Comparison of Reliability of Norberg Angle and Distraction Index as Measurements for Hip Laxity in Dogs

Julius Klever[1] Andreas Brühschwein[1] Silvia Wagner[1] Sven Reese[2] Andrea Meyer-Lindenberg[1]

[1] Clinic of Small Animal Surgery and Reproduction, Centre of Veterinary Clinical Medicine, LMU Munich, Munich, Germany
[2] Institute of Veterinary Anatomy, Histology and Embryology, LMU Munich, Munich, Germany

Address for correspondence Julius Klever, Dr. med. vet., Clinic of Small Animal Surgery and Reproduction, Centre of Veterinary Clinical Medicine, LMU Munich, Veterinärstrasse 13, D-80539 Munich, Germany (e-mail: klever@chir.vetmed.uni-muenchen.de).

Vet Comp Orthop Traumatol 2020;33:274–278.

Abstract

Objective The main purpose of the study was to compare reliability of measurements for the evaluation of hip joint laxity in 59 dogs.

Materials and Methods Measurement of the distraction index (DI) of the PennHIP method and the Norberg angle (NA) of the Fédération Cynologique Internationale (FCI) scoring scheme as well as scoring according to the FCI scheme and the Swiss scoring scheme were performed by three observers at different level of experience. For each dog, two radiographs were acquired with each method by the same operator to evaluate intraoperator-reliability.

Results Intraoperator-reliability was slightly better for the NA compared with the DI with an intraclass correlation coefficient (ICC) of 0.962 and 0.892 respectively. The ICC showed excellent results in intraobserver-reliability and interobserver-reliability for both the NA (ICC 0.975; 0.969) and the DI (ICC 0.986; 0.972). Thus, the NA as well as the DI can be considered as reliable measurements. The FCI scheme and the Swiss scoring scheme provide similar reliability. While the FCI scheme seems to be slightly more reliable in experienced observers (Kappa FCI 0.687; Kappa Swiss 0.681), the Swiss scoring scheme had a noticeable better reliability for the unexperienced observer (Kappa FCI 0.465; Kappa Swiss 0.514).

Clinical Significance The Swiss scoring scheme provides a structured guideline for the interpretation of hip radiographs and can thus be recommended to unexperienced observers.

Keywords
- canine hip dysplasia
- distraction index
- Norberg angle
- radiography

Introduction

Canine hip dysplasia is a common orthopaedic disease in dogs.[1] The prevalence varies in different breeds between 2 and 80%.[2] Canine hip dysplasia is a polygenetic and multifactorial condition[3–6] and heritabilities of 0.14 to 0.43 are reported.[7,8] Phenotypic breeding stock selection is aimed to reduce the incidence based on the genetic component. Increased hip joint laxity is one of the most important factors in the assessment of canine hip dysplasia. There are numerous radiographic methods for the detection of canine hip dysplasia in the world.[9] The most widely used method in Europe is the five grade (A-E) Fédération Cynologique Internationale (FCI) scheme,[10] which is based on evaluation of various radiographic findings, including signs for osteoarthritis, and the Norberg angle (NA) as an objective indicator for hip laxity. A line between both femoral head centres and the corresponding craniolateral acetabular margins on each side form the NA.[11] In contrast to the FCI method, the PennHIP method relies on the identification of

received
July 9, 2019
accepted
February 16, 2020
published online
April 29, 2020

DOI https://doi.org/10.1055/s-0040-1709460.
ISSN 0932-0814.

osteoarthritis and, for those without signs for osteoarthritis, assessment of the passive hip joint laxity expressed by the distraction index (DI).[12] Laxity is measured on radiographs with a distraction device causing the femoral head to displace laterally. The DI is calculated using the distance between the acetabular and the femoral head centre divided by the radius of the femoral head.

The FCI grading system has relatively poor interobserver agreement[13,14] although the reproducibility of the NA seems to be sufficient.[15] For the PennHIP method, a study was published and showed high within- and between-examiner repeatability.[16] One study showed a high repeatability of DI measurements when comparing the official results to results of trained researchers.[17] A recent study revealed substantial variability for the NA but not for the DI.[18]

Measurements should be both reliable and valid to evaluate the radiographic phenotype. Accuracy, also referred to as validity, demonstrates how close a measurement is to the true value based on the gold standard. Reliability, also referred to as precision or consistency, determines how close the measurements are to each other and is therefore negatively correlated to variability. Reliability can be evaluated by repeated measurements.[19]

To evaluate the reliability of radiographic measurements, different factors have to be taken into account. An error may derive from differences in the radiograph due to positioning, projection or different forces applied during acquisition. This effect can be assessed by acquiring two identical sets of radiographs and is also referred to as repeatability, also termed intraoperator reliability or -agreement, if the radiographs is taken by the same person or reproducibility (also termed interoperator reliability or agreement) if the radiographs are taken by different persons. Furthermore, an error can be derived from the measurement itself. This can be evaluated measuring twice using the same radiograph and is also termed repeatability (intraobserver or intrarater reliability or agreement) or reproducibility (interobserver or interrater reliability or agreement) depending if the measurements are made by the same or different persons.[19]

In the available literature, to date there is no study that directly compares the reliability between measurements of NA and DI in a structured and comparable form that takes repeatability and reproducibility into consideration. The aim of the study was to evaluate intraoperator-reliability as well as intra- and interobserver reliability of the NA and DI measurements.

Materials and Methods

A total of 59 dogs that were presented for official hip screening were included after the owner's consent was given. The dogs had to fit the minimum weight requirement of 8 kg for evaluation with the PennHIP distractor. To comply with the FCI criteria for official screening, the minimum age was 12 months. All animals underwent injection anaesthesia using dexmedetomidine (0.01–0.02 mg/kg dexdomitor 0.5 mg/mL; Orion Pharma GmbH, Hamburg, Germany), medetomidine (0.01–0.04 mg/kg Dorbene Vet 1 mg/mL, Zoetis Deutschland GmbH, Berlin, Germany) or diazepam (0.1–0.5 mg/kg Ziapam 5 mg/mL, Ecuphar GmbH, Greifswald, Germany) intravenously followed by the administration of propofol (1–8 mg/kg Narcofol 10 mg/mL, CP-Pharma GmbH, Burgdorf, Germany) until the dogs were fully anaesthetized with adequate muscle relaxation.[20]

For each dog five radiographs were taken in the same order on a direct digital radiography system (Siemens Axiom Luminos dRF; Siemens Healthcare AG, Erlangen, Germany) without the use of positioning devices. All radiographs were obtained by the same PennHIP-certified veterinarian. A standard ventrodorsal projection of the pelvis with extended hips also known as the FCI position 1 and the ventrodorsal projection of the pelvis with limbs in neutral position with distraction of the femoral joint using a PennHIP distractor (PennHIP distraction view) were repeated, while the PennHIP compression view was performed once. Images were anonymized by a person not involved in scoring of the radiographs and evaluations were performed at the earliest 1 month after acquisition of the images. Before the study was conducted, every observer trained measuring the DI and the NA in 10 cases with known official results. The FCI and Swiss scheme scoring of the hips as well as measurements of the NA and the DI was performed twice, after a 2-month interval, by a first year imaging resident with 5 years of experience in diagnostic imaging, once by an European specialist in veterinary diagnostic imaging and member of the German association of scrutineers and one intern without experience in veterinary diagnostic imaging. The measurements were made in the same digital environment in the same order by all observers, using specific tools for measurement of the NA and DI provided by the commercial software (Dicom PACS, Oehm & Rehbein GmbH, Rostock, Germany) used in the institution. The 'distraction index tool' consists of two circles that can be manually adjusted to fit the femoral head and the acetabulum and automatically calculates the DI value. The 'Norberg angle tool' consists of two circles that need to be drawn over each femoral head and a line that needs to be adjusted to the cranial acetabular edge on each side. The NA for each side is displayed subsequently.

Results were stored for each hip joint separately in an excel spreadsheet (Office 2010 Excel; Microsoft, Redmond, Washington, United States). Statistical analysis was conducted using commercial statistical software (MedCalc; MedCalc Software, Ostend, Belgium). Intraclass correlation coefficient (ICC) was calculated to evaluate reliability of intraoperator, intraobserver as well as interobserver measurements. This test allows comparison between samples of different scales, such as the NA (degree) and the DI (unitless) values.[10,12] An ICC of 1 indicates perfect agreement, whereas an ICC of 0 indicated not more than random agreement. Intraclass correlation coefficient values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9 and greater than 0.90 can be interpreted as poor, moderate, good and excellent reliability, respectively.[21] Cohens weighted kappa was calculated to compare the observer agreement between the categorical FCI classification and classification made using the Swiss scoring scheme.[22] A kappa of 1 indicates perfect agreement, whereas a value of 0

Table 1 Comparison of intraclass correlation coefficient for the reliability of Norberg angle and distraction index

	Norberg angle	Distraction index
Intraoperator	0.962	0.892
Intraobserver	0.975	0.986
Interobserver	0.969	0.972

indicates not more than random agreement and negative values represent a negative correlation. Values of 0.21 to 0.40, 0.41 to 0.60, 0.61 to 0.80 and greater than 0.81 can be interpreted as fair, moderate, substantial and as almost perfect agreement, respectively.[23]

Results

The 59 dogs included 20 different breeds (10 German Shepherd Dogs, 7 Labrador Retriever, 6 Golden Retriever, 4 Doberman Pinscher, 4 Flat Coated Retriever, 3 Small Münsterländer, 3 Belgian Shepherd Dogs, 3 Entlebucher Mountain Dogs, 2 Akita, 2 Australian Shepherd Dogs, 2 Border Collies, 2 Nova Scotia Duck Tolling Retriever, 2 Schnauzer, 2 Vizsla, 2 White Shepherd Dogs, 1 Pyrenean Shepherd Dog, 1 Bernese Mountain Dog, 1 German Wirehaired Pointer, 1 Eurasian Dog, 1 Keeshond). Of all dogs 32.2% ($n = 19$) were scored FCI grade 'A' (no evidence of hip dysplasia), 42.4% ($n = 25$) were scored FCI grade 'B' (borderline), 18.6% ($n = 11$) were scored FCI grade 'C' (mild hip dysplasia) and 6.8% ($n = 4$) FCI grade 'D' (moderate hip dysplasia.

Results of the statistical analysis for intraoperator reliability, intraobserver reliability and interobserver reliability are provided in ►**Table 1**.

Intraoperator Reliability

Intraclass correlation coefficient for the NA was 0.962 with a 95% confidence interval from 0.941 to 0.975 and for the DI 0.892 with a 95% confidence interval from 0.833 to 0.931.

Intraobserver Reliability

Intraclass correlation coefficient for the NA was 0.975 with a 95% confidence interval from 0.964 to 0.983 and for the DI 0.986 with a 95% confidence interval from 0.979 to 0.990.

The weighted kappa for the agreement between both measurements for the classification according to the FCI scheme was 0.699 with a 95% confidence interval from 0.609 to 0.789 and for the classification according to the Swiss scheme 0.661 with a 95% confidence interval from 0.556 to 0.767.

Interobserver Reliability

Intraclass correlation coefficient between all three observers for the NA was 0.969 with a 95% confidence interval from 0.957 to 0.978 and for the DI 0.972 with a 95% confidence interval from 0.950 to 0.983.

Intraclass correlation coefficient between both experienced observers (AB and JK) for the NA was 0.983 with a 95% confidence interval from 0.969 to 0.990 and for the DI 0.980 with a 95% confidence interval from 0.972 to 0.986.

Intraclass correlation coefficient between one experienced and one unexperienced observer (AB, SW) for the NA was 0.936 with a 95% confidence interval from 0.895 to 0.959 and for the DI 0.947 with a 95% confidence interval from 0.865 to 0.973.

The weighted Kappa for the agreement between both experienced observers (AB and JK) for the classification according to the FCI scheme was 0.687 with a 95% confidence interval from 0.596 to 0.778 and for the classification according to the Swiss scheme 0.681 with a 95% confidence interval from 0.588 to 0.774. The weighted Kappa for the agreement between one experienced and one unexperienced observer (AB and SW) for the classification according to the FCI scheme was 0.465 with a 95% confidence interval from 0.344 to 0.585 and for the classification according to the Swiss scheme 0.514 with a 95% confidence interval from 0.392 to 0.635.

Discussion

Repeated radiographs and measurements were performed to evaluate reliability of DI and NA.[19] The intraoperator reliability of the DI was slightly lower (ICC 0.892), but still a good, almost excellent result. The NA seems to generate slightly more precise results in between two repeated radiographs. Although our operators are PennHIP-certified, they are much more trained in the more frequently used standard ventrodorsal radiograph compared with the distraction radiographs. This experience may influence the repeatability. Subjectively distraction radiographs are more difficult because besides patient positioning, additional attention has to be paid to the handling of the distraction device. The slight differences in repeated radiographs may derive from a combination of various factors such as the forces applied, pelvic tilting, muscle relaxation, central beam position or other unknown random effects.[20,24,25]

The ICC showed minimally better results for the intra- and interobserver-reliability of the DI compared with the NA. This complies with a recent study where variability of the NA was higher than of the DI.[26] In contrast to the other study, we found no substantial difference and excellent reliability ($ICC > 0.90$) for both methods and the differences seem negligible. Based on the small sample size of only 10 dogs in the other study, their higher variability for the NA might be caused by outliers. Another main influence on intra- and interobserver-reliability is probably caused by the precise definition of measurement points with special focus on common anatomic variations. The availability or the lack of a detailed and in-depth description of measurement points and procedures, also with special regard to anatomical variants, may contribute to the variations in the results of various studies of inter-observer agreement. Norberg angle and DI are based on the measurement of perfect circles. Based on our agreement for the NA, the femoral head circle was defined by two points on the cranial and craniolateral projected surface and one point on the centre of the caudomedial projected surface of the femoral head on the

radiograph, neglecting and bridging the depression or flattening of the acetabular fossa and the junction to the femoral neck. Neither the femoral head nor the craniolateral acetabular rim of the *facies semilunata* were always projected as perfect circle segments on radiographs, this can be due to distortion caused by divergence of the X-ray or just normal anatomical variation.[27] But we were able to fit freely adjustable circles to these structures by approximation. In our experience, it was frequently hard to precisely define the measurement point of the caudolateral acetabular edge for the DI as well as the craniolateral acetabular edge for the NA. This can be explained variability in the visibility of the measurement points in different radiographs, probably mainly due to anatomic variation and positioning. Another feature that might influence the precision of the measurements is the severity of osteoarthritis in the population. It is probably easier to generate reliable results in hips without evidence of osteoarthritis.

For the measurement process, digital environment may play an important role, like thin or thick, dotted or continuous tool-line, screen-size and level of magnification. Use of a three-point circle as alternative to freely adjustable circles might also have an influence.[27] We used standard commercially available 24-inch high definition flat panel screen computer monitors with high, but undefined zoom levels of the radiographic image and thin continuous coloured tool-lines (1px) in our setting.

Comparing the interobserver reliability of NA and DI, there was no substantial difference related to the level of experience and both methods show excellent reliability (ICC $>$ 0.90). The interobserver agreement of the FCI scheme and the Swiss scheme is similar. There was almost no difference in the comparison between experienced observers with a good agreement (Kappa 0.687 and 0.681, respectively). In the comparison between one experienced and one unexperienced observer, the agreement was still moderate. Kappa for the FCI scheme was considerably lower than for the Swiss scheme (0.465 and 0.514, respectively). This implies the Swiss scoring scheme enables better results in unexperienced observers than the FCI system. It has to be considered that in our study only three different observers scored the images. To make a recommendation, follow-up studies should be performed with a higher number of observers. And even if it is unlikely to have unexperienced observers in an official hip screening scenario, it is obviously easier for the beginner to adopt and successfully implement the structured approach of the Swiss scoring scheme than the categorical FCI grading system. It is probably easier and more consistent to work through a table of predefined anatomical structures, with a description and predefined scoring of individual findings that sum up to a final result than to match a complex joint into a single category based on a global description.

Conclusion

The intraoperator reliability was slightly better for the NA than for the DI. Intra- and interobserver reliability showed excellent results for both, the NA and the DI. Therefore, both methods can be considered highly and equally reliable. The influence of the positioning seems to have slightly more impact on the result than the measurement itself. The FCI and the Swiss scheme seem to be equally reliable in experienced observers, but based on the better results for the unexperienced observer, we suggest novices at hip scoring to favour the Swiss scoring system.

Authors' Contributions

Julius Klever and Andreas Brühschwein contributed to conception of study, study design, acquisition of data and data analysis and interpretation. Silvia Wagner contributed to acquisition of data. Sven Reese contributed to data analysis and interpretation. Andrea Meyer-Lindenberg contributed to conception of study and study design. All authors drafted, revised and approved the submitted manuscript.

Conflict of Interest

The authors report a grants from the Gesellschaft zur Förderung Kynologischer Forschung e.V. (GKF), during the conduct of the study.

Acknowledgement

We would like to thank the Gesellschaft zur Förderung Kynologischer Forschung e.V. (GKF) for financial support provided.

References

1 Johnson J, Austin C, Breur G. Incidence of canine appendicular musculoskeletal disorders in 16 veterinary teaching hospitals from 1980 through 1989. Vet Comp Orthop Traumatol 1994;7:56–69
2 Leppänen M, Saloniemi H. Controlling canine hip dysplasia in Finland. Prev Vet Med 1999;42(02):121–131
3 Kealy RD, Lawler DF, Ballam JM, et al. Evaluation of the effect of limited food consumption on radiographic evidence of osteoarthritis in dogs. J Am Vet Med Assoc 2000;217(11):1678–1680
4 Leighton EA. Genetics of canine hip dysplasia. J Am Vet Med Assoc 1997;210(10):1474–1479
5 Powers MY, Karbe GT, Gregor TP, et al. Evaluation of the relationship between Orthopedic Foundation for Animals' hip joint scores and PennHIP distraction index values in dogs. J Am Vet Med Assoc 2010;237(05):532–541
6 Zhang Z, Zhu L, Sandler J, et al. Estimation of heritabilities, genetic correlations, and breeding values of four traits that collectively define hip dysplasia in dogs. Am J Vet Res 2009;70 (04):483–492
7 Freeman B, Evans VB, McEwan NR. Canine hip dysplasia in Irish water spaniels: two decades of gradual improvement. Vet Rec 2013;173(03):72–72
8 Silvestre AM, Ginja MMD, Ferreira AJ, Colaço J. Comparison of estimates of hip dysplasia genetic parameters in Estrela Mountain Dog using linear and threshold models. J Anim Sci 2007;85(08): 1880–1884
9 Verhoeven G, Fortrie R, Van Ryssen B, Coopman F. Worldwide screening for canine hip dysplasia: where are we now? Vet Surg 2012;41(01):10–19
10 Flückiger M. Scoring radiographs for canine hip dysplasia - the big three organisations in the world. Eur J Companion Anim Pract 2007;17(02):135–140
11 Norberg I. Höftledysplasi hos hund. Nordisk Medicin 1963;69: 246

12 Smith GK, Biery DN, Gregor TP. New concepts of coxofemoral joint stability and the development of a clinical stress-radiographic method for quantitating hip joint laxity in the dog. J Am Vet Med Assoc 1990;196(01):59–70
13 Verhoeven G, Coopman F, Duchateau L, Saunders JH, van Rijssen B, van Bree H. Interobserver agreement in the diagnosis of canine hip dysplasia using the standard ventrodorsal hip-extended radiographic method. J Small Anim Pract 2007;48(07):387–393
14 Geissbühler U, Drazovic S, Lang J, Howard J. Inter-rater agreement in radiographic canine hip dysplasia evaluation. Vet Rec 2017;180 (14):357. Doi: 10.1136/vr.104053
15 Comhaire FH, Schoonjans FA. Canine hip dysplasia: the significance of the Norberg angle for healthy breeding. J Small Anim Pract 2011;52(10):536–542
16 Smith GK, LaFond E, Gregor TP, Lawler DF, Nie RC. Within- and between-examiner repeatability of distraction indices of the hip joints in dogs. Am J Vet Res 1997;58(10):1076–1077 AccessedApril302012
17 Ginja MMD, Ferreira AJ, Silvestre M, Gonzalo-Orden JM, Llorens-Pena MP. Repeatability and reproducibility of distraction indices in PennHIP examinations of the hip joint in dogs. Acta Vet Hung 2006;54(03):387–392
18 Broeckx BJG, Vezzoni A, Bogaerts E, et al. Comparison of three methods to quantify laxity in the canine hip joint. Vet Comp Orthop Traumatol 2018;31(01):23–29
19 Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. Theriogenology 2010;73(09):1167–1179
20 Genevois JP, Chanoit G, Carozzo C, Remy D, Fau D, Viguier E. Influence of anaesthesia on canine hip dysplasia score. J Vet Med A Physiol Pathol Clin Med 2006;53(08):415–417
21 Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 2016;15(02):155–163
22 Flückiger M. How to take and read hip joint radiographs in a structured way. Eur J Companion Anim Pract 2007;17(02): 133–134
23 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(01):159–174
24 Bausman JA, Wendelburg KL. Evaluation of the effect of pelvic tilt in the coronal plane on the Norberg angle measured in ventrodorsal radiographic views of a canine hip joint bone model. Am J Vet Res 2010;71(11):1348–1353
25 Genevois JP, Cachon T, Fau D, et al. Canine hip dysplasia radiographic screening. Prevalence of rotation of the pelvis along its length axis in 7,012 conventional hip extended radiographs. Vet Comp Orthop Traumatol 2007;20(04):296–298
26 Broeckx BJ, Verhoeven G, Coopman F, et al. The effects of positioning, reason for screening and the referring veterinarian on prevalence estimates of canine hip dysplasia. Vet J 2014;201(03): 378–384
27 Bertal M, Vezzoni A, Houdellier B, et al. Intra- and inter-observer variability of measurements of the laxity index on stress radiographs performed with the Vezzoni-modified Badertscher hip distension device. Vet Comp Orthop Traumatol 2018;31(04):246–251