Reproducibility of instrumented knee joint laxity measurement in healthy subjects

M. van der Esch¹, M. Steultjens¹,²,³, R. W. J. G. Ostelo³,⁴, J. Harlaar²,⁵ and J. Dekker¹,²,³

Objective. To determine the reproducibility of frontal plane knee joint laxity measurement through the assessment of intra- and inter-rater reliability coefficients and intra- and inter-rater agreement coefficients.

Methods. Two raters independently assessed the laxity of the knee joint in the frontal plane by three repeated measurements. Fourteen days later the assessment was repeated. Complete data were obtained from 20 healthy subjects. Laxity was assessed using a device which consisted of a chair with a free-moving arm that supported the subject’s lower leg. Medial and lateral loads were applied, resulting in a varus and valgus movement in the knee joint. The intra- and inter-rater reliability coefficients [intraclass correlation coefficients (ICC)] were estimated, as were the intra- and inter-rater agreement parameters [standard error of measurement (SEM) and minimal detectable difference (MDD)].

Results. Adequate intra-rater reliability (ICC > 0.80) was calculated for each rater’s measurements of laxity. The inter-rater reliability was less adequate (ICC = 0.65) when calculated using the first day’s measurements. However, inter-rater reliability was adequate (ICC = 0.88) when calculated using the day 14 measurements. The intra-rater measurement error calculated across occasions was 1.3° for individual subjects. This resulted in an MDD of 3.7°. The inter-rater measurement error, i.e. the SEM and MDD, was higher (1.5° and 4.3°, respectively).

Conclusions. Intra-rater reliability of knee joint laxity measurement is good. Adequate training of raters establishes the basis for good inter-rater reliability. In clinical trials, it is preferable for one trained rater to perform the laxity measurement. The measurement of knee joint laxity is limited due to its relatively high measurement error in individual subjects; therefore, this measurement should be restricted to group assessment rather than individual patient assessment.

Key words: Reproducibility, Reliability, Knee joint laxity, Osteoarthritis.

Frontal plane knee joint laxity may play an important role in knee osteoarthritis (OA). Laxity can be defined as the angular deviation of the tibia-femoral joint in the frontal plane after varus–valgus load is applied [1, 2]. Laxity is related to radiographic progression and to poor functional outcome [3–8]. Although laxity has been identified as an important factor in OA of the knee, detailed information on the clinimetric properties of its measurement is unavailable.

Measuring laxity equates to measuring small differences in varus–valgus deviations. To detect minimal differences in laxity, high-precision measurement with high reproducibility is essential.

Reproducibility concerns the degree to which repeated measurements in a constant situation provide similar answers. For the quantification of reproducibility, two types of measure can be distinguished: measurements of reliability and measurements of agreement. Reliability parameters assess whether persons in a group can be distinguished from each other, despite measurement errors [9]. Reliability is expressed as the intraclass correlation coefficient (ICC), ranging from 0 to 1 [9]. A high ICC represents a sufficient distinguishing capacity of the instrument regardless of measurement error. In order to identify precise measurement, the absolute measurement error has to be taken into account. Expressing the measurement error in scale points is often referred to as agreement. Agreement parameters assess how close the results of the measurements are within individual subjects by estimating the absolute measurement error in repeated measurements [10, 11]. Agreement in measuring joint laxity is expressed as the standard error of measurement (SEM) in degrees and the minimal detectable difference (MDD) in degrees.

Currently, there is very limited information regarding the reproducibility of the measurement of knee joint laxity. In two studies [4, 5], the reliability was tested on four and five patients, respectively, with an intra-rater reliability of 0.92 (ICC). Sharma et al. [1, 2, 6, 7] presented reliability scores ranging from 0.84 to 0.90 (ICC). Information regarding inter-rater reliability and agreement parameters is presently unavailable. For this reason there is an evident need to examine the reproducibility of the measurement of laxity in the knee.

The objective of this study was to establish (i) the intra- and inter-rater reliabilities and (ii) the intra- and inter-rater agreement parameters of the measurement of knee joint laxity.

Methods

Subjects

Twenty healthy young volunteers (10 males, 10 females) participated in the study. The mean ± S.D. age of the subjects was

Received 23 May 2005; revised version accepted 4 November 2005.

Correspondence to: M. van der Esch, Jan van Breemen Institute, Center for Rehabilitation and Rheumatology, Jan van Breemenstraat 2, 1056 AB Amsterdam, The Netherlands. E-mail: M.van.der.Esch@planet.nl

© The Author 2005. Published by Oxford University Press on behalf of the British Society for Rheumatology. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
22.9 ± 3.0 yr. The inclusion criteria were no current knee pain; no previous injury in the hip-knee region; no analgesics or anti-depressive medication; and, for women, regular menstrual cycles for the 3 months prior to the study. All of the above criteria may influence the degree of laxity. Ethical review board approval was obtained, and all participants provided written informed consent.

Design

Two raters (a physical therapist and a human movement scientist), both trained in clinical measurements by a clinician, independently performed all the laxity measurements. The subjects were scheduled for the two experimental sessions (days 1 and 14). On both occasions the raters measured the subjects. Each rater measured the same knee of each subject three times. In 10 subjects the right knee was measured and in 10 other subjects the left knee.

Each rater made three consecutive measurements and the subjects remained seated and fixed between measurements. The loading was applied and then released after each measurement. The deviation in the subject's knee was recorded digitally. After the first rater had assessed the joint laxity, all fixation points were removed and the subject stood up. Subsequently the second rater seated, fixed and assessed the same subject. To avoid bias, the second rater waited in another room while the first rater performed the measurements.

After 14 days the procedure was repeated; the order of raters was reversed. Both raters were blinded to the results of the reproducibility analyses of the day 1 measurements.

Equipment

An electronic device (Fig. 1) was used to measure knee varus-valgus laxity. A chair with an attached free-moving arm, which supported the subject’s lower leg, was used to seat the subject. The subject was seated comfortably in the measurement chair, which had a back support. The device was constructed in such a manner that throughout the study the knee joint was held in 20° of flexion.

The thigh, lower leg and ankle were fixed to the device. No medial or lateral movement of the lower leg and thigh or internal and external rotation of the hip was possible using these fixation techniques. The thigh and lower leg were fixed at five places. The foot and distal part of the lower leg were fastened to the arm using clamps at the ankle and at the distal part of the leg (Fig. 1; points 1 and 2). Below the knee the lower leg was fixed to the device with a Velcro bandage (Fig. 1; point 3). The distal/lower part of the thigh was fixed using two clamps (Fig. 1; point 4). The upper thigh was fastened to the chair using a Velcro bandage (Fig. 1; point 5).

The joint of the arm moved with minimal friction. The axis of rotation of the free-moving arm was centrally located directly under the tibiofemoral joint of the subject (i.e. the middle of the popliteal fossa). To supply a steady moment to the knee of 7.7 Nm, a dead-weight was used. This weight was attached to the free-moving arm by a cord. The cord was attached 0.68 m from the axis of rotation of the arm. The load could be applied to the lower leg both medially and laterally, resulting in varus or valgus movement in the knee joint. An electronic measurement system digitally recorded the end point of the varus or valgus movement, after 4 s.

Joint laxity measurement

All measurements of laxity were performed in accordance with our protocol, including the use of anatomical landmarks for patient positioning, patient instructions and the examiner’s position.

Joint laxity measurement

Anatomical landmarks of the knee were palpated to localize the medial and lateral joint spaces and the middle of the popliteal fossa. These anatomical structures give an indication of the position of the varus-valgus rotation axis of the tibia-femoral joint of the knee. The electronic meter was positioned in line with the varus–valgus rotation axis (Fig. 1; point B).

To avoid increased muscle tone resulting from pain during the fixation or measurement, subjects were instructed to relax as much as possible and to report the onset of pain.

Raters were seated behind the patient and applied the load slowly by hand to the lower leg in a standardized manner.

Analysis

The mean score in degrees for laxity obtained from the three measurements was used for analysis. Reproducibility was assessed using the following sources of variance: subject, rater, time of measurement and interaction between these variables. To express reproducibility, the following parameters were established [11, 12].

Intra-rater reliability. The ICC (2,k) was calculated as the ratio of variance between subjects within one rater, in relation to the relative measurement error (including all sources of variance: rater, subject, time of measurements and the absolute measurement error).

Intra-rater agreement. The SEM concerns the absolute measurement error in measuring an individual. It assesses the proximity of the scores on repeated measures [10, 11, 13]. The amount of measurement error can be expressed as the SEM. The SEM was derived by taking the square root of the error variance of the following sources of variance: time of measurement, interaction between subject and time of measurement, interaction between rater and time of measurement, and interaction between subject, rater and time of measurement. The SEM was calculated across both occasions. The SEM was used to calculate the MDD. The MDD is the smallest measurable difference that can be interpreted as a real difference between
two measurements, i.e. beyond zero [10, 11]. To compute the MDD as the 95% confidence limit of the SEM, the SEM has to be multiplied by 1.96 [for the 95% confidence interval (CI)] and by the square root of 2 for the difference scores \((1.96 \times \sqrt{2} \times \text{SEM})\). The MDD expresses the uncertainty of the difference between two observed scores [14].

**Inter-rater reliability.** The ICC was calculated as the ratio of variance (rater, subject, time of measurement and the absolute measurement error) between subjects and between the two raters, in relation to the relative measurement error.

**Inter-rater agreement.** The SEM was calculated to establish the absolute measurement error across raters and occasions, calculated according to the generalizability theory [9]. The SEM was derived by taking the square root of the error variance of the following sources of variance: rater, time of measurement, interaction between rater and subject, interaction between rater and time of measurement, interaction between subject and time of measurements, and the interaction between subject, rater and time of measurement. The SEM was used to calculate the MDD. The MDD was also calculated across raters and occasions.

In order to visualize the difference between raters against the corresponding mean of the two raters for each subject, a limits-of-agreement plot was constructed, as proposed by Bland and Altman [15].

For reliability, an ICC of \(>0.70\) was regarded as adequate [16]. Confidence intervals were presented as an indication of the precision of the point estimate. To calculate the ICC, the SEM and the MDD, a two-way random effects model of analysis of variance (ANOVA) was performed, using the Statistical Package for the Social Sciences (SPSS) version 12.0. Windows (SPSS, Chicago, IL, USA).

### Results

**Subjects**

The study sample consisted of 20 healthy subjects. The demographic data of the subjects are presented in Table 1. For rater A, the mean scores in knee joint laxity on the first (day 1) and second assessment (day 14) were 5.5 and 6.2\(^\circ\), respectively. For rater B, the mean scores were 5.5 and 6.5\(^\circ\) at day 1 and 14, respectively.

**Intra-rater reliability**

The ICC for rater A was 0.84 (95% CI 0.61, 0.94) and 0.93 (95% CI 0.81, 0.97) for rater B.

**Intra-rater agreement**

Generalized across occasions for the same fixed rater, the measurement error, expressed as the SEM, was 1.35\(^\circ\) and the MDD was 3.73\(^\circ\) (Table 2).

**Inter-rater reliability**

The ICC was 0.65 (95% CI 0.13, 0.86) for the assessment on day 1 and 0.88 (95% CI 0.70, 0.95) for the assessment on day 14.

**Inter-rater agreement**

The SEM was 1.55\(^\circ\) and the MDD 4.30\(^\circ\), generalized across raters and occasions. This result represents the absolute measurement error when a subject has been measured on a first occasion by a rater and the same subject is also measured on a second occasion by a second rater. The agreement coefficients are presented in Table 2.

Figure 2 shows the difference between raters on day 14 plotted against the mean value of both raters for each subject for varus-valgus laxity of the knee joint. Solid line shows the mean difference \((-0.026\); dashed lines show the 95% limits of agreement \((3.38 to -3.44\)).

### Discussion

In this study the reproducibility of knee joint laxity measurement was quantified using generalized reliability parameters and
agreement parameters in healthy, stable subjects. The ICC as an intra-rater reliability coefficient expresses the measured variance within one rater on two occasions. In our study the ICCs were found to be adequate for both raters (0.84 and 0.93, respectively). The ICC as an inter-rater reliability coefficient expresses the measured variance between two raters on the first and second occasions. The ICC was low (0.65) on the first occasion (day 1) and adequate (0.88) on the second occasion (day 14). Measurement of intra-rater agreement parameters is important in quantifying measurement error. In our study the intra-rater SEM was 1.39.

When the measurement was repeated by the same rater on the same subject the MDD was 3.79. This expresses (with an uncertainty of <5%) that a difference between two measurements of less than 3.79 is attributable to measurement error and can therefore not be interpreted as a real difference. Only a difference in measurements made by the same rater exceeding 3.79 is likely to signify a real change in laxity.

Inter-rater agreement parameters express the absolute measurement error when a rater measures an individual subject on one occasion and a second rater measures the same subject on a second occasion 14 days later. In our study the inter-rater SEM was 1.59. This indicates the absolute measurement error generalized for occasions and raters. The MDD was 4.39, which indicates that this is the smallest difference between two measurements made by different raters at different times that can be interpreted as a genuine change.

To assess reproducibility we used healthy, stable subjects. It was assumed that the biological variation in the group, i.e. the variability in laxity of the knee joint, was small. The raters were well instructed, trained and measured in accordance with a given protocol. Compared with other studies involving clinical subjects [1, 2, 4, 5], our intra-rater reliability coefficients were lower. The heterogeneity of the population in previously conducted clinical studies could explain the difference from our study. A small range of laxity in healthy subjects makes the distinction between subjects more difficult, compared with a patient population with higher variability. In a patient population the subjects are easier to rank, because the difference between subjects is greater than the difference between subjects in a healthy population. Consequently, the ICC will be lower in healthy subjects than the ICCs in clinical studies. To compare the inter-rater reliability, no other studies are available. The inter-rater reliability coefficient was substantially higher on day 14 compared with day 1. In the day 14 session reliability was good. Although the raters had some previous experience in knee assessment, it is conceivable that experience gained through the knee measurements in this study resulted in a higher reliability coefficient. The increased experience could explain the higher reliability on day 14.

Inter-rater reliability was lower than intra-rater reliability. Therefore, using one trained rater to perform all laxity measurements is recommended.

One source of error could be the fixation of the lower leg and thigh of subjects. The lower leg and thigh were fixed in five places. Possible reasons for variation in the fixation points which can account for measurement variance are: (i) small differences in the positioning of the leg during fixation between raters and between occasions, and (ii) possible pain in the lower leg and thigh during the measurement. In our study a load of 7.7 Nm was used. Sharma et al. [1, 2] used a load of 12 Nm. In a study applying this load to patients with OA of the knee, we found that a load of 12 Nm induced pain in some patients [8]. Hence, we decided to reduce the applied load to 7.7 Nm. This reduced load was not painful for any of the patients tested. However, it is recommended that the patient’s exposure should be limited to the minimum number of measurement readings needed to obtain a reliable result, and that attention should be paid to possible discomfort or pain during the measurements, in order to prevent any adverse effects. Although the mean scores of laxity in our study are similar to those in the Sharma et al. study [1, 2, 7], the results are not comparable because of the technical differences between the devices and the different loads used in the measurement.

Our results suggest that laxity measurements are of limited use in clinical practice, because of considerable measurement error. However, in research precision can be increased by including more subjects. For clinical trials related to laxity, an adequate number of subjects should be included, based on a power analysis.

In conclusion, these results on the reproducibility of knee joint laxity measurement indicate that the intra-rater reliability is good. The inter-rater reliability is less adequate on the first test occasion and good on the second test occasion. In a setting in which both raters are well trained, it is possible to achieve acceptable inter-rater reliability. The interpretation of results of the measurement of frontal plane knee laxity at the individual level is limited because of measurement error.

Acknowledgements
We gratefully acknowledge Ms L. Pernet and Ms M. C. Boucher for their assistance in obtaining the data, Ms K. Fiedler for her assistance in correcting the manuscript, Mr D. Knol for his assistance with the statistical analysis and Mr A. v. Vark of Enraf-Nonius, Delft, The Netherlands, for manufacturing the knee laxity measuring device.

The authors have declared no conflicts of interest.

References
Clinical Vignette

Nailfold capillaroscopic picture by chance

A 71-yr-old man was referred for evaluation of complaints of Raynaud’s phenomenon. He had a 3-yr history of cold-induced bluish discoloration, and some broadening and progressive (although not yet disabling) stiffness of his fingers. He underwent bypass grafting because of intermittent claudication due to atherosclerosis of the aorta and iliac arteries and was treated for his hyperhomocysteinaemia. He did not remember a tick bite or erythema chronicum migrans. Physical examination did not reveal sclerotic skin lesions or arthritis.

Capillaroscopy of all nailfolds was performed. A normal capillary density of short capillary loops was observed in all nailfolds. Remarkably, the left fifth finger showed marked architectural disarrangement of the capillaries apart from a normal appearance of the distal rows; there were slightly enlarged capillary loops surrounding an avascular area. Serological testing revealed the presence of antibodies to Borrelia burgdorferi and confirmed the diagnosis acrodermatitis chronica atrophicans (ACA), a feature of Lyme borreliosis. Antinuclear antibody tests were negative. After treatment with doxycycline (100 mg twice daily for 30 days), capillary microscopy showed a completely normalized pattern.

Capillaroscopy is a simple, uninvasive imaging technique to study the microcirculation and is one of the diagnostic tools for early detection of systemic sclerosis and related disorders. This capillaroscopic image can be added to the morphological features of the diseases in the rheumatological field. ACA is more frequently observed in Europe than in the USA. Chronic joint and bone involvement underlying the skin lesions have been noted in long-standing ACA. Radiographs showed periostal thickening of bones, suggesting periostitis, and subluxations of finger or toe joints in some cases. It might be expected that adequate prevention of tick bites and rapid antibiotic treatment of early recognized cases decreases the incidence of the late features.

P. M. HOUTMAN, T. L. T. A. JANSEN

Medical Centre Leeuwarden, Rheumatology, Leeuwarden, The Netherlands

Correspondence to: P. M. Houtman.
E-mail: n.houtman@znb.nl