femoral condyle, medial and lateral tibial plateaus) in four plastic knees (figs 2 and 3). The lesions drawn on the plastic knees corresponded to typical

Abbreviations: ANOVA, analysis of variance; ICC, intraclass correlation coefficients; SDD, smallest detectable difference; VAE, variable angle elongated (probes)

Figure 2 A plastic knee simulation model viewed in a constructed state.

lesions seen in actual arthroscopic procedures. The knees were filled with water and observers performed arthroscopy to measure the diameter of lesions in orientations parallel and then orthogonal to the shaft of the probe. Two sets of measurements were made by each observer with the conventional probe without knowledge of a future study using improved probes. The knees were then disarticulated and true lesion diameter determined.

Five months later the same two observers were asked to repeat the measurements of the same 20 ovoid lesions, now using VAE probes. The observers were allowed to choose from among probes with foot processes set at a range of different angles (0, 30, 45, 60, and 90 degrees). Neither observer saw the plastic knees in a state of disarticulation. No specific feedback about the results of measurements with the conventional probe was given in the interim.

Statistical methods

Three complementary statistical methods were used to evaluate the accuracy of the observed diameter and the reliability of the measurements within and between observers.

Method 1

Intraclass correlation coefficients (ICCs) were determined from analysis of the components of variance (ANOVA). Although they are usually used as a measure of reliability, they can be used as a measure of accuracy when the true lesion diameter is known. We have used a random effects ICC as described by Shrout and Fleiss.[12] ICCs are a relative measure of agreement and therefore are biased towards higher coefficients (1.0 is perfect agreement, 0.0 is no agreement) if the lesions measured vary over a larger range of values.

Method 2

The 95% confidence limits of the standard deviation of the difference of the true diameter minus the observed diameter was also used to estimate accuracy and reliability. The statistic has been called the smallest detectable difference (SDD)[13] and is derived from the limits of agreement method[14] and quantifies random error, whereas the mean of the difference scores (also known as the paired *t* test) quantifies systematic error (also known as bias). The ICC cannot discriminate between random and systematic error. However, the SDD, unlike the ICC, is an absolute measure of agreement, therefore it is biased toward smaller values (SDD of 0 is perfect agreement, and there is no convention that anchors the upper limit) if the lesions are measured over a narrower range of values. Accuracy was defined as the observed minus the true diameter. Thus positive difference scores represent overestimation and negative scores underestimation.

Method 3

ANOVA[15] was used to determine which components of the variance were statistically significant.

RESULTS

With the conventional probe, observer 1, an experienced orthopaedic surgeon, had overall satisfactory accuracy and intraobserver reliability (table 1). His ICCs were all greater than 0.75. He had essentially no bias (three of four mean difference scores were 0, but there was moderate random error (SDD ±3 and ±5 for the two orientations) for both accuracy and reliability.

To illustrate the difference between the ICC and SDD approaches, it should be noted, firstly, that there was a larger range of values for the orthogonal (4–36 mm) than the parallel (5–17 mm) orientation. The ICC was better (that is, larger) in the orthogonal than the parallel (0.95 *v* 0.77), whereas the SDD was better (that is, smaller) in the parallel than in the orthogonal (±3 *v* ±5) orientation.

Observer 2, in contrast, had poor accuracy but satisfactory intraobserver reliability; in other words he was consistent in his inaccuracy with the conventional probes (table 1). The poor accuracy was due to both considerable systematic bias (underestimation of 4 and 6 mm in the two orientations) and random error (±5 and ±6). The interobserver reliability was poor. There were substantial systematic bias and random differences between observer 1 and observer 2.

Upon re-test several months later, observer 1 showed modest additional improvement in both accuracy and reliability with use of VAE probes. Random error, although less, was still present; its relative importance is discussed later. With this repeat study, observer 2 showed a substantial improvement in

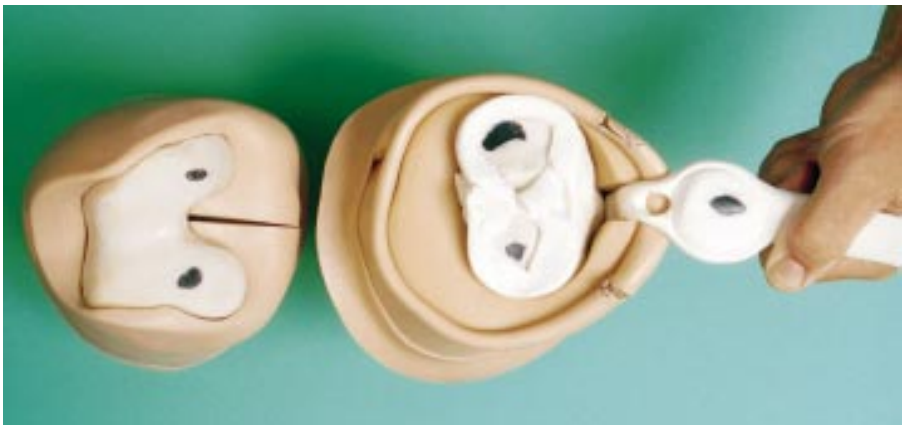

Figure 3 A plastic knee simulation model disarticulated to show "lesions" drawn on "cartilage surfaces".

Table 1 Accuracy and reliability of diameter measurements

Performance orientation	True diameter mean min–max (n=20)	Observer 1						Observer 2						Interobserver: observer 1 v observer 2					
		Conventional probe			VAE probes			Conventional probe			VAE probes			Conventional probe			VAE probes		
		ICC	$Diff_{mean}$	± SDD (± % SDD)	ICC	$Diff_{mean}$	± SDD (± % SDD)	ICC	$Diff_{mean}$	± SDD (± % SDD)	ICC	$Diff_{mean}$	± SDD (± % SDD)	ICC	$Diff_{mean}$	± SDD (± % SDD)	ICC	$Diff_{mean}$	± SDD (± % SDD)
Accuracy																			
Parallel	8 (5–17)	0.82	0	3 (18)	0.93	0	2 (12)	0.14	−4	5 (29)	0.66	−1	4 (24)						
Orthogonal	14 (4–36)	0.93	1	5 (14)	0.96	0	4 (11)	0.59	−6	6 (17)	0.93	−2	4 (11)						
Reliability																			
Parallel	8 (5–17)	0.77	0	3 (18)	0.92	0	2 (12)	0.63	0	2 (12)	0.71	0	4 (24)	0.18	−4	4 (24)	0.58	0	5 (29)
Orthogonal	14 (4–36)	0.95	0	5 (14)	0.98	1	3 (8)	0.94	0	4 (11)	0.96	0	4 (11)	0.54	−8	6 (17)	0.95	2	4 (11)

VAE probe, variable angulated elongated probe (see text); accuracy, observed diameter minus true diameter; intraobserver reliability, agreement between 1st versus 2nd sets of estimates by each observer; interobserver reliability, agreement between the 1st sets of estimates by each observer; orientation, orientation of probe in relation to lesion (see fig 1); ICC, intraclass correlation coefficient (2, 1) for agreement between observed diameter and true diameter (accuracy) and between repeated measurements; $diff_{mean}$, mean of the difference scores (observed diameter minus true diameter); SDD, ± 1.96 standard deviations of the difference scores; (% SDD), SDD expressed as a percentage of maximum lesion size ([SDD/max diameter] × 100).

accuracy, now approaching that seen with observer 1. Specifically, he no longer underestimated lesion diameter (underestimation reduced from 50% to 13% in the parallel and 43% to 14% in the orthogonal orientation), and this accounted for the bulk of his improvement. His reliability (random error) did not improve. However, interobserver reliability improved. Both systematic bias and random differences decreased, although the random component remained sizeable, particularly in the parallel orientation.

These results were confirmed by ANOVA, which evaluated, singly and in combination, sources of variability (observer, location of lesion, orientation of probe, and probe type; table 2). There was significant difference in accuracy between observers (p=0.00) because observer 2 was influenced by orientation (p=0.00) and type of probe (p=0.00) whereas observer 1 was not. Location of the lesion within the joint did not affect the accuracy (p=0.10). Assessment of within observer and between observer reliability showed that the repeated measurements of observer 1 significantly differed only by orientation of measurement (p=0.00), whereas those of observer 2 differed by location (p=0.01), orientation, and probe type (p=0.00). These three factors were also significant sources of interobserver variation. One way ANOVA with multiple comparisons indicated that there was a significant difference between variance of measurements in the parallel orientation on the medial versus the lateral tibia (p=0.004). There was no difference in true lesion diameter in this orientation.

The possibility of continued improvement (that is, memory or learning effect) was assessed when the observers performed a third set of measurements (data not shown) using the conventional probe. There was no improvement in accuracy or reliability over time.

DISCUSSION

This study determined the accuracy and reliability and explored sources of variability of arthroscopic measurements of cartilage lesion diameter in an artificial knee model. Overall, with the use of the VAE probes, accuracy and reliability were both good. The percentage SDD varied from 8 to 24%, signifying that arthroscopic measurements performed as well as most clinical and radiographic measures.[16] Major sources of variability could be traced to the probe type, the characteristics of the operator, and the orientation of the lesion in relation to the probe; the lesion location itself did not affect variability.

We showed that systematic over- or underestimation of lesion diameter (bias) with conventional probes was operator dependent, but bias and, consequently, differences between operators decreased substantially with the use of improved (VAE) probes. Accuracy, as determined by agreement statistics (degree of random error), also improved with the use of the VAE probes, and although this improvement was less dependent on the operator, it was influenced somewhat by the method of statistical analysis employed. Intraobserver reliability was also unaltered with the VAE probes but, importantly, interobserver reliability improved considerably.

This study was conducted in a highly artificial, albeit favourable, situation to study the effects of image distortion, among other factors, on arthroscopic measurements. Studies in plastic knees have a number of advantages. Lesions on "cartilage surfaces" were dark coloured with clearly defined margins. In such circumstances one would expect good accuracy and reliability. Comparison with a "gold standard" measurement upon disarticulation may be performed, and measurements of reliability can incorporate variations in procedural technique such as the ability to reproduce the positioning of the arthroscope and the probe. These studies are difficult to undertake in vivo because conditions change during the procedure as synovium becomes oedematous and obscures vision, and cartilage damage may change between

Table 2 Effects upon accuracy and reliability analysed by ANOVA

	Accuracy						Reliability					
	Both Observers		Observer 1		Observer 2		Both Observers		Observer 1		Observer 2	
Source	F	p	F	p	F	p	F	p	F	p	F	p
Model	23.69	0.00	1.97	0.07	21.94	0.00	20.02	0.00	10.34	0.00	12.03	0.00
Observer	139.15	0.00	N/A	N/A	N/A	N/A	33.89	0.00	N/A	N/A	N/A	N/A
Location	2.19	0.07	1.31	0.27	2.06	0.09	4.78	0.00	2.15	0.08	3.23	0.01
Orientation	6.64	0.01	3.40	0.07	27.73	0.00	79.31	0.00	53.04	0.00	30.03	0.00
Probe type	32.14	0.00	3.19	0.08	95.65	0.00	7.85	0.01	0.39	0.53	29.25	0.00
R^2	0.37		0.07		0.46				0.29		0.32	

Accuracy, difference between true diameter and observed diameter; reliability, repeated measurements; N/A, not applicable.

repeated arthroscopic procedures. By the nature of the experiment, observers were aware of the type of probe used. However, observers were unaware that future study was planned with improved probes, they were always unaware of the true lesion size, and the time between measurements using the conventional and VAE probes was large.

Several source of variability were demonstrated. One factor was training and experience of the operator. The better accuracy of observer 1, an orthopaedic surgeon, may be explained by his greater arthroscopic and open knee surgery operating experience. This would improve spatial awareness within the knee joint, thereby compensating for distortions created by arthroscopy. By contrast, observer 2, was a rheumatologist with experience with over 200 knee arthroscopies but no experience with open knee surgery and compensation for distortion.

Another factor was orientation of the lesion in relation to the probe. Overall, both systematic and random accuracy were less in the parallel than in the orthogonal orientation. Although this seems counterintuitive, it is probably due to a greater parallax effect because the conventional probe cannot easily be lain flat on the cartilage surface, especially by a less experienced operator. The design of the VAE probe dealt with this difficulty and improved both the accuracy and interobserver reliability. The improved accuracy with the VAE probes may also be due to its finer calibrations (2 mm) and elongated foot processes, which permit easier measurement. However, the greater improvement still occurred in the parallel orientation, not the orthogonal.

These methods are yet to be tested in vivo, where additional challenges will also be encountered. Different grades of cartilage lesion with poorly defined margins are common, so one would expect reduced accuracy and reliability. However, it may be that in this context the true value of the VAE probes is apparent. Furthermore, other easily implemented techniques may be found to improve accuracy and reliability of arthroscopic measurements.

The implications of these types of studies of the fundamentals of measurement are considerable. Arthroscopic outcome measures will continue to play an important part in test-of-concept studies and for validation of full scale clinical trials of agents directed at cartilage protection. These trials are likely to be multicentre and involve arthroscopists of differing backgrounds and differing levels of experience. Although arthroscopy is an invasive procedure, it remains the most sensitive method for detecting early cartilage changes at a time when cartilage protection may have the greatest benefit. It also remains the only method to directly assay physical and biomechanical properties of cartilage, which may be important outcomes themselves. Thus arthroscopy remains of interest as a highly informative outcome measure in arthritis.

ACKNOWLEDGEMENTS

Dr Oakley is supported by a research scholarship from the Arthritis Foundation of Australia. We thank the staff at the operating theatres of St George Hospital for the use of their facilities, the St George Hospital Department of Orthopaedic Surgery for the use of the plastic knee teaching models, and Kempsey Valley Arks for their financial support.

Authors' affiliations
S P Oakley, I Portek, M N Lassere, Department of Rheumatology, St George Hospital, University of New South Wales, Sydney, Australia
Z Szomor, A Turnbull, G A C Murrell, Department of Orthopaedics, St George Hospital
B W Kirkham, Department of Rheumatology, Guy's and St Thomas' Hospital Trust, London, UK

Correspondence to: Dr S Oakley, Department of Rheumatology, St George Hospital, Gray St, Kogarah, NSW 2217 Australia; z8425086@student.unsw.edu.au

Accepted 7 January 2002

REFERENCES

1 **Listrat V**, Ayral X, Patarnello F, Bonvarlet JP, Simonnet J, Amor B, *et al*. Arthroscopic evaluation of potential structure modifying activity of hyaluronan (Hyalgan) in osteoarthritis of the knee. Osteoarthritis Cartilage 1997;5:153–60.
2 **Listrat V**, Dougados M, Ayral X. Effects of intra-articular (IA) injections of hyaluronic acid (HA) (Hyalectin) on the anatomical course of OA (OA) [abstract]. Rev Rhum Engl Ed 1993;60:583.
3 **Frizziero L**, Govoni E, Bacchini P. Intra-articular hyaluronic acid in the treatment of osteoarthritis of the knee: clinical and morphological study. Clin Exp Rheumatol 1998;16:441–9.
4 **Brittberg M**, Lindahl A, Nilsson A, Ohlsson C, Isaksson O, Peterson L. Treatment of deep cartilage defects in the knee with autologous chondrocyte transplantation. N Engl J Med 1994;14:889–95.
5 **Franssen M**, Boerbooms A, Karthaus RP. Treatment of pigmented villonodular synovitis of the knee with yttrium-90 silicate: prospective evaluations by arthroscopy, histology and Tc pertechnate uptake measurements. Ann Rheum Dis 1989;48:1007–13.
6 **Fujisawa Y**, Masahura K, Shiomi S. The effects of high tibial osteotomy on OA of the knee. An arthroscopic study of 54 knee joints. Orthoped Clin North Am 1979;10:585–608.
7 **Raatikainen T**, Vaananen K, Tamelander G. Effect of glycosaminoglycan polysulfate on chondromalacia patellae. A placebo controlled one year study. Acta Orthopaed Scand 1990;61:443–8.
8 **Dougados M**, Ayral X, Listrat V, Gueguen A, Bahuaud J, Beaufils P, *et al*. The SFA system for assessing articular cartilage lesions at arthroscopy of the knee. Arthroscopy 1994;10:69–77.
9 **Noyes FR**, Stabler CL. A system for grading articular cartilage lesions at arthroscopy [review]. Am J Sports Med 1989;17:505–13.
10 **Klashman D**, Ike R, Moreland L, Skovron M, Kalunian K. Validation of an OA data report form for knee arthroscopy [abstract]. Arthritis Rheum 1995;38 (suppl 9):154.
11 **Outerbridge R**. The etiology of chondromalacia patellae. J Bone Joint Surg Br 1961;43:752–7.
12 **Shrout P**, Fleiss J. Intra-class correlations: using in assessing rater reliability. Psychol Bull 1979;86:420–8.
13 **Lassere M**, Boers M, van der Heidje D, Boonen A, Edmonds J, Saudan A, *et al*. Smallest detectable difference in radiological progression. J Rheumatol 1999;26:731–9.
14 **Bland JM**, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i:307–10.
15 **STATA Corporation**. Intercooled STATA version 6. 1999.
16 **Lassere M**, van der Heidje D, Johnson K, Boers M, Edmonds J. The reliability of measures of disease activity and disease damage in rheumatoid arthritis: implications for the smallest detectable difference (SDD), the minimum clinical important difference (MCID) and the analysis of treatment effects in randomised controlled trials. J Rheumatol 2001;28:892–903.