PAPER

The Finnish Canine Stifle Index: responsiveness to change and intertester reliability

Heli K Hyytiäinen ,[1] Mikael Morelius,[1] Anu K Lappalainen,[1] Anna F Bostrom,[1] Kirsti A Lind,[2] Jouni J T Junnila ,[3] Anna Hielm-Björkman,[1] Outi Laitinen-Vapaavuori[1]

Abstract

Background The responsiveness and the intertester reliability of the Finnish Canine Stifle Index (FCSI) were tested, and a cut-off between compromised and severely compromised performance level was set.

Methods Three groups of dogs were used, 29 with any stifle dysfunction (STIF), 17 with other musculoskeletal disease except stifle (OTHER) and 11 controls (CTRL). All dogs were tested with the FCSI by the same physiotherapist at three occasions, at baseline, at six weeks and 10 weeks, and once also by another physiotherapist.

Results Dogs in the STIF group demonstrated significantly higher ($P<0.001$) FCSI scores than in OTHER or CTRL groups at baseline. Only the STIF group showed a significant ($P<0.001$) change in FCSI score at all time points, indicating responsiveness to change. There were no significant differences between the evaluators ($P=0.736$), showing good intertester reliability, supported by moderate to good (0.78) intraclass correlation coefficient (ICC). The evaluator performing the FCSI did not have a significant effect when comparing the groups of dogs ($P=0.214$). The 95 per cent confidence intervals of the ICC per group were 0.79 (0.60, 0.91) for STIF, 0.83 (0.53, 0.96) for OTHER 0.78 (0.64, 0.88) for all dogs. A cut-off differentiating a severely compromised from a compromised performance was set at 120, having sensitivity of 83 per cent and specificity of 89 per cent.

Conclusion The FCSI is a recommendable measure of dogs' stifle functionality.

Introduction

The Finnish Canine Stifle Index (FCSI)[1] was generated to provide professionals working with canine stifle patients with a new outcome measure for assessing the level of stifle function, including a functional as well as an objective aspect. The testing battery was composed of several individual items,[2] and it was aimed at quantifying the level of dysfunction in stifle diseased patients. Dysfunction is defined as an abnormality or impairment in the operation of a specified bodily organ or system.[3] Although the individual items comprising the battery have been validated previously, the testing battery still has to be assessed as a whole.[4] Moreover, it is important for the user of a measure or a test to be aware of the measurement properties of that test. This is to ensure appropriate use of test and reliable results, which, unless reliable and correctly interpreted, can lead to distorted knowledge of the patient's situation, and thus have an adverse effect to the patient through misled treatment decisions.

The FCSI has not yet been tested for its responsiveness nor for its reliability. When the testing battery is meant to measure the effect of treatment, it is important to study its ability to detect change over time corresponding to the recovery process, that is, the responsiveness.[5] Responsiveness includes both internal as well as external aspects. The first is based on differences in groups over a prespecified time frame.[6] The latter, in turn, is about the amount of change in a measure in comparison with the change in another measure.[6] This also relates to minimal clinically important difference (MCID),[7] which is an important factor to consider when quantifying a dysfunction in a patient. The MCID is the smallest change that is meaningful to the patient, that is, the smallest change in a treatment outcome that

Veterinary Record (2019) doi:10.1136/ vetrec-2018-105030

[1]Department of Equine and Small Animal Medicine, Faculty of Veterinary Medicine, University of Helsinki, Helsinki, Finland
[2]Veterinary Teaching Hospital, University of Helsinki, Helsinki, Finland
[3]Statistics, 4Pharma, Helsinki, Finland

E-mail for correspondence: Dr Heli K Hyytiäinen; heli.hyytiainen@helsinki.fi

Provenance and peer review Not commissioned; externally peer reviewed.

Received May 22, 2018
Revised October 8, 2019
Accepted October 18, 2019

would be considered important and would indicate a change in the patient's status. A common criterion, by which testing batteries are evaluated, is the intertester reliability. This tests whether several evaluators obtain similar results using the same testing battery on the same patient at the same time.[8] Some evaluators may be stricter in their judgement, and the level of the evaluator's experience may affect the results.[9] The test–retest reliability can be relative, meaning the ratio of total variability between measurements, or absolute, which means the variability of scores between measurements.[10]

The hypothesis of the study was that the FCSI would be responsive to change in stifle diseased patients' level of stifle function. Another hypothesis was that the FCSI would have a good[11] intertester reliability, where the experience level of the evaluator would have no effect on the FCSI result. In addition, a cut-off between compromised and severely compromised performance level, that is, MCID measured with FCSI, would be defined.

Materials and methods

The study was an experimental, longitudinal prospective clinical study, performed on June 1, 2013–April 1, 2014. Dog owners were free to discontinue the study at any time point.

Three groups of dogs were included in the study: dogs with any stifle dysfunction (STIF), dogs with some musculoskeletal disease other than stifle dysfunction (OTHER) and control dogs with no known musculoskeletal disease (CTRL). The group descriptions are presented in table 1. Recruitment of the STIF and OTHER dogs was done by asking all physiotherapy patients of the Veterinary Teaching Hospital of University of Helsinki, meeting the inclusion criteria, to participate in the study. CTRL dogs were recruited by an advertisement on the veterinary students' intranet.

Dogs with dysfunction

All dogs in the STIF and OTHER groups were clients referred by veterinary surgeons to the physiotherapy department of the Veterinary Teaching Hospital of University of Helsinki. The referral letters included the diagnosis, the orthopaedic history, and the clinical and radiographical findings of the dog. A full orthopaedic examination was performed on all dogs at baseline. Inclusion criterion for both groups was a referral from the veterinary surgeon, for the STIF group a diagnosis of a stifle disease and for the OTHER group any other but stifle-related orthopaedic disease. Dogs with neurological deficits were excluded from the study.

Table 1 Description of the dogs participating in the study

Study group	n	Sex (male/ female)	Age (years), mean±sd	Weight (kg), mean±sd
STIF	29	17/12	5.7±2.9	16.0±14.3
OTHER	17	8/9	5.2±3.2	21.5±9.1
CTRL	11	6/5	3.7±2.6	18.5±5.8

Breeds in each group are presented in online supplementary appendix 1.
CTRL, control dogs with no known musculoskeletal disease; OTHER, dogs with some musculoskeletal disease other than stifle dysfunction; STIF, dogs with any stifle dysfunction .

For STIF group dogs, the reasons for referring the patient to physiotherapy, as described by the referring veterinarian in the medical record, were surgical treatment of patellar luxation (n=9), surgical treatment of cranial cruciate ligament rupture (n=14), surgical treatment of a combination of the patellar luxation and cranial cruciate ligament rupture (n=1), surgical treatment of the caudal cruciate ligament (n=1), surgical treatment of both cranial and caudal cruciate ligament and meniscal injury (n=1), conservative treatment for bilateral patellar luxation (n=1), conservative treatment for unilateral patellar luxation (n=1), and stifle osteoarthritis (n=1).

For the OTHER group dogs, the reasons for referring the patient to physiotherapy, as described by the referring veterinarian in the medical record, were femoral head ostectomy (n=2), front limb lameness (n=2), *musculus gluteus medius* injury (n=1), painful back (n=1), avulsion fracture of the tarsal malleolus (n=1), radius and ulna fracture (n=1), spondylosis and a sprained toe in the front limb (n=1), tarsal arthrodesis (n=1), glenohumeral arthroscopy (n=1), bilateral hip dysplasia (n=1), bilateral hip dysplasia and a hemivertebra in thoracic spine (n=1), hip dysplasia and osteoarthritis (n=1), hip osteoarthritis and spondylosis (n=1), hip and glenohumeral osteoarthritis (n=1), and bilateral osteoarthritis of the elbow joint (n=1).

Control dogs

The dogs in the CTRL group consisted of 16 dogs owned by veterinary students and were subjectively healthy. All dogs were evaluated by radiographs and by pressure-sensitive walkway.

Radiological evaluation

Dogs were sedated with dexmedetomidine (0.005 mg/kg) and butorphanol (0.1 mg/kg), and a ventrodorsal hip radiograph with hindlimbs extended was taken according to the radiographic procedure of Fédération Cynologique Internationale (FCI).[12] The images were evaluated and graded (AKL) according to FCI, where grade A is given to a normal, grade B to a nearly normal, grade C to a mildly dysplastic, grade D to a moderately dysplastic and grade E to a severely dysplastic hip joint. Only dogs with grade A or B hip joints were eligible for the study. Stifle joints were radiographed in mediolateral and craniocaudal projections to rule out osteoarthritis and osteochondrosis.

Pressure-sensitive walkway analysis

A pressure-sensitive walkway (GAITRite Electronic Walkway, Peekskill, USA) was used to determine whether the dogs in the CTRL group had any temporospatial asymmetries in their movement. The walkway has an active area of 60.96 x 609.6 cm (90 x 700 cm total area), and an inactive 90 x 125 cm mat was placed at each end of the walkway to minimise any surface change effect on movement. Accompanying software recorded and interpreted the pressure changes in the walkway sensors (GAITRite Manual V.3.9, CIR Systems, Sparta). A scan rate of 240 Hz was used.[13]

The dogs were first acclimatised to the walkway during one to three passes over it. All dogs trotted four to six times over the walkway at a comfortable trotting speed, with no eye contact with the owner, no pull on the leash, and as freely as possible, led by their owners. Runs in both directions were recorded and data of at least 12 full gait cycles from three separate valid runs were collected. The results of the walkway analysis were used to verify that the dog was not lame. Total pressure index, stance time as well as step length were evaluated for obvious asymmetries.

All dogs

For all three groups (STIF, OTHER and CTRL), an orthopaedic examination was done at the time of referral (STIF and OTHER) or at the same time as the initial physiotherapeutic examination (CTRL). Also, all groups were tested with the FCSI at their first physiotherapy appointment (baseline). All dogs were then retested at six weeks and 10 weeks after initial scoring, according to standard orthopaedic veterinary re-evaluation schedule used in the Veterinary Teaching Hospital of University of Helsinki for surgically treated cruciate ligament rupture patients treated with osteotomy techniques or for fracture patients.

Orthopaedic examination

In the subjective clinical lameness examination, the dogs were walked and trotted on a straight line and on a circle in both directions. The surface of the floor was even and non-slippery. The palpatory examination was performed on a standing and a laterally recumbent dog, evaluating muscle symmetry, joint effusions, range of motion and pain. The examinations were scored as follows: lameness on a scale from normal (0) to non-weightbearing (4),[14] and the rest of the evaluations as mild, moderate or severe. To ensure the patient's safety during examination and handling of the patient, the examiners were not blinded to the patient's disease.

Finnish Canine Stifle Index

The FCSI consists of eight tasks, which are the evaluation of the positions of the dog's hindlimbs in sitting and lying positions, the subjective evaluation of the symmetry of the thrust of the hindlimbs in relation to each other as the dog rises from sitting and lying positions, subjective evaluation of the symmetry of the thigh circumference of the dog's hindlimbs in a standing position, measurement of the symmetry of the static weightbearing with bathroom scales, and the measurement of the passive range of motion (flexion and extension) of the dog's stifles with a universal goniometer. In each task the dog's performance is scored with a final result of 0–263. The total score has a cut-off at 60, dividing the scores to adequate and compromised performance level.[1]

Study protocol

The FCSI was used by all three physiotherapists working at the physiotherapy department at the Veterinary Teaching Hospital of University of Helsinki. All of them are specialised in animal physiotherapy, two with over 10 years of experience (HKH, AFB) and one with one-year experience (KAL). One of the therapists (HKH) was very familiar with the FCSI, whereas the others were not. All physiotherapists were taught how to use the testing battery and score the performances in a standardised manner. Both written instructions and a practical introduction session were provided before commencing the study. Each dog was tested using the FCSI at their first physiotherapy appointment (baseline) and after six weeks and 10 weeks from baseline by the same physiotherapist. In addition, at one of the three evaluations, another physiotherapist performed the test as well. The selection of the other physiotherapist was random, based on the physiotherapist's availability. In some cases the test was not performed by another physiotherapist due to either unavailability of another physiotherapist or because physiotherapy was ended before the end of the study period. Therapists were blinded to each others' results, as well as to their own previous results. For patients' safety, they were not blinded to the disease of the patient they were evaluating. The equipment used for the FCSI items (bathroom scales and goniometers) during the trial was the same between all evaluators and times.

Statistical methods

The internal responsiveness was evaluated as follows: the differences between groups (STIF, OTHER, CTRL) in total FCSI score were assessed using a linear mixed effects model for repeated measures, where group, visit and interaction term between group and visit was used as fixed effect and dog as a random effect. Between-group and within-group comparisons were estimated from this model using contrasts.

The relationship and differences between the three evaluators, that is, intertest reliability, were evaluated in three ways. First, to validate the primary group comparisons, a similar linear mixed effects model was fitted as above for the full data, added with the fixed effect of the evaluators (an insignificant tester effect shows that no significant bias is introduced to the group comparisons due to the evaluator evaluating the dog). Secondly, using only the data where two parallel ratings had been made, an analysis of variance model was fitted, where the sole fixed effect was the tester pair for the corresponding dog. Thirdly, a random effects model was fitted to estimate the variance component related to the evaluator. The model included the dog as a random effect and the group as a fixed effect (to avoid overestimation of the variation between dogs). The variance components related to dogs and evaluators were estimated from the model and the proportions of total variation were calculated for the components, that is, intraclass correlation coefficient (ICC). The random effect modelling was repeated separately in the STIF and OTHER groups to investigate the variance components within group. These models included only the dog as random effect with no fixed effects. For the CTRL group, this within-group evaluation was not possible due to low variation between the dogs.

The diagnostic ability of the FCSI total score in differentiating severely compromised and compromised dogs (namely, STIF v OTHER) was investigated using receiver operating characteristic (ROC) curve. The optimal cut-off value for the FCSI score was defined as the point where the sum of sensitivity and specificity of the score reached its maximum value. In addition, the previously set cut-off level between adequate and compromised performance level in the FCSI was retested in the present study population.[1]

A P value of less than 0.05 was considered statistically significant, and 95 per cent confidence intervals (CI) were calculated for the estimates of group differences in FCSI score and for the estimated ICCs. All statistical analyses were done using the same statistical program (SAS System for Windows, V.9.3, SAS Institute, Cary, North Carolina).

Results

Initially, 16 dogs were enrolled into the CTRL group, but five of them were excluded due to findings in their radiographic evaluation, which were hip osteoarthritis with hips graded as C/C (n=3), hips graded C/C with no osteoarthritis (n=1) and asymmetrical lumbosacral transitional vertebra (n=1). In addition one of the excluded five dogs had also mild findings in the orthopaedic examination and two in the pressure-sensitive walkway evaluation. The remaining 11 dogs were finally included into the study as the CTRL group. The changes in the group sizes between three testing times and the reasons leading to the changes are presented in table 2.

Of the dogs in the STIF group, 26 had had a surgical treatment of one of their stifles. The time from surgical treatment to the FCSI baseline measurement was a median of 17.5 days (minimum 10, maximum 78 days). Two of the STIF dogs were treated conservatively, and the time from diagnosis and start of treatment to the FCSI baseline measurement was a median of 48 days (minimum 1, maximum 95 days). In the OTHER group, the diseases were treated surgically in four cases, with a median of 17.5 days from diagnosis to the first FCSI measurement (minimum 9, maximum 44 days). Thirteen of the OTHER dogs were treated conservatively. However, three dogs had no information on the actual date of diagnosis, so only 10 of the dogs' timeline from diagnosis and start of medical treatment to the start of physiotherapy and the first FCSI measurement could be counted. In these dogs, the median was 13 (minimum 5, maximum 55) days.

Table 2 Description of the groups during the study

	STIF	OTHER	CTRL
Number of dogs at baseline	29	17	11
Number of dogs at six weeks	25	11	10
Number of dogs at 10 weeks	19	11	11
Physiotherapy ended before the end of the study period due to relapse of the disease or complications	1	1	NA
Physiotherapy ended before the end of the study period due to owner decision	6	4	NA
Acute trauma front limb lameness	NA	NA	1
Intertester reliability not tested due to logistical reasons (ie, unavailability of another physiotherapist or physiotherapy aborted before the end of the study period)	7	8	1

CTRL, control dogs with no known musculoskeletal disease; NA, not applicable; OTHER, dogs with some musculoskeletal disease other than stifle dysfunction; STIF, dogs with any stifle dysfunction.

The mean FCSI score at baseline was 154.7±60.9 in the STIF group, 59.4±54.3 in the OTHER group and 17.0±22.9 in the CTRL group, respectively. The difference between all groups was significant (P<0.001). All of the mean scores can be seen in table 3.

Dogs were tested at a mean of 5.7±1.9, and 10.3±1.4 weeks from the baseline. The largest change in mean±sd total score between baseline and at six weeks and 10 weeks was in the STIF group: 48.8±44.6 and 93.3±62, respectively. Only the STIF group showed a significant (P<0.001) change at both six weeks and 10 weeks (figures 1 and 2).

When evaluating the internal responsiveness, a significant difference between the STIF and the other two groups was seen both at baseline and at six weeks. At 10 weeks the difference was significant only between the STIF and the CTRL groups (P=0.002) (table 4). Differences in the FCSI total score between groups were consistently highest when STIF and CTRL groups were compared (table 4).

Based on the baseline results, a cut-off point to differentiate a severely compromised from a compromised performance level was set to 120, which had a sensitivity of 83 per cent and specificity of 89 per cent (figure 3). The previously set cut-off value of 60 between a compromised and an adequate performance level[1] resulted in a sensitivity of 72 per cent and a specificity of 91 per cent in this study population (figure 3). A receiver operator curve also illustrates the above-mentioned specificity and sensitivity values (for differentiating severely compromised (STIF) from compromised (OTHER) dogs) as well as the AUC: 0.905 (95 per cent CI 0.829, 0.982) (figure 4).

Table 3 Means of the FCSI scores at all measurement points

Measurement/group/ dogs (n)	Mean of FCSI score (±sd)	Change from baseline score (±sd)
1. STIF/29	154.7 (±60.1)	NA
2. STIF/25	108.7 (56.9)	−48.8 (±44.6)
3. STIF/19	58.6 (±44.9)	−93.3 (±62.0)
1. OTHER/17	59.4 (±54.3)	NA
2. OTHER/11	43.2 (±52.8)	−26.1 (±38.1)
3. OTHER/11	39.8 (±37.0)	−29.5 (±39.6)
1. CTRL/11	17.0 (±22.9)	NA
2. CTRL/10	15.4 (±13.9)	−3.34 (±13.6)
3. CTRL/11	5.3 (±11.9)	−11.7 (±21.0)

CTRL, control dogs with no known musculoskeletal disease; FCSI, Finnish Canine Stifle Index; NA, not applicable; OTHER, dogs with some musculoskeletal disease other than stifle dysfunction; STIF, dogs with any stifle dysfunction.

No significant differences were observed between the different evaluators (P=0.736). The evaluator performing the FCSI did not have a significant effect when comparing the groups (P=0.214). The random effects model showed that the proportion of total variance was 78.4 per cent due to variation between dogs (within each problem group) and 21.6 per cent due to variation between the evaluators, calculated as an ICC of 0.78. The 95 per cent CIs of the ICC per group were 0.79 (0.60, 0.91) for STIF, 0.83 (0.53, 0.96) for OTHER and 0.78 (0.64, 0.88) for all dogs.

Discussion

Based on the results of this study, the FCSI was seen to be responsive to changes in the dogs' level of dysfunction in the STIF group. The change over time in the FCSI score was significant (P<0.001) and largest in the STIF group. It is noteworthy that all dogs in STIF and OTHER groups received physiotherapy, and although the effect of therapy was not studied here a change seen in the STIF group (93.3 (±62)) was clearly more evident than the one in the OTHER group (29.5 (±39.6)). This indicates that the FCSI is sensitive to stifle dysfunction over other dysfunctions, and that there are more stifle-related than other joint-related items in the FCSI. Nevertheless, the possible effect of others, such as tarsus or hip-related disease, should be kept in mind, as they might affect the FCSI. However, as the FCSI is not used to diagnose a disease, this should not be a problem. The evaluation of internal responsiveness is based on differences between groups over a specified time frame. This leads to assumption of treatment effect over time in studies like this, as is the case in the present study too. No other research-related outcome measures were used to verify the change of stifle functionality over time than the FCSI, but progress was assumed to happen. The possible other measures used with the patients as part of their treatment were not recorded nor compared with the results of the present study. Previous studies have shown the positive effect of physiotherapy on postoperative rehabilitation on surgically treated cranial cruciate ligament patients' outcome.[15–18] Towards the final testing, the results of STIF dogs started to resemble the results of the OTHER

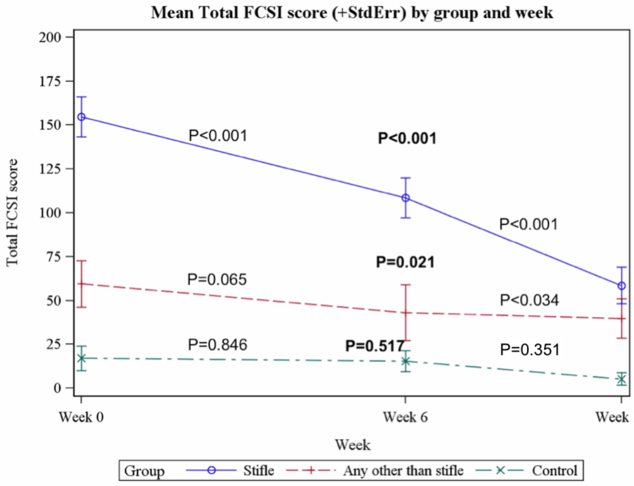

Figure 1 Descriptive statistics for the mean total FCSI score for the three groups at baseline and after six weeks and 10 weeks. Significance of change within group between two testing times is marked above each corresponding line. The significance of change from baseline to 10 weeks is marked with bold above the line of the group. FCSI, Finnish Canine Stifle Index.

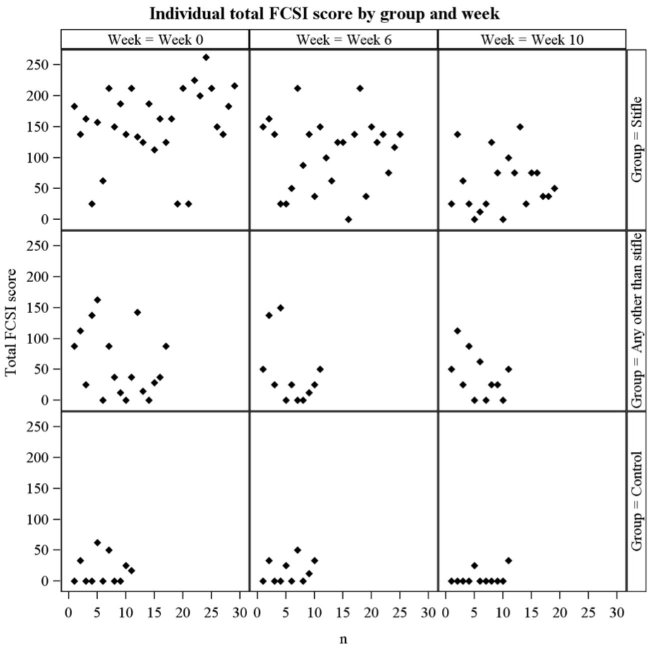

Figure 2 Total individual FCSI scores per group at three different testing times. FCSI, Finnish Canine Stifle Index.

Table 4 Differences between groups in FCSI score by testing times

		Estimate of difference in FCSI score	se	95% CI		P value
				Upper	Lower	
Difference between groups at baseline	STIF v OTHER	95.2	15.2	64.9	125.5	<0.001*
	STIF v CTRL	137.6	17.6	102.5	172.8	<0.001*
	OTHER v CTRL	42.4	19.3	4.0	80.8	0.031*
Difference between groups at six weeks from baseline	STIF v OTHER	70.4	16.9	36.8	103.9	<0.001*
	STIF v CTRL	92.6	18.2	56.5	128.7	<0.001*
	OTHER v CTRL	22.2	20.7	−18.9	63.3	0.287
Difference between groups at 10 weeks from baseline	STIF v OTHER	29.9	17.4	−4.5	64.3	0.088
	STIF v CTRL	57.9	18.3	21.6	94.3	0.002*
	OTHER v CTRL	28.0	20.4	−12.6	68.7	0.174

*Denotes significance.
CI, confidence interval; CTRL, control dogs with no known musculoskeletal disease; FCSI, Finnish Canine Stifle Index; OTHER, dogs with some musculoskeletal disease other than stifle dysfunction; STIF, dogs with any stifle dysfunction.

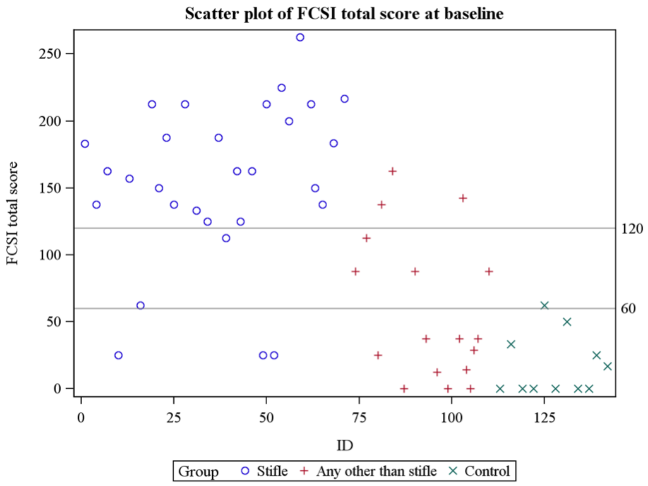

Figure 3 Scatter plot of the mean FCSI score at baseline. The figure presents the two cut-off lines for the three performance levels at the FCSI total score: adequate below 60, compromised between 60 and above 120, and severely compromised above 120. FCSI, Finnish Canine Stifle Index.

dogs and were almost even with the results of the CTRL dogs. The results may have been even clearer had the testing times been even wider apart, or if there had been a fourth measurement time. This, however, would not have been realistic due to owner compliance.

A thing to consider is the ceiling effect, which means that the maximum result of the test is often reached. This, based on the results of the present study, does not seem to be a problem with FCSI. Of all tested dogs, only one was near maximal score, despite several severely dysfunctional patients being included. This tells that the upper scale of the test is sufficient to be used with this type of a patient group. Floor effect, in turn, means that most of the subjects would score the minimum result. This would not seem to be a problem either, with FCSI, when testing stifle patients. That being said, when the patients reach 'near normal' functionality, there may be some level of floor effect, and the responsiveness to change may decrease and the amount of change gets less. This can already be seen in the results of the present study, in the CTRL group and to an extent with the OTHER group. However, clinically, this would no longer be a problem to a rehabilitating stifle patient, as at that stage the level of functionality would be acceptable.

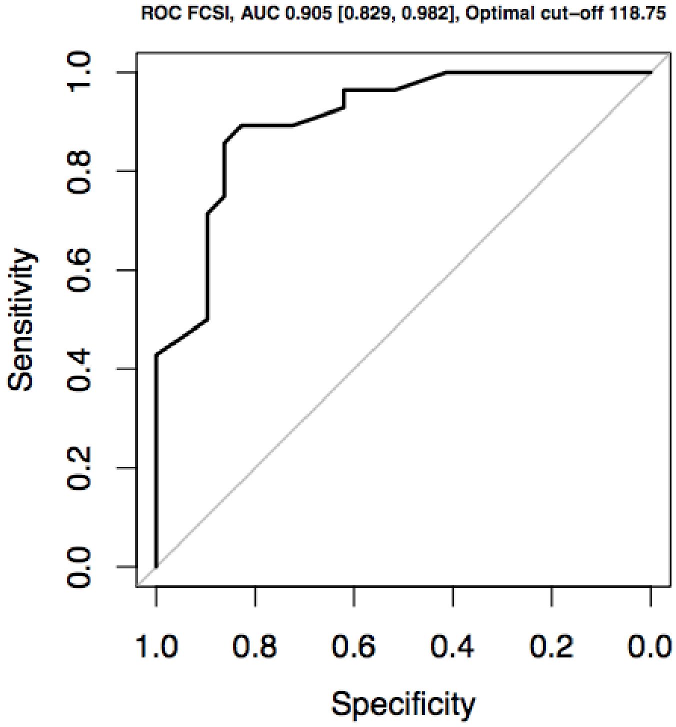

Figure 4 ROC curve representing the sensitivity and specificity of FCSI. FCSI, Finnish Canine Stifle Index; ROC, receiver operating characteristic.

A cut-off between 'adequate' and 'compromised' performance according to the FCSI total score had been set in a previous publication,[1] and it was confirmed in this study, separating the CTRL dogs from the others, with moderate sensitivity (72 per cent) and high specificity (91 per cent). In addition, the cut-off between 'compromised' and 'severely compromised' was set with high sensitivity (83 per cent) and specificity (89 per cent). The total score of the testing battery now has a descriptive aspect to it, as the result is not merely numerical, but also describes the clinical state of the patient qualitatively. Similar cut-offs and definitions have been used in human knee testing batteries.[19–21] The importance of these values does not lie only in their statistical significance, but primarily on the clinical significance. The concept of MCID is of great value to the patient itself, and thus at the core of functionality. Although this concept is often applied in human patients' quality of life questionnaires, in animals, where assessment of clinical signs is left to human interpretation, FCSI could represent an equivalent through defining levels of dysfunction. One method of establishing the MCID is through equal sensitivity and specificity, and the ROC curves,[22] like the one presented in the present study. The cut-off line presented is the point of MCID. Based on the surface below the ROC curve, one can identify the probability of correct discrimination between the improved and not improved stifle patients. In the present study, the AUC was within the reference values of 0.8 and 0.9, meaning excellent discrimination ability[22] for FCSI.

The intertester reliability of the FCSI (0.78) was found to be good,[11] and as ICC over 0.70 is considered to represent adequate degree of reliability in a clinically used test[23] this demand is clearly met. Further, no differences were seen between the experienced and unexperienced evaluators, nor was there a difference in the results between the evaluator who was previously familiar with the testing battery and the ones who were not. Although some lower limb-related human testing batteries have been studied for their intertester reliability,[24–26] many have not.[26] The CIs of the ICC were rather wide in the present sample of dogs, for example, the lower CI for the STIF group being 0.6, implying only moderate reliability. Most probably the wide intervals were mainly due to the low number of dogs used in this study. Had there been more dogs enrolled, the intervals might have been narrower. Nevertheless, based on the present study, physiotherapists specialised in animal physiotherapy would be able to use the FCSI testing battery after familiarisation with the protocol.[27]

The FCSI has been developed based on studies on dogs weighing over 17.5 kg,[1 2] and it was unclear whether or not the test would work on smaller sized dogs. The population in this study therefore was chosen to be heterogeneous. Although size was not considered to be a factor *per se*, the results of the FCSI's reliability and responsiveness are explicit, even with dogs weighing between 2.7 kg and 58.4 kg.

Having both surgically and non-surgically treated dogs in the STIF group may have influenced the results of the present study. The non-surgically treated dogs may have been slower to improve, thus possibly inhibiting the improvement seen in the STIF group's FCSI score over time. Further, the fact that there were more surgically treated dogs in the STIF group (26) in comparison with the OTHER group (6) may also have affected the results. However, as this was a clinical study, and all available patients during the study period were included, we had no control over the proportions of surgically or conservatively treated patients. The time interval from surgical treatment or time of diagnosis and start of medical treatment in conservatively managed cases was equal in both groups. However, the nature of disease as well as the treatment (conservative v surgical) were different. In the OTHER group the diseases can be considered to have been generally more chronic in nature, in comparison with the STIF groups' postsurgical stage, and this may have increased the difference between the groups at the baseline. Presumably surgically treated patients will show a steeper healing curve than the conservatively treated ones. Thus, in addition to making the conclusion that the test is more sensitive to change in 'stifle' than 'other' diseases, one could also argue that it is actually more sensitive to change in postsurgical patients than patients with more stable orthopaedic disease. However, despite the nature of diseases being different, it could be assumed that even the chronic orthopaedic diseases are likely to have been painful and in an acute phase at the time when the owner sought veterinarian help and when a diagnosis was made and treatment started.

The dogs with bilateral problems in their hindlimbs may also have affected the results of the study. It should be emphasised that some of the FCSI testing battery's items (thrust up from sitting and lying, and thigh circumference symmetry) are comparative, giving a score only to the weaker of the hindlimbs, and therefore always scores at least one of the limbs as 'adequate'. Other items (hindlimb position in sitting and lying, static weightbearing, range of motion) score both limbs, independently of each other, meaning that both limbs can get a score. This may be confusing when there is a bilateral problem. Although one can, to an extent, score both limbs, only the total score of the limb that is worse at that time will be reliable. This is because the better hindlimb may provide misleadingly good results due to the comparative items. Therefore one should always be aware that the FCSI is a test for one hindlimb, comparative in nature, and works most accurately on dogs whose other hindlimb is healthy or at least clearly better than the diseased one.

Another factor to consider are the dogs in the OTHER group with dysfunction in their hip or tarsus. They also may have affected the results to some extent, as the tarsal and hip joints are connected to the stifle joint both anatomically and biomechanically, and dysfunction in either of these would, potentially, affect the stifle—as would therapy of these joints. Nevertheless, a significant difference in total FCSI score was seen between the groups, although half of the dogs in the OTHER group did have a dysfunction in either their hip or tarsal joints.

Relying on the referring veterinarians' diagnosis and to accurately exclude any concurrent pathologies, for example any stifle disease in the OTHER group, does introduce a random factor to the present study. In addition, the STIFLE and OTHER groups' treatment response was not confirmed by any gold standard measurement. It was expected that they would improve over time and due to rehabilitation. The authors do recognise these factors as weaknesses of the study. However, at baseline the FCSI scores in the STIF group were significantly higher than in the other two groups, suggesting that FCSI is able to differentiate the dogs with stifle dysfunction from other dogs. Moreover, dogs in both study groups were referred to physiotherapy, and therefore progress of rehabilitation was at all times controlled by the veterinarian during routine veterinary controls, such as for the tibial plateau levelling osteotomy patients at eight weeks. In case of unprogressive rehabilitation, the therapist would have reacted by contacting the referring veterinarian. This was a clinical study, and the situation corresponds to the one with which physiotherapists work daily.

The FCSI is a responsive measurement method with moderate to good intertester reliability in all dogs and moderate to excellent intertester reliability in dogs with stifle disease. A cut-off point for MCID has been defined. The FCSI can be recommended as an outcome measure and an assessment method when evaluating the level of stifle functionality in stifle diseased dogs.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Ethics approval The study protocol was approved by the University of Helsinki Viikki Campus Research Ethics Committee, and written consent was obtained from all owners.

Data availability statement All data relevant to the study are included in the article.

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/vetrec-2018-105030).

ORCID iDs

Heli K Hyytiäinen http://orcid.org/0000-0002-7903-1672
Jouni J T Junnila http://orcid.org/0000-0003-2703-0798

References

1 Hyytiäinen HK, Mölsä SH, Junnila JJT, *et al*. Developing a testing battery for measuring dogs' stifle functionality: the Finnish Canine Stifle Index (FCSI). *Vet Rec* 2018;183.

2 Hyytiäinen HK, Mölsä SH, Junnila JT, *et al*. Ranking of physiotherapeutic evaluation methods as outcome measures of stifle functionality in dogs. *Acta Vet Scand* 2013;55:29–37.

3 . Available: https://en.oxforddictionaries.com/ [Accessed 12.12.2018].

4 Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006;119:166.e7–166.e16.

5 Stratford PW, Binkley JM, Riddle DL. Health status measures: strategies and analytic methods for assessing change scores. *Phys Ther* 1996;76:1109–23.

6 Johnston BC, Ebrahim S, Carrasco-Labra A, *et al*. Minimally important difference estimates and methods: a protocol. *BMJ Open* 2015;5:e007953.

7 Husted JA, Cook RJ, Farewell VT, *et al*. Methods for assessing responsiveness: a critical review and recommendations. *J Clin Epidemiol* 2000;53:459–68.

8 Bartko JJ, Carpenter WT. On the methods and theory of reliability. *J Nerv Ment Dis* 1976;163:307–17.

9 Blonna D, Zarkadas PC, Fitzsimmons JS, *et al*. Accuracy and inter-observer reliability of visual estimation compared to clinical goniometry of the elbow. *Knee Surg Sports Traumatol Arthrosc* 2012;20:1378–85.

10 Serbetar I. Establishing some measures of absolute and relative reliability os a motor tests. *Croat J of Educ* 2016;17:37–48.

11 Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15:155–63.

12 Flückiger M. Scoring radiographs for canine hip dysplasia – the big three organisations in the world. *EJCAP* 2007;17:135–40.

13 Light VA, Steiss JE, Montgomery RD, *et al*. Temporal-Spatial gait analysis by use of a portable walkway system in healthy Labrador Retrievers at a walk. *Am J Vet Res* 2010;71:997–1002.

14 Mostafa AA, Griffon DJ, Thomas MW, *et al*. Morphometric characteristics of the pelvic limbs of Labrador Retrievers with and without cranial cruciate ligament deficiency. *Am J Vet Res* 2009;70:498–507.

15 Marsolais GS, Dvorak G, Conzemius MG. Effects of postoperative rehabilitation on limb function after cranial cruciate ligament repair in dogs. *J Am Vet Med Assoc* 2002;220:1325–30.

16 Monk ML, Preston CA, McGowan CM. Effects of early intensive postoperative physiotherapy on limb function after tibial plateau leveling osteotomy in dogs with deficiency of the cranial cruciate ligament. *Am J Vet Res* 2006;67:529–36.

17 Baltzer WI, Smith-Ostrin S, Warnock JJ, *et al*. Evaluation of the clinical effects of diet and physical rehabilitation in dogs following tibial plateau leveling osteotomy. *J Am Vet Med Assoc* 2018;252:686–700.

18 Romano LS, Cook JL. Safety and functional outcomes associated with short-term rehabilitation therapy in the post-operative management of tibial plateau leveling osteotomy. *Can Vet J* 2015;56:942–6.

19 Lequesne MG, Mery C, Samson M, *et al*. Indexes of severity for osteoarthritis of the hip and knee. Validation--value in comparison with other assessment tests. *Scand J Rheumatol Suppl* 1987;65:85–9.

20 Lequesne M. Indices of severity and disease activity for osteoarthritis. *Semin Arthritis Rheum* 1991;20:48–54.

21 Lequesne MG. The algofunctional indices for hip and knee osteoarthritis. *J Rheumatol* 1997;24:779–81.

22 Copay AG, Subach BR, Glassman SD, *et al*. Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J* 2007;7:541–6.

23 Tickle-Degnen L. Communicating evidence to clients, managers, and funders. In: Law M, MacDermid J, eds. Evidence based rehabilitation. A guide to practice. 2nd edn. New Jersey, USA: SLACK Incorporated, 2008.

24 Frohm A, Heijne A, Kowalski J, *et al*. A nine-test screening battery for athletes: a reliability study. *Scand J Med Sci Sports* 2012;22:306–15.

25 Mikkelsen LR, Mikkelsen S, Søballe K, *et al*. A study of the inter-rater reliability of a test battery for use in patients after total hip replacement. *Clin Rehabil* 2015;29:165–74.

26 Haitz K, Shultz R, Hodgins M, *et al*. Test-retest and interrater reliability of the functional lower extremity evaluation. *J Orthop Sports Phys Ther* 2014;44:947–54.

27 Finch E, Brooks D, Stratford PW, *et al*. Physical rehabilitation outcome measures. A guide to enhanced clinical decision making. 2nd edn. Ontario, Canada: BC Decker, 2002.