Prospectively recorded versus medical record-derived spinal cord injury scores in a cohort of dogs with intervertebral disk herniation

Emiko Y. Van Wie, Geoffrey T. Fosgate, Joseph M. Mankin, Nicholas D. Jeffery, Sharon C. Kerwin, Gwendolyn J. Levine, Hillary H. Greatting, Annie V. Chen, Andrew K. Barker and Jonathan M. Levine

From the Department of Small Animal Clinical Sciences (Van Wie, J. Levine, Mankin, Kerwin, Greatting), the Department of Veterinary Pathobiology (G. Levine), College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, TX; the Department of Production Animal Studies (Fosgate), University of Pretoria, Onderstepoort, South Africa; the Department of Small Animal Clinical Sciences (Jeffery, Barker), College of Veterinary Medicine, Iowa State University, Ames, IA; the Department of Small Animal Clinical Sciences (Chen), College of Veterinary Medicine and Biomedical Sciences, Washington State University, Pullman, WA

Running head: SCI scores in dogs

Keywords: neurotrauma, gait scores, modified Frankel score, canine

Abbreviations

CI	Confidence interval
ICC	Intra-class correlation
IQR	Inter-quartile range

IVDH	Intervertebral disk herniation
MFS	Modified Frankel score
SCI	Spinal cord injury
TSCIS	Texas spinal cord injury score

Address correspondence to Dr. Jonathan M. Levine; Department of Small Animal Clinical Sciences, Texas A&M University, 4474 TAMU, College Station, TX 77843 USA. e-mail: jlevine@cvm.tamu.edu; phone: (979)845-2351

The majority of the study was conducted at Texas A&M University where dogs were initially admitted, medical records were abstracted, and data were collected. Neurologic scores were reconstructed by board certified neurologists at Texas A&M University, Iowa State University and Washington State University from digitized abstracted medical records generated at Texas A&M University. Statistical analysis and portions of the study design were performed at the University of Pretoria.

The study was not supported by extra-mural or intra-mural grant funding.

This study has not been presented at a national or local meeting.

Acknowledgements: The authors wish to thank Amanda Garner, RVT and Ms. Alisha Selix for assisting in the identification of medical records.

Background: Validated spinal cord injury (SCI) scores have been established for veterinary species but are not uniformly used in practice.

Hypothesis/objectives: To determine the level of agreement of SCI scores at the time of admission versus those assigned from reconstructed medical records in a population of dogs with intervertebral disk herniation (IVDH).

Animals: Eighty-six client-owned dogs with confirmed IVDH.

Methods: Retrospective study. Medical records were reviewed for history, physical examination, neurologic examination and recorded Modified Frankel score (MFS) and Texas spinal cord injury score (TSCIS) at the time of admission. Three raters, all board certified neurologists, assigned MFS and TSCIS based on digitized abstracted medical records to each patient. These scores were then compared to the recorded score at the time of admission.

Results: Actual agreement for MFS and TSCIS derived from medical records by the 3 raters compared to prospectively derived MFS and TSCIS was 77.9% and 51.2%, respectively. A kappa value of 0.572 (95% CI 0.450, 0.694; $P < 0.001$) and an ICC of 0.533 (95% CI 0.410, 0.646; $P < 0.001$) was calculated for MFS scores. A kappa value of 0.100 (95% CI 0.000, 0.222; $P = 0.107$), and an ICC of 0.503 (95% CI 0.377, 0.620; $P < 0.001$) was calculated for TSCIS scores.

Conclusions and Clinical Importance: Results showed that SCI scores recorded at the time of admission often do not agree with those retrospectively abstracted from medical records. Agreement was less when using the more complex TSCIS scale and therefore the MFS scale might be more appropriate for use in retrospective studies.

In humans with spinal cord injury (SCI), ordinal physical examination-based scales reflecting the severity of neurological impairment have commonly been used for the last 4 decades.[1] Accepted systems have been validated via assessment of inter-rater agreement, correlation to neuro-imaging findings, and association with long-term functional outcome.[2-3] At human neurotrauma centers, SCI scores are recorded in a standardized manner for all patients to facilitate more accurate determination of initial injury severity and clinical improvement.[4]

In veterinary medicine, the vast majority of studies characterizing outcomes following SCI are retrospective investigations.[5] Physical examination-based SCI scores are frequently used in these reports to classify initial injury severity and recovery. Many of the described scoring systems have not been formally validated. Additionally, SCI scores reported in veterinary retrospective studies are often not entered into medical records at the time an animal is clinically examined.[6] Instead, they are frequently constructed by medical records abstractors who compile subjective qualifiers present in history, physical examination, or neurological examination sheets.[6] Despite this practice, few questions have been raised concerning the validity of retrospective determinations of SCI severity in veterinary studies.

Since 2008, dogs with SCI admitted to our institution have been assigned two validated SCI scores (modified Frankel score and Texas Spinal Cord Injury gait score) at the time of initial evaluation.[7] Similar to human SCI centers, scores are entered into standardized worksheets and cross checked by a clinician and technician for internal validity. Additionally, all animals have subjective neurological examination data, physical examination findings, and history recorded in an electronic record.

The purpose of this report was to determine the level of agreement between SCI scores recorded in medical records at the time of admission versus those reconstructed via record abstractors in a population of dogs with thoracolumbar intervertebral disk herniation (IVDH). We hypothesized that: 1) agreement would be poor between prospective scores and those abstracted from medical records when complex gait scores (i.e. TSCIS) were examined and 2) inter-rater agreement would be poorest amongst record abstractors when trying to qualify mild to moderate levels of SCI.

Materials and Methods

Cases

Two populations of dogs admitted to Texas A&M University Veterinary Medical Teaching Hospital between August 2008 and November 2009 with surgical thoracolumbar IVDH were utilized for this clinical study. One population was enrolled in an on-going clinical trial investigating a novel neuroprotective agent (Animal Use Protocol 2011-057). The second population consisted of dogs admitted during this same time frame that were either not eligible for trial participation or were not enrolled due to lack of owner consent. Dogs from both populations were required to meet the following inclusion criteria: 1) IVDH confirmed by advanced imaging and surgical biopsy of compressive material; 2) extra-dural IVDH-associated SCI located between the T3-L7 vertebral articulations; and 3) available medical data recorded on the day of admission, including SCI scores, typed history, physical examination forms, and neurological examination sheet.

SCI Scores

All dogs in this study were assigned 2 SCI scores at the time of admission, the modified Frankel score (MFS) and Texas Spinal Cord Injury Gait Score (TSCIS).[7] Both systems have high inter-rater agreement, correlate with the T2 weighted magnetic resonance imaging features of SCI, and predict functional outcome in dogs with thoracolumbar IVDH.[7] The MFS was defined as paraplegia with no deep nociception (grade 0), paraplegia with no superficial nociception (grade 1), paraplegia with nociception (grade 2), non-ambulatory paraparesis (grade 3), ambulatory paraparesis and ataxia (grade 4), spinal hyperesthesia only (grade 5), or no dysfunction. The TSCIS evaluates each pelvic limb individually and the gait component evaluated in this report was defined as no limb movement (grade 0), limb protraction with no ground clearance (grade 1), limb protraction with inconsistent ground clearance (grade 2), limb protraction with ground clearance >75% of steps (grade 3), ambulatory with consistent ground clearance and moderate ataxia-paresis (falls occasionally) (grade 4), ambulatory with consistent ground clearance and mild ataxia-paresis (does not fall), and normal gait (grade 6).

Procedures

Medical records from all included dogs were obtained for review. In a central database, age (in years), gender, breed, TSCIS gait score at admission, and MFS at admission were recorded. A veterinarian (HG), not involved in the delivery of medical care to included patients, compiled data from medical records for rater review. . Digitally recorded history, physical examination summary from admission, neurological examination summary from admission, and neurological examination sheet generated at the day of admission were placed into an electronic

document for each patient. The MFS and TSCIS were redacted from these abstracted medical records.

Three independent raters (JMM, NDJ, AVC), all board certified neurologists and two of which were not located at Texas A&M University, assigned MFS and TSCIS based on digitized abstracted medical records to each patient. Each rater was provided with a database that had numerical patient identifiers to enter SCI scores. Rater data were recorded in the central database for statistical analysis.

Statistical Analysis

Prospectively recorded MFS and TSCIS gait scores were considered the true measures and bias was calculated as the subtraction of this true value from the value obtained from medical records reconstruction. Bias was descriptively evaluated using modified Bland-Altman plots and calculation of 95% prediction intervals. Repeatability of score reconstruction was assessed using kappa statistics, intra-class correlation coefficient (ICC), and their corresponding 95% CI. Average bias (over all abstractors) was compared between trial and non-trial dogs (and other factors) using Mann-Whitney U tests. Kappa was calculated by entering standard formulas into a spreadsheet program and all other statistical analyses were performed using commercially available software (IBM SPSS Statistics Version 20, International Business Machines Corp., Armonk, NY). Statistical results were evaluated at the 5% level of significance.

Results

A total of 86 dogs met the inclusion criteria. Fifty of the 86 dogs were enrolled in the on-going clinical trial at the time of admission. The median age was 6 years (range 2 - 14 years). There were 7 sexually intact females, 33 spayed females, 14 sexually intact males, and 32 castrated males. Breeds included Dachshund (n=53), Beagle (n=4), Basset Hound (n=3), Shih Tzu (n=3), mixed breed (n=3), and 15 other breeds with <2 dogs each (n=20). The majority of dogs were chondrodysplastic (69/86).[5] The median MFS was 2 (IQR 2-3) and the median TSCIS

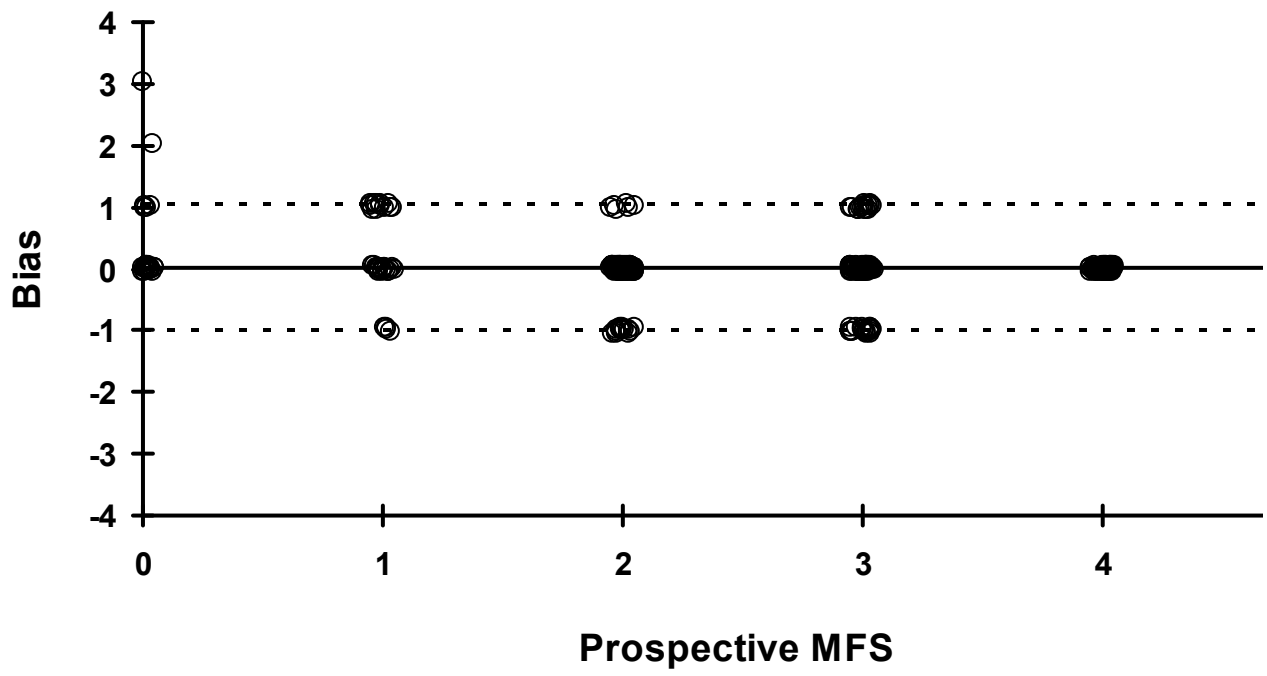

Figure 1. Combined Bland-Altman plot for three neurologists evaluating bias in medical-records derived modified Frankel scores (MFS) relative to prospectively derived scores. The solid line corresponds to the mean bias and dashed lines represent limits of the 95% prediction interval.

Table 1. Comparison of modified Frankel scores (MFS) and Texas spinal cord injury scale (TSCIS) values and mean bias calculated for three neurologists compared to prospectively derived scores.

Factor	Level	n	MFS		TSCIS	
			Median (IQR)	Median bias (IQR)	Median (IQR)	Median bias (IQR)
Overall		86	2.0 (1.7, 3.0)	0.0 (0.0, 0.0)	0.7 (0.7, 3.7)	0.3 (-0.7, 0.7)
Dog	Trial	50	2.0 (1.6, 3.0)	0.0 (0.0, 0.1)	0.7 (0.6, 2.3)	0.5 (0.0, 0.7)
	Non-trial	36	3.0 (2.0, 4.0)	0.0 (0.0, 0.0)	2.3 (0.7, 8.0)	0.0 (-1.3, 0.6)
	P value		0.034	0.320	0.021	0.018
Sex	Female	40	2.0 (1.4, 3.9)	0.0 (0.0, 0.0)	0.7 (0.7, 6.0)	0.3 (-0.3, 0.7)
	Male	46	2.3 (1.7, 3.0)	0.0 (0.0, 0.0)	1.0 (0.6, 2.8)	0.3 (-1.1, 0.7)
	P value		0.947	0.439	0.923	0.563
Breed	CD	69	2.0 (1.7, 3.0)	0.0 (0.0, 0.0)	0.7 (0.5, 3.0)	0.3 (-0.2, 0.7)
	Non-CD	17	3.0 (1.5, 4.0)	0.0 (0.0, 0.0)	2.3 (0.7, 7.5)	0.0 (-1.5, 0.7)
	P value		0.428	0.505	0.397	0.307
Age	≤ 5 yrs	48	2.0 (1.0, 3.0)	0.0 (0.0, 0.0)	0.7 (0.3, 2.3)	0.3 (-0.3, 0.7)
	> 5 yrs	38	3.0 (2.0, 4.0)	0.0 (0.0, 0.0)	2.3 (0.7, 6.6)	0.3 (-0.8, 0.7)
	P value		0.058	0.924	0.091	0.958
Duration	≤ 1 day	53	2.0 (1.0, 3.0)	0.0 (0.0, 0.0)	0.7 (0.3, 2.3)	0.3 (0.0, 0.7)
	> 1 day	33	3.0 (2.0, 4.0)	0.0 (0.0, 0.0)	2.3 (0.7, 8.0)	0.0 (-1.3, 0.7)
	P value		0.009	0.880	0.006	0.181

IQR, inter-quartile range; CD, chondrodysplastic.

was 1 (IQR 1-4) (Table 1). Dogs participating in the clinical trial and dogs with a ≤ 1day duration of SCI had significantly lower MFS and TSCIS compared to non-trial participants and dogs with >1 day duration of SCI, respectively.

The MFS derived from medical records by the 3 raters compared to prospectively derived MFS had an actual agreement of 77.9%, a kappa value of 0.572 (95% CI 0.450, 0.694; P < 0.001), and an ICC of 0.533 (95% CI 0.410, 0.646; P < 0.001) (Table 2). The kappa value and

Table 2. Agreement among medical records derived modified Frankel scores for three neurologists relative to prospectively derived values.

Scale	Level	n	Agreement*(%)	Kappa* (95% CI)	P value	ICC (95% CI)	P value
MFS	Overall	86	77.9	0.572 (0.450, 0.694)	<0.001	0.533 (0.410, 0.646)	<0.001
	MFS 0-2	44	75.0	0.434 (0.264, 0.605)	<0.001	0.431 (0.247, 0.608)	<0.001
	MFS 3,4	42	81.0	0.743 (0.568, 0.917)	<0.001	0.710 (0.571, 0.819)	<0.001
TSCIS	Overall	86	51.2	0.100 (0.000, 0.222)	0.107	0.503 (0.377, 0.620)	<0.001
	TSCIS 0	44	74.2	-0.347 (-0.518, -0.176)	<0.001	-0.059 (-0.198, 0.127)	0.744
	TSCIS 1-10	42	27.0	0.033 (0.000, 0.208)	0.709	0.554 (0.379, 0.708)	<0.001

CI, confidence interval, ICC, intra-class correlation coefficient.

* Responses that were the same as the prospectively determined scores.

ICC indicated good inter-rater agreement.[8] The mean bias of the retrospectively abstracted MFS was 0.0 (95% prediction interval -1.0, 1.0); dogs with a prospective MFS of 4 had a bias of 0 whereas dogs with a prospective MFS of 0 had larger deviations (Figure 1).

The TSCIS derived from medical records by the 3 raters compared to prospectively recorded TSCIS had an actual agreement of 51.2%, a kappa value of 0.100 (95% CI 0.000, 0.222; P = 0.107), and an ICC of 0.503 (95% CI 0.377, 0.620; P < 0.001) (Table 2). The kappa value indicated poor agreement whereas the ICC suggested good agreement between evaluators.[8] The mean bias of retrospectively abstracted TSCIS was -0.2 (95% prediction interval -3.7, 3.3) (Figure 2).

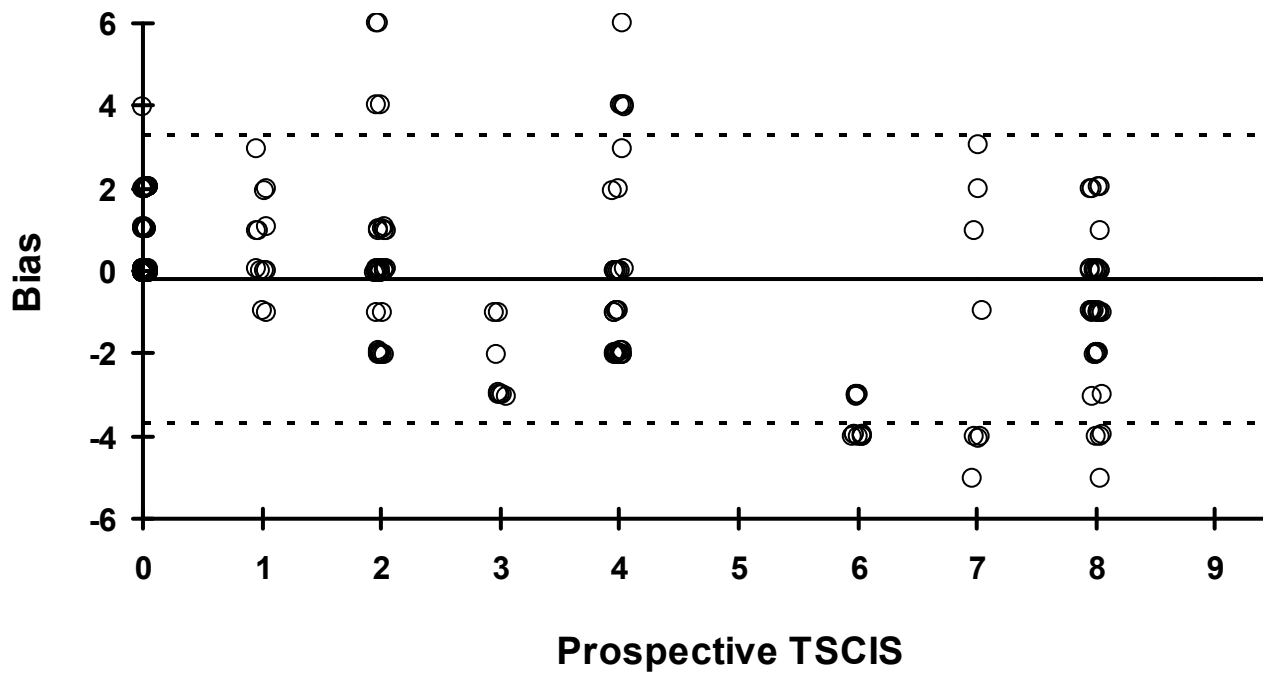

Figure 2. Combined Bland-Altman plot for three neurologists evaluating bias in medical-records derived Texas spinal cord injury scale (TSCIS) values relative to prospectively derived scores. The solid line corresponds to the mean bias and dashed lines represent limits of the 95% prediction interval.

Discussion

The present study showed that SCI scores abstracted from medical records often did not agree with those that had been prospectively recorded at the time of admission. These results parallel those from a recent study on human stroke, where actual agreement between prospectively recorded and abstracted severity scores approximated 70%.[9] While these data and those from the study reported here suggest abstraction of injury scores is feasible, mis-classification is common and could impact retrospective studies that require stratification by initial injury severity or that utilize scores to determine functional outcome.

The most commonly used ordinal SCI grading systems in veterinary medicine are based on the Frankel scale which was developed for humans with traumatic myelopathy in 1969.[1] Typical veterinary ordinal SCI systems have 3-6 severity strata determined by ambulatory status, limb movement, and nociception. The MFS used here was designed to be clinically practical; it requires minimal training, has been previously validated, and grades function 0-5.[6,7] The actual agreement between record abstractors and prospectively derived scores was 77.9% and inter-rater agreement was good. The limit of the bias for MFS abstraction was +/- 1, which is not surprising given the broad categories of this scale. The mean bias of the abstracted MFS was zero suggesting that it would be valid to use this measure for retrospective studies. However, the added variability due to abstractor errors would increase the necessary sample size over a similarly designed prospective study. The limited number of injury strata, which likely enhance the validity of abstracted data, do reduce the ability of the MFS to detect even moderate changes in neurologic dysfunction between groups. For example, dogs within the MFS 3 category may have minimal ability to advance a limb or may be capable of robust, non-weight bearing stepping movements.

The TSCIS was developed and validated to provide a more precise method to grade SCI compared to the MFS. It is similar to the 14 point open field gait score that has been used in dogs, but permits independent limb scoring and has a more limited number of strata.[7,10] The actual agreement between abstractors and prospective scores was 51.2%. The mean bias of the abstracted TSCIS was slightly negative suggesting that this score, like the MFS, would be valid to use for retrospective studies. However, the bias in TSCIS appeared to vary more extensively than MFS with more severely affected dogs tending to have a positive bias and less severely affected dogs having a negative bias. Overall, inter-abstractor variability was seen for all scores in the TSCIS. The trend in bias over level of dysfunction in addition to the added imprecision compared to the MFS should be a concern for researchers designing retrospective studies that utilize complex gait scores such as the TSCIS. These limitations cannot be overcome through large sample size alone; to use complex gait assessment tools prospective recording of scores in medical records would appear to be important.

Although the authors of the present study hypothesized that injury severity would impact agreement between abstractors and prospective scores, this was not uniformly the case. For the TSCIS bias was present qualitatively in all severity groups, but varied in direction by degree of dysfunction. In the MFS, bias was more limited in dogs that were ambulatory compared to those animals that were non-ambulatory. Data from this report would suggest that abstractors have fairly little error in determining from records whether dogs are voluntarily walking, however, bias is possible when trying to sub-classify ambulatory dogs based on the amount of ataxia and paresis that is present (TSCIS grades 8-10). Errors were even observed when differentiating dogs that lacked pelvic limb nociception (MFS 0) from those with intact nociception. These limitations are important to consider when interpreting recovery rates in retrospective series

focusing on certain sub-groups as mis-classification is likely and this may inflate reported recovery rates in severe SCI.

The study reported here has several limitations. Although a standardized history and neurologic examination form were used for data recording, those recording the information had different levels of training ranging from fourth year veterinary students to faculty clinicians. In an institutional setting, history recording is performed by fourth year veterinary students and not the clinician. Follow-up questions are commonly done by the clinician but recording is still performed by students. Thus, errors can occur with hard copy data collection. Neurologic examinations are performed by clinicians; however, recording of exam findings can vary by the individual, even when using standardized recording forms. Another limitation is that medical abstractors for this study were from multiple institutions resulting in a potential lack of familiarity with our institutional medical records and varying exposure to the MFS and TSCIS scoring systems. Finally, board certified neurologists were used as the medical abstractors. Most data collection and abstraction for retrospective studies are done by technicians, veterinary students, interns or residents, all of whom have less clinical experience. Thus, using board certified neurologists in this study might have underestimated the amount of bias and estimated more agreement than would be expected within a typical research situation.

Medical-record derived neurological injury scores are frequently used in retrospective studies focused on outcomes following SCI. While an argument can be made that all SCI studies assessing relationships with function should be prospective, this is not always feasible especially if a disease has a low prevalence, a treatment is uncommonly implemented, or a controlled prospective study would be unethical. In human medicine, the use of validated SCI scores in all neurotrauma facilities permits reliable injury data to be harvested retrospectively without

difficulty. In veterinary medicine, the use of validated SCI scores and prospective score recording for dogs is not a uniform practice. Without the use of validated SCI scoring systems and appropriate prospective recording, classification of function from medical records introduces error into assessments, especially if the assessment tool is complex. Thus, our data suggest that the simpler MFS score might be more appropriate for use in retrospective studies lacking prospectively derived scores.

References

1. Frankel HL, Hancock DO, Hyslop G, et al. The value of postural reduction in the initial management of closed injuries of the spine with paraplegia and tetraplegia. Paraplegia 1969;7:179-192.

2. Jonsson M, Tollback A, Gonzales H, et al. Inter-rater reliability of the 1992 international standards for neurological and functional classification of incomplete spinal cord injury. Spinal Cord 2000;38:675-679.

3. Yavuz N, Tezyurek M, Akyuz M. A comparison of two functional tests in quadriplegia: the quadriplegia index of function and the functional independence measure. Spinal Cord 1998;36:832-837.

4. Anonymous. International Standards for Neurological Classification of Spinal Cord Injury, revised 2002. In. Chicago, IL: 2002.

5. Brisson BA. Intervertebral disc disease in dogs. The Veterinary clinics of North America 2010;40:829-858.

6. Levine JM, Fosgate GT. Medical record-derived functional assessments of spinal cord injury. J Small Anim Pract 2009;50:507-508.

7. Levine GJ, Levine JM, Budke CM, et al. Description and repeatability of a newly developed spinal cord injury scale for dogs. Prev Vet Med 2009;89:121-127.

8. Landis JR, Koch GC. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.

9. Ishida K, Raser-Schramm JM, Wilson CA, et al. Convergent validity and interrater reliability of estimating the $ABCD^2$ score from medical records. Stroke 2013;44:803-805.

10. Olby NJ, De Risio L, Munana KR, et al. Development of a functional scoring system in dogs with acute spinal cord injuries. Am J Vet Res 2001;62:1624-1628.