PAPER

Reader performance in radiographic diagnosis of signs of mitral regurgitation in cavalier King Charles spaniels

OBJECTIVES: To measure accuracy and variability of diagnosis by radiography of heart enlargement (HE) and heart failure (HF) in mitral regurgitation (MR).

METHODS: Sixteen readers representing four levels of experience evaluated 50 sets of radiographs with varying severity of MR for presence or absence of HE, left atrial enlargement (LAE) and HF. The performance of the readers was compared with a reference standard, using area under the curve (AUC) of receiver operating characteristic (ROC) curves. The interreader agreement value kappa (K) was calculated. A subset of difficult cases of HF was analysed before and after removing an outlying reader from each group.

RESULTS: AUC for HE was 0·89, for LAE it was 0·93 and for HF it was 0·92. Experience increased certainty of diagnosis but not accuracy. K ranges were HE, 0·53 to 0·67; LAE, 0·61 to 0·69 and HF, 0·49 to 0·58. When only difficult cases of HF were read, accuracy decreased and experienced readers performed better than inexperienced. When outlying readers were excluded, the differences between experienced and inexperienced readers increased.

CLINICAL SIGNIFICANCE: LAE, not HE, should be used to evaluate the heart size and indirectly the severity of MR on radiographs. For HF, agreement among individual readers was only moderate. Studies of reader accuracy should consider the effects of interreader variability.

K. HANSSON, J. HÄGGSTRÖM, C. KVART* AND P. LORD

Journal of Small Animal Practice (2009)
50 (Suppl. 1), 44–53
DOI: 10.1111/j.1748-5827.2009.00669.x

Department of Clinical Sciences and *Department of Animal Physiology, Swedish University of Agricultural Sciences, Box 7054, SE-750 07 Uppsala, Sweden

INTRODUCTION

Thoracic radiographs are important in assessing the severity of mitral regurgitation (MR) caused by myxomatous mitral valve disease by determining general heart enlargement (HE) and left atrial enlargement (LAE), and presence of pulmonary oedema as evidence of left-sided heart failure (HF) (Häggström and others 1997, Kittleson and Kienle 1998a, Lord and Suter 1999, Sisson and others 1999). Radiographic diagnosis of pulmonary oedema has been used as the sole criterion of HF in clinical studies (Atkins and others 2007, MacDonald and others 2003), in clinical trials with multiple readers at different hospitals (Atkins and others 2007) or as part of a modified New York Heart Association (NYHA) classifications (Häggström and others 2000). Lamb and others (2000) found that experience improved the accuracy of diagnosing HE. However, only one reader represented each level of experience. Diagnosis of LAE may be more reliable than of HE (Kittleson and Kienle 1998b). Criteria for determining HE, LAE and HF are similar in the literature, but are subjectively applied, and affected by positioning, exposure and individual variations among dogs. The aim of the present study was to investigate the accuracy of readers compared with expert consensus diagnosis and the variability among readers of varying experience in subjective evaluation of HE, LAE and signs of HF in dogs with MR.

MATERIALS AND METHODS

Materials

Fifty sets of left lateral and ventrodorsal (VD) thoracic radiographs of privately owned unsedated cavalier King Charles spaniels from one to 12 years of age were selected from a large number of examinations as part of several studies on progression and treatment of MR in cavalier King Charles spaniels (Häggström and others 1997, Hansson and others 2002, Kvart and others 2002). The sets of films were assigned by two of the authors (K. H. and P. L.) in a consensus opinion to one of the five following classes, 10 sets in each class: normal, normal cardiopulmonary structures; I, slight LAE and slight general HE; II, moderate HE and LAE without HF; II+, moderate HE and LAE with HF; III+, severe HE and LAE with HF

Table 1. Radiological criteria used to classify heart enlargement and failure for reference standard

Class	Left atrial enlargement	General heart enlargement: normal group VHS	Heart failure: vascular structures and pulmonary parenchyma
I	Lateral view: straight caudal border or slight concavity at the level of atrioventricular junction. Minimal dorsal deviation of left main stem bronchus on lateral view. VD view: normal	Increased width of ventricular area with a rounded apex on lateral and VD view	Normal
II	Lateral view: straight caudal border. Dorsal deviation and slight compression of left main stem bronchus VD view: with or without bulging left atrium on left side	Trachea dorsally displaced but not more than parallel to the spine on lateral view. Increased width of ventricular area with a generally rounded appearance on both lateral and VD view	Normal
II+	Lateral view: straight caudal border. Dorsal deviation and slight compression of left main stem bronchus. VD view: with or without bulging left atrium on left side	Trachea dorsally displaced but not more than parallel to the spine on lateral view. Increased width of ventricular area with a generally rounded appearance on both lateral and VD view	Diffuse opacity mainly in the caudal lung lobes. Possibly dilated pulmonary veins. Possibly air bronchograms
III+	Lateral view: obvious dorsal deviation and compression of left main stem bronchus on lateral view. VD view: bulging left atrium on left side	Trachea dorsally displaced, towards the spine on lateral view. Heart silhouette occupying the majority of the thoracic cavity on both lateral and VD views	Diffuse opacity mainly in the caudal lung lobes. Possibly dilated pulmonary veins. Possibly air bronchograms

VD Ventrodorsal

(Table 1). The clinical evaluation made at radiographic examination included physical examination, auscultation of the heart, electrocardiography, thoracic radiography and echocardiography. All dogs were radiographed at the University Animal Hospital. All radiographs were exposed at peak inspiration or as close as possible to peak inspiration, using the same exposures for each dog and standard automatic processing. In all dogs with a heart murmur, LAE and HE, echocardiography confirmed the cause to be MR caused by myxomatous mitral valve disease. Radiographs with uncommon thoracic conformation and extreme obesity were excluded.

Reference standards for HF

Two authors classified the radiographs using the radiographic criteria given in Table 1. The determination of whether or not HF was present was based on the presence of all three of:

1 Radiographic signs of pulmonary oedema: the criteria were greater than normal background opacity reducing the contrast between lung interstitium and pulmonary vessels, and in severe cases, presence of air bronchograms. The radiographs were compared with those made on previous examinations no more than six months before the current one and with radiographs made after treatment. For the diagnosis of oedema to be made, the radiographs had to have greater diffuse opacity and less contrast between background and lung vessels than the preceding or following ones.

2 Clinical evidence of HF: clinical evidence came from owners' statements concerning dyspnoea on exertion, cough, nocturnal restlessness and exercise intolerance,

Table 2. The frequency of use of radiological signs for each diagnosis as reported by the readers. Most of the radiological signs were used

Diagnosis	Radiological sign used	Number of observers using the sign
Heart enlargement	Impression of a relative increase in cardiac length and width	10
	Dorsal displacement of the trachea	9
	Comparison with the number of intercostal spaces covered by the heart	7
	Round cardiac silhouette	5
	Straight caudal margin	3
	Cardiac width exceeding two-third of the width of the thoracic cavity	3
	Reversed "D" shape	1
	Vertebral heart scale measurement	1
	Decreased distance between the heart and the diaphragm	1
Left atrial enlargement	Dorso-caudally located bulging soft tissue opacity on the lateral view	16
	Left side bulge on the VD view	13
	Dorsally displaced trachea and left main bronchus	13
Heart failure	Increased opacity in the caudo-dorsal lungfield	16
	Dilation of pulmonary veins	13
	General cardiomegaly	3
	Enlarged left atrium	2

VD Ventrodorsal

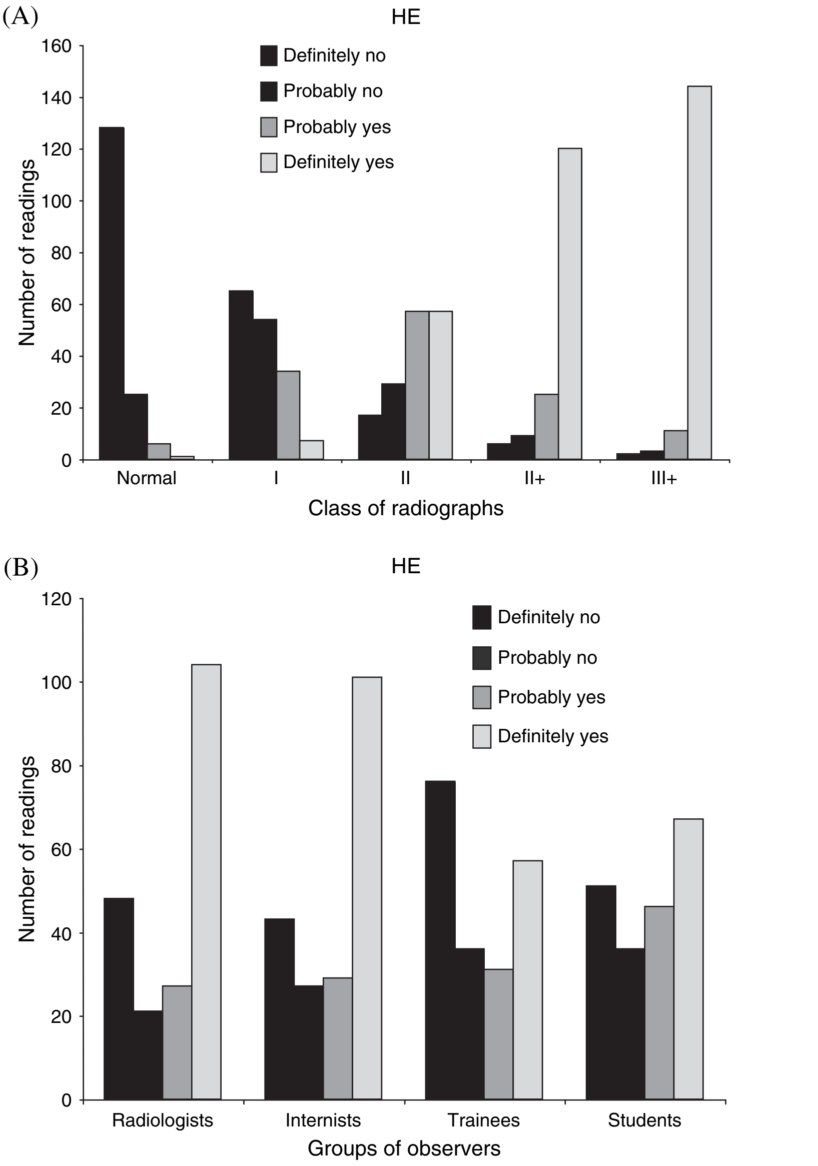

FIG 1. Reader confidence for diagnosing heart enlargement (HE). The true frequency was 40 negative and 160 positive diagnoses. (A) Confidence related to class of radiograph. Radiologists and internists were more definite in their diagnosis, and the students less definite than average. Most uncertainty is with mild (I) and moderate (II) enlargement without failure. The certainty of diagnosis in the II+ class was much higher than in the II class. When failure was present (II+), the readers were probably biased towards diagnosing HE. Agreement among groups of readers was greatest in the normal and III+ groups. (B) Distribution of the number of observations of HE present for each group of readers. The trainees had the highest number of false-negative observations. Students were less certain than the others

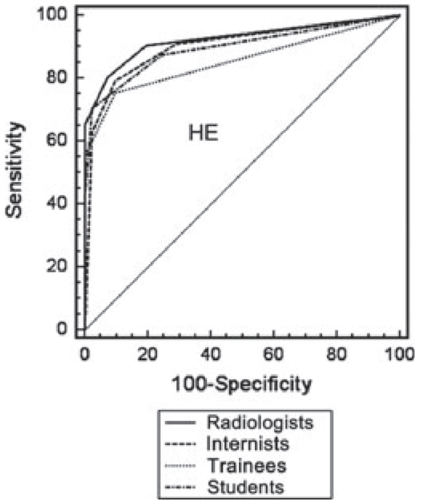

FIG 2. Receiver operating characteristic curves of each group of readers for heart enlargement (HE). Note that all curves have a similar shape and are very close to each other where sensitivity increases rapidly without much loss of specificity. Radiologists were significantly better than trainees (P=0·01). See Table 3 for statistics of these curves

and on physical examination, tachycardia (heart rate >140 beats per minute), tachypnoea (respiration rate >28 breaths per minute), loss of sinus arrhythmia (Häaggström and others 1996), dyspnoea and increased lung sounds typical for oedema.

3 Response to treatment with furosemide, evaluated by owner's opinion, and clinical and radiographic examinations. If the radiographs were equivocal for pulmonary oedema, the cardiologists' clinical evaluation determined the decision to treat the dog.

Each set of radiographs was assigned a random number between 1 and 50 and sorted according to number. The material consisted of 320 (20 sets, 16 readers) true-positive sets of HF (classes II+ and III+) and 480 (30 sets, 16 readers) true-negative sets of HF (classes 0, I and II) and 640 true-positive sets of HE and LAE (classes I, II, II+ and III+) and 160 true-negative sets of HE and LAE (class 0). They were presented in the same random order to all readers.

Readers

The groups of readers represented four levels of experience with four individuals per level. The experience levels were (1) radiologists, European Diplomates in Veterinary Diagnostic Imaging each from a different country; (2) internists, small animal clinicians with more than 15 years of experience, each from a different practice in Sweden; (3) trainees, clinicians enrolled in a three year national (Swedish) training programme towards specialisation in canine and feline diseases, all in their second or third year of training, each from a different practice and (4) students, fifth year veterinary students who volunteered to participate, and whose education in small animal medicine and radiology took place in the fourth year.

Instructions to readers

The 16 readers evaluated the radiographs following instructions presented immediately before the evaluation. The readers were informed of the breed and that the

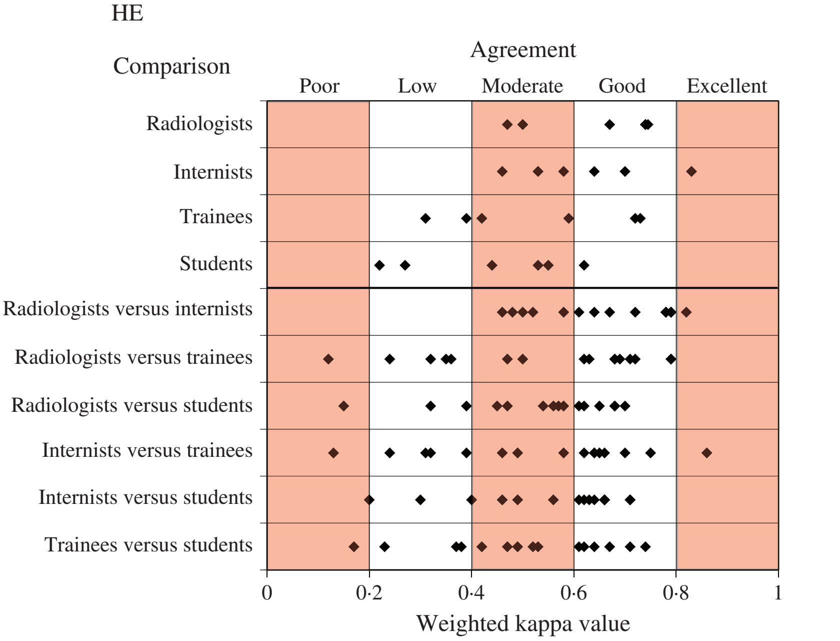

FIG 3. Weighted kappa values for heart enlargement (HE). Each diamond represents one comparison between two readers within (six pairs) or between groups (15 pairs). Radiologists agreed better with each other and with internists than other groups

dogs were either normal or had various degrees of MR with or without HF in the form of pulmonary oedema. The proportion of normal to diseased dogs was not given. The questions asked for each set were (1) Is the heart generally enlarged? (2) Is the left atrium enlarged? and (3) Has the dog had HF? For each of these questions, the readers chose one of the following four alternative degrees of confidence: (1) definitely no, (2) probably no, (3) probably yes and (4) definitely yes. Each person evaluated the radiographs sequentially and on one occasion. At the end of the session, the readers were asked to list the criteria he/she used to define HF, HE and LAE. No time limit was set for completion.

Measurements of examination performance

Readers' responses were compared with the reference standard. Receiver operating characteristic (ROC) curves and the area under the curve (AUC) were calculated (Langlotz 2003, Metz 1978, 2006, Obuchowski 2003). Statistical analyses were performed with statistical software programs JMP 4.02 (SAS Institute Inc., 2000), and MedCalc software version 9.3.8.0 (1993-2007; Mariakerke, Belgium). AUCs were compared in MedCalc by the method of Hanley and McNeil (1983) with a P value of 0·05 as the level of significance. ROC curves incorporate both sensitivity and specificity. AUC, by incorporating all degrees of certainty of diagnosis, gives a general idea of the accuracy of a diagnostic test (Langlotz 2003, Metz 2006, Obuchowski 2003). An area of 0·50 indicates a diagnostic test which is no better than a guess, whereas 1·00 is a perfect test. AUC is independent of the criteria of strictness of diagnosis, that is, whether the reader tends to over- or underdiagnose the condition. A sensitive examination is particularly valuable when the consequences of a false-negative result are undesirable (Häggström and others 2000, Lamb 2007a, Obuchowski 2003). All combinations of pairs of readers within each group and between groups were compared for agreement by linearly weighted kappa, K. K is the agreement between observers which is greater than chance. The K value can be interpreted as follows (Altman 1991): value of K<0·20, poor; 0·21 to 0·40, fair; 0·41 to 0·60, moderate; 0·61 to 0·80, good and 0·81 to 1·00, very good agreement. Linear weighting classifies ordered differences by weighting them according to the degree of difference, each step of difference being weighted equally.

As we found large interreader variability and no significant effect of experience for HF, we analysed a subset of difficult cases (classes II and II+). As the effect of outliers can be considerable when reader variability is large (Gur and others 2005), we tested the effect of removing an outlier from each group. Outliers were determined as the AUC value that differed most from the mean of the group. If there were two equal values, the lower one was selected.

RESULTS

The evaluations took three to five hours. The frequency of use of radiological signs used for each diagnosis as reported by the readers is listed in Table 2. The analyses of the decisions of the readers for HE is shown in Figs 1–3, for LAE in Figs 4–6 and for HF in Figs 7–9. Figures 3, 6 and 9 are the distribution of K values within and between groups of readers. Table 3 shows the AUCs for HE and LAE. Table 4 shows the AUC for HF for all classes and for difficult classes, before and after removing the outlying readers. Table 5 lists the P values of comparisons of the groups of readers.

For HE, the trainees had the highest number of false-negative observations. Students were less certain than the others (Fig 1B). All groups of readers were more confident and more accurate in diagnosing LAE than HE, and their agreement was better (Figs 1–6, Table 3). For HE, agreement was related to experience, but for LAE, only radiologists and trainees had close within-group agreements. For HF, radiologists and internists were most confident, students the least (Fig 7). Experience did not affect the AUCs for HF (Fig 8). Two readers, a radiologist and an internist, were the main cause of overdiagnosed HF in class II (Fig 7A). These two outlying readers lowered the AUC and K values for the radiologists and internists (Table 4 and Fig 9). For HF, in the subset of difficult cases, accuracy decreased (Tables 4, 5). The lower P values of the comparisons of the experienced and inexperienced readers suggested that when the diagnosis was difficult, experience made a difference, although the number of sets was lower (Table 5). After removing the outlying readers from the evaluation of

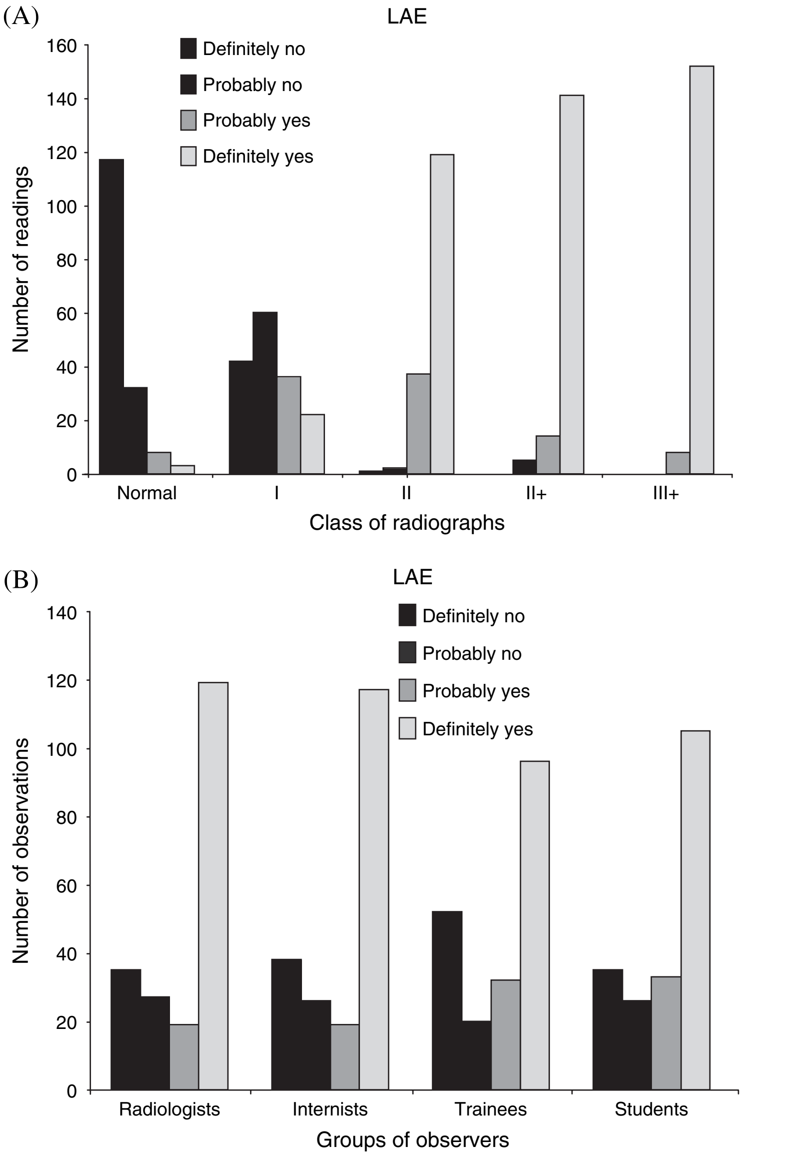

FIG 4. Reader confidence for diagnosing left atrial enlargement (LAE). The true frequency was 40 negative and 160 positive diagnoses. (A) Confidence related to class of radiograph. The readers were more confident in diagnosing LAE than heart enlargement (HE), particularly in the clinically important classes of moderate enlargement (II and II+). (B) Distribution of the number of observations LAE present for each group of readers. All groups of readers had a high number of "definitely yes" decisions. Trainees and students were more confident in diagnosing LAE than HE

all classes and from the difficult classes, the differences were even greater, experience becoming significant in the difficult classes, although the number of sets and readers was reduced (Table 5).

DISCUSSION

The ROCs were characterised by the desirable trait of rising to a high sensitivity in the range of low false-positive rate (100-specificity) at the left side of the graph (Obuchowski 2000, 2003). The overall good results of AUC hid considerable individual differences in interpretation that are revealed by the K values. Interreader agreement in this kind of study is usually in the range of 0·5 to 0·7 (Albaum and others 1996, Gierada and others 2008, Hopstaken and others 2004, Manninen and others 1991, Quekel and others 2001b). Interreader variability is the largest component of variability in this type of study (Gur and others 2005).

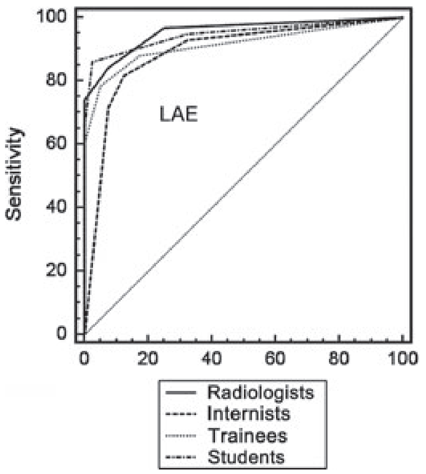

FIG 5. Receiver operating characteristic curves of each group of readers for left atrial enlargement (LAE). Radiologists were significantly better than all clinicians but not students. See Table 3 for statistics of these curves

HE and LAE

Signs of change of shape of the heart and the position of the left main bronchus are easier to evaluate than size comparison with the thorax, which varies with breathing and conformation of the thorax, so that LAE is a better radiological sign for indirectly assessing MR than HE. In general practice, the thorax is more likely to be rotated slightly than in our material, leading to overdiagnosis of a falsely rounded heart shadow. For the diagnosis of HE in human beings, interreader K values were also low (Manninen and others 1991), and in children, radiographic evaluation of HE was not as sensitive as echocardiography (Satou and others 2001), indicating the generality of the problem. The underdiagnosis of HE and LAE in dogs with slight enlargement has little clinical significance, as the diagnosis of presence of MR is made by auscultation and echocardiography.

In many practices, the right lateral and dorsoventral views may be used instead of the left lateral and VD views we used. If the readers were unused to the left lateral projection, in which the heart appears rounder than in the right lateral, they could overdiagnose HE, but instead, mild HE was actually underdiagnosed (Fig 1A). The high specificity for HE (Fig 2) also refutes this possibility. In the VD view, the heart appears more elongated and less round than in the dorsoventral view (Ruehl and

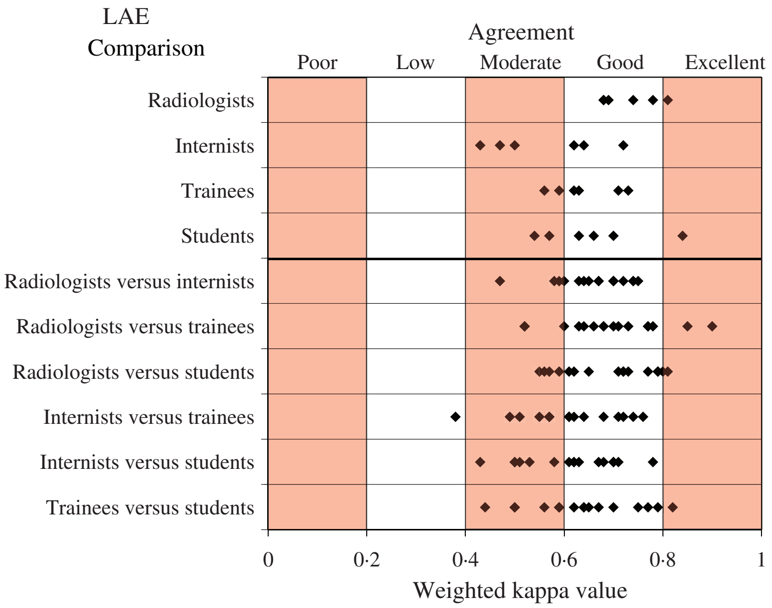

FIG 6. Kappa values for left atrial enlargement (LAE). Each diamond represents one comparison between two readers or groups. Within- and between-group values were significantly higher than for heart enlargement, and the spread was less, especially within radiologists

Thrall 1981). If the readers were not used to the VD view, this difference might have contributed to the underdiagnosis of HE as roundness is a criterion of enlargement. All the readers, except one of the radiologists, were used to interpreting heart size using the views we used, and it is unlikely that radiologists would have been unaware of the differences between left and right lateral views.

Heart failure

The initial analysis showed overall good accuracy for all groups of readers, but only moderate agreement within and among the groups of readers, and, surprisingly, no positive effect of experience. Experience improved confidence of diagnosis, confirming other studies (Butela and others 2001, Lamb and others 2000, Monnier-Cholley and others 2004, Thompson and others 2007). We had expected that experience would improve accuracy (Butela and others 2001, Kido and others 1994, Thompson and others 2007), but this became apparent only with the subset of difficult groups. When variability was large, the effect on evaluation of experience of removing low AUC outliers illustrates the importance of the individuals in each group (Gur and others 2005, Obuchowski 2000, Obuchowski and Zepp 1996). Removing outliers add information about the sensitivity of the results to individual readers. The reason for the variability is the difficulty of diagnosing pulmonary oedema. Mild pulmonary oedema is a subtle sign, and subtle signs as a subset of cases had decreased accuracy and greater variability in reader performance (Rockette and others 1995). If exposure factors or film type were different in our cases, with less contrast or less exposure than the reader was familiar with, the lungs would have been more radio-opaque and may have caused the consistent overdiagnosis of HF made by one radiologist and one internist, which affected the performance of their groups.

The increase in false-positive diagnosis of HF from classes I to II (Fig 5A) is likely to be the result of bias induced by the presence of a larger heart. HE and LAE would cause false-positive diagnoses of HF if the size of the heart were weighted more strongly than mild pulmonary oedema. As few observers stated that they used HE and LAE in their evaluation of HF (Table 2), the effect would have been unconscious.

In comparison with the NT-proANP (a natriuretic peptide) test for discriminating compensation from decompensation in cavalier King Charles spaniels, the AUC for HF was 0·92, AUC but for proANP was 0·99±0·1 (Häggström and others 2000). The 8 per cent difference was probably mostly because of diagnostic errors in the difficult groups. The performance of the readers in our study was approximately the same (0·84±0·01 to 0·91±0·01) as that of radiologists and residents reading subtle interstitial abnormalities (Kido and others 1994), a task of similar difficulty to reading the lungs for oedema. The moderate interreader agreements cast doubt on the reliability of criteria for HF, which rely solely or heavily on radiological diagnosis of pulmonary oedema, as in the modified NYHA classification, particularly if different veterinarians, even if experts, interpret the radiographs, as in multicentre trials.

Study design and limitations

As the purpose of training is to approach the standard of experts, a consensus of experts is a valid standard to compare groups with different levels of training. The use of consensus opinion as a reference standard for creating ROC curves is well established in radiology, as exemplified by studies of interstitial lung abnormalities (Kido and others 1994), stomach cancer (Iinuma and others 2000) and computed tomographic colonoscopy (Taylor and others 2007). The criteria we used for HF are well-established clinical end-points (Häggström 2000, Kvart and others 2002).

The prevalence of disease in the study was much higher than in general practice, which is usual in reader performance studies to keep the number of cases as low as possible, reducing the risk of reader fatigue. We chose a wide spectrum of severity that included easy as well as difficult decisions, and 40 per cent of the hearts were normal or only slightly enlarged, and the ratio of negative to positive HF cases was 30:20, typical for this type of study (Gur and others 2003, Metz 2006, Obuchowski 2000). High prevalence of disease compared with low prevalence in ROC studies of lung diseases did not affect accuracy but confidence was very slightly but statistically significantly decreased (Gur and others 2007).

It is possible that the results were affected by learning during the reading, causing a change in the criteria for decisions as the reader adjusted to the radiographs. If this had been a significant factor, the experienced readers would probably have been

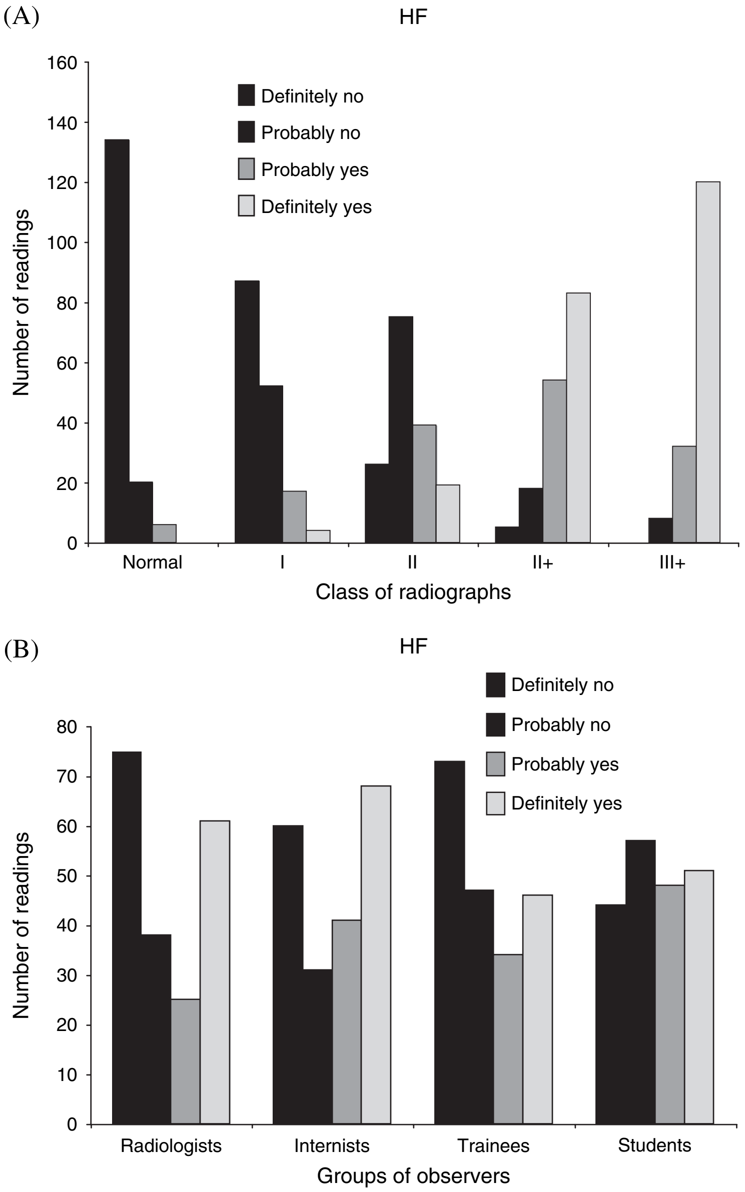

FIG 7. Reader confidence for diagnosing heart failure (HF). The true frequency was 120 negative and 80 positive diagnoses. (A) Confidence related to class of radiograph. It was high only at the extremes of heart size. The certainty of diagnosis was low for class II radiographs. (B) Distribution of the number of observations of HF present for each group of readers. Radiologists and experienced clinicians were most confident, students the least

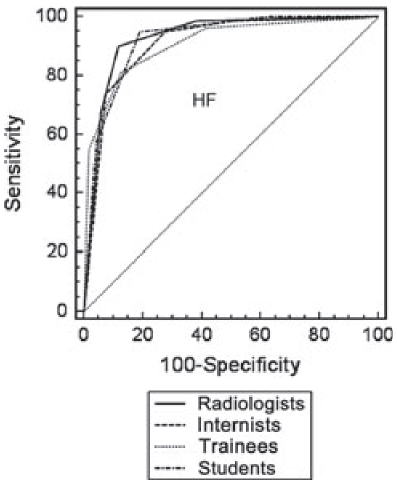

FIG 8. Receiver operating characteristic curves of each group of readers for heart failure (HF). All groups were similar, with no significant difference in shape or area under the curve (AUC). There were no significant differences among the AUCs of the readers (Table 4), nor between experienced (radiologists and experienced clinicians) and inexperienced groups

less affected by it than the inexperienced readers, yet we found that their agreement was not better. It is possible to "train" the readers with sample cases before starting the readings. The one paper on the subject found no effect of training on interstitial lesions but an effect on nodule diagnosis and warned against the effects of such "interventions" (Gur and others 1990).

Ideally, the readers should be representative of the populations of the groups of readers in practice, but "practice" has to be defined. It may vary, for example, by country, degree of specialisation, case type and frequency. All these factors cannot be taken into account in a test design. Readers may behave differently in practice than in a laboratory environment (Gur 2004). The radiologists were from different countries and training centres, yet three of the four closely agreed with each other, while one overdiagnosed HF. The internists came from different practices, and although they all had similar training as students, that was at least 15 years earlier, three of the four closely agreed with each other. The students were the most homogeneous group, all from the same class, but their agreement was closer only when diagnosing HF. The students had recently been taught by the radiologists and cardiologists establishing the reference standard, and this may have compensated for their lack of experience. As the students were volunteers, they were likely to have been better than the average student. Their recent training may have made them a more homogenous group than the others. That and their less-definite choices would explain the lesser spread of K. Although bias and study design may have affected the absolute numbers, the conclusion of the study that LAE is a better sign of MR than general HE, is not affected. The effect of removing the outlying AUC of a reader in each experience group in the HF tests shows that the study did not have a sufficient number of readers to minimise the effects of variability within groups; the level of difference was smaller than we had estimated because the variability was larger. This is a common problem: a review of 29 multiple-reader studies found that only two had more than three readers, and in only seven were differences among readers' interpretations described

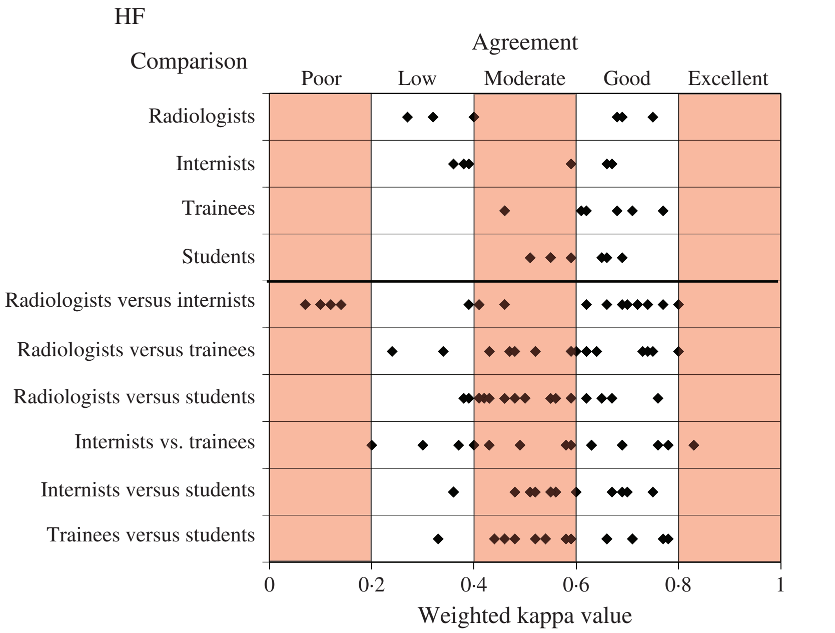

FIG 9. Kappa values for heart failure (HF). Each diamond represents one comparison between two readers or groups. The agreement between groups is widely spread. The values of low agreement within the radiologists and internists were caused by one reader in each group. Students agreed better with each other than the other readers because they were less confident and chose "probably" more. These differences were weighted less than "definitely"

Table 3. Area under the curve for HE and LAE according to experience

Reader groups (n=200)	HE	LAE
Radiologists	0·93±0·02^{1}	0·96±0·01^{2,3}
Internists	0·90±0·02	0·89±0·02^{2,4}
Trainees	0·86±0·03^{1}	0·92±0·02^{3}
Students	0·90±0·02	0·95±0·02^{4}
All (n=800)	0·89±0·01^{5}	0·93±0·01^{5}

HE heart enlargement, LAE left atrial enlargement
LAE was more accurately diagnosed than HE, and radiologists performed slightly better than the other readers. Superscript numbers indicate significantly different comparisons

(Obuchowski and Zepp 1996). Tables to calculate the numbers of readers and patients have been published based on the accuracy and desired and suspected level of difference (Obuchowski 2000).

In the present paper, we used weighted K because we had ordered categorical data. Simple K would not differentiate between small and large disagreements. Nevertheless, K analysis is sensitive to the number of categories and the number of observations within each category, making it difficult to compare the results of this study with others (Altman 1991); hence the vague distinctions of the levels of agreement previously cited.

The radiographs had less variability in exposure, contrast and positioning than in a typical clinical practice, reducing errors because of these factors. Only one breed was used, so the variations in heart size and shape were less than if many breeds had been used, and the results were almost certainly better than would be obtained in a hospital population.

This study was limited to the three main radiological signs in MR. In practice, the task of the reader may be not only to help determine if a dog has HE, LAE or HF, but also to differentiate HF from respiratory diseases causing similar clinical signs in dogs with a systolic murmur (DeFrancesco and others 2007, Prosek and others 2007). Although a difference in performance would be expected when the tasks are different, no difference in AUC was found between free choice of diagnosis and choice of three defined diagnoses, one of which included interstital lung disease (Rockette and others 1990), an appearance similar to oedema, which suggests that the results of this study would not have changed much if other diseases had been included.

This study measured the accuracy and variability of readers for the signs seen in MR, the second level of six in the hierarchic model for diagnostic efficacy of the National Council on Radiation Protection and Measurements (Metz 2006). The accuracy of a test does not allow a judgement of its effect on diagnosis in practice unless the frequency of the condition tested is known for the population of the practice (pretest probability) being

Table 4. AUC for HF showing the effects of selecting difficult cases only (classes II and II+) and removing outlying AUCs

Reader groups	HF (n=200)	AUC of each reader for all classes of radiographs	AUC of each group of readers minus outliers (n=150)	AUC of each group of readers for difficult cases, classes II and II+ (n=80)	AUC of each reader for classes II and II+	AUC of each group of readers for classes II and II+ minus outliers (n=60)
Radiologists	0·93±0·02^{1}	(0·88), 0·97, 0·97, 0·99	0·98±0·01^{1,3}	0·83±0·05^{3,11}	(0·61), 0·91, 0·97, 0·98·	0·95±0·03^{11}
Internists	0·91±0·02	(0·85), 0·98, 0·95, 0·93·	0·96±0·02^{4}	0·80 ±0·05^{4}	(0·65), 0·95, 0·88, 0·80·	0·88±0·05
Trainees	0·91±0·02	0·97, 0·91, (0·88), 0·94·	0·94±0·02^{5}	0·75 ±0·06^{5}	0·87, 0·75, (0·64), 0·83·	0·81±0·06
Students	0·92±0·02	(0·97), 0·93, 0·89, 0·91·	0·94±0·02^{6,8}	0·78 ±0·05^{6}	0·92· 0·80, (0·67), 0·76·	0·81±0·06^{8}
All	0·92±0·01^{2} (n=800)		0·95 ±0·01^{2,7,9} (n=600)	0·79 ±0·03^{7,10} (n=320)		0·86±0·03^{9,10} (n=180)

AUC area under the curve, HF heart failure
After removing one outlying reader from each group, the effect of experience increased (see Table 5 for P values). Accuracy when reading a subset of only difficult cases (classes II and II+) decreased, more for inexperienced readers. When an outlier (the worst reader) was removed from each of these two groups, the accuracy increased, despite the decrease in the sample sizeOutliers are given in parentheses. Superscript numbers indicate significantly different comparisons

Table 5. P values for comparisons of AUC between groups of readers of HF

Comparison	P value (four readers, n =200)	P value, outliers removed (three readers, n=150)	Classes II and II+ P value (four readers, n=80)	Classes II and II+ P value, outliers removed (three readers, n=60)
Radiologists-internists	0·40	0·47	0·63	0·30
Radiologists-trainees	0·46	0·06	0·19	0·02*
Radiologists-students	0·82	0·10	0·36	0·04*
Internists-trainees	0·96	0·29	0·34	0·23
Internists-students	0·66	0·29	0·62	0·03*
Trainees-students	0·56	0·92	0·64	0·90
Experienced-inexperienced	0·80 (n=400) (0·92 *versus* 0·91)	0·07 (n=200) (0·96 *versus* 0·92)	0·31 (n=160) (0·81 *versus* 0·76)	0·003* (n=80) (0·94 *versus* 0·79)

AUC area under the curve, HF heart failure
Removing outliers greatly decreased P values of comparisons between experienced (radiologists and internists) and inexperienced (trainees and students) readers, and some differences became significant (*)

considered (Lamb 2007a, b), which allows calculations of the after test probability of the disease being present or absent. The desired outcome influences the choice of specificity and sensitivity to be selected on the ROC curve (Lamb 2007a, c, Metz 2006, Obuchowski 2003). These allow assessment at the third and fourth levels, diagnostic-thinking efficacy and therapeutic efficacy.

Measures to improve accuracy

The methods used to set the reference standards could be adapted to practice. Consensus diagnosis in equivocal cases could be performed in large or specialist practices. Three readers but not two, improved accuracy (Hessel and others 1978, Norgaard and others 1990, Quekel and others 2001a). A comparison of baseline radiographs with radiographs made at the time of suspected failure improved accuracy (Snashall and others 1981), as did chronological reading of successive films of suspected rheumatoid arthritis (van Der Heijde and others 1999). If the suspicion is high enough for the dog to be treated, radiographs should be compared with follow-up radiographs to see if the suspected oedema regressed after treatment.

The vertebral heart scale (VHS) measurement might be expected to improve the accuracy of diagnosing HE, but Lamb and others (2000) found no improvement when three readers used it as an aid. The VHS method is not accurate enough to reliably differentiate mild HE from normal (Hansson and others 2005, Lamb and others 2000). Veterinarians should not try to diagnose the presence or absence of mild MR by radiography, but use the presence and quality of the heart sounds and murmurs (Häggström and others 1995). Radiographs should be used to help determine if the LA is enlarged in cases of suspected HF because of clinical signs and/or lung infiltrate to determine if MR is the cause of these signs.

Conclusions

LA is a more reliable radiographic sign of MR than HE. HF in the form of pulmonary oedema was generally accurately diagnosed but considerable individual discrepancies warrant measures to improve accuracy when the heart is enlarged. Experienced readers were more confident of their diagnosis and performed better than inexperienced readers in difficult cases. Studies using HF as an end-point should not rely solely on radiographic determination of pulmonary oedema. Studies of reader performance using ROC analysis require sufficient numbers of readers to minimise the effect of outliers.

Acknowledgements

The authors thank Swedish Medical Research Council Project no. 3392, Agria Insurance Ltd, Thure F and Karin Forssberg Foundation.

References

Albaum, M. N., Hill, L. C., Murphy, M., Li, Y. H., Fuhrman, C. R., Britton, C. A., Kapoor, W. N. & Fine, M. J. (1996) Interobserver reliability of the chest radiograph in community-acquired pneumonia. PORT Investigators. *Chest* **110**, 343–350

Altman, D. G. (1991) Practical Statistics for Medical Research. London: Chapman & Hall

Atkins, C. E., Keene, B. W., Brown, W. A., Coats, J. R., Crawford, M. A., DeFrancesco, T. C., Edwards, N. J., Fox, P. R., Lehmkuhl, L. B., Luethy, M. W., Meurs, K. M., Petrie, J. P., Pipers, F. S., Rosenthal, S. L., Sidley, J. A. & Straus, J. H. (2007) Results of the veterinary enalapril trial to prove reduction in onset of heart failure in dogs chronically treated with enalapril alone for compensated, naturally occurring mitral valve insufficiency. *Journal of American Veterinary Medical Association* **231**, 1061–1069

Butela, S. T., Federle, M. P., Chang, P. J., Thaete, F. L., Peterson, M. S., Dorvault, C. J., Hari, A. K., Soni, S., Branstetter, B. F., Paisley, K. J. & Huang, L. F. (2001) Performance of CT in detection of bowel injury. *AJR American Journal of Roentgenology* **176**, 129–135

DeFrancesco, T. C., Rush, J. E., Rozanski, E. A., Hansen, B. D., Keene, B. W., Moore, D. T. & Atkins, C. E. (2007) Prospective clinical evaluation of an ELISA B-type natriuretic peptide assay in the diagnosis of congestive heart failure in dogs presenting with cough or dyspnea. *Journal of Veterinary Internal Medicine* **21**, 243–250

van Der Heijde, D., Boonen, A., Boers, M., Kostense, P. & van Der Linden, S. (1999) Reading radiographs in chronological order, in pairs or as single films has important implications for the discriminative power of rheumatoid arthritis clinical trials. *Rheumatology (Oxford)* **38**, 1213–1220

Gierada, D. S., Pilgram, T. K., Ford, M., Fagerstrom, R. M., Church, T. R., Nath, H., Garg, K. & Strollo, D. C. (2008) Lung cancer: interobserver agreement on interpretation of pulmonary findings at low-dose CT screening. *Radiology* **246**, 265–272

Gur, D. (2004) Imaging technology and practice assessments: diagnostic performance, clinical relevance, and generalizability in a changing environment. *Radiology* **233**, 309–312

Gur, D., Rockette, H. E., Good, W. F., Slasky, B. S., Cooperstein, L. A., Straub, W. H., Obuchowski, N. A. & Metz, C. E. (1990) Effect of observer instruction on ROC study of chest images. *Investigative Radiology* **25**, 230–234

Gur, D., Rockette, H. E., Warfel, T., Lacomis, J. M. & Fuhrman, C. R. (2003) From the laboratory to the clinic: the "prevalence effect". *Academic Radiology* **10**, 1324–1326

Gur, D., Rockette, H. E., Maitz, G. S., King, J. L., Klym, A. H. & Bandos, A. I. (2005) Variability in observer performance studies experimental observations. *Academic Radiology* **12**, 1527–1533

Gur, D., Bandos, A. I., Fuhrman, C. R., Klym, A. H., King, J. L. & Rockette, H. E. (2007) The prevalence effect in a laboratory environment: changing the confidence ratings. *Academic Radiology* **14**, 49–53

Häggström, J., Kvart, C. & Hansson, K. (1995) Heart sounds and murmurs: changes related to severity of chronic valvular disease in the Cavalier King

Charles spaniel. *Journal of Veterinary Internal Medicine* **9**, 75–85

Häggström, J., Hamlin, R. L., Hansson, K. & Kvart, C. (1996) Heart rate variability in relation to severity of mitral regurgitation in Cavalier King Charles spaniels. *Journal of Small Animal Practice* **37**, 69–75

Häggström, J., Hansson, K., Kvart, C., Karlberg, B. E., Vuolteenaho, O. & Olsson, K. (1997) Effects of naturally acquired decompensated mitral valve regurgitation on the renin-angiotensin-aldosterone system and atrial natriuretic peptide concentration in dogs. *American Journal of Veterinary Research* **58**, 77–82

Häggström, J. H. K., Kvart, C., Pederson, H. D., Vuolteenaho, O. & Olsson, K. (2000) Relationship between different natriuretic peptides and severity of naturally acquired mitral regurgitation in dogs with chronic myxomatous valve disease. *Journal of Veterinary Cardiology* **2**, 7–16

Hanley, J. A. & McNeil, B. J. (1983) A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **148**, 839–843

Hansson, K., Häggström, J., Kvart, C. & Lord, P. (2002) Left atrial to aortic root indices using two-dimensional and M-mode echocardiography in cavalier King Charles spaniels with and without left atrial enlargement. *Veterinary Radiology and Ultrasound* **43**, 568–575

Hansson, K., Häggström, J., Kvart, C. & Lord, P. (2005) Interobserver variability of vertebral heart size measurements in dogs with normal and enlarged hearts. *Veterinary Radiology and Ultrasound* **46**, 122–130

Hessel, S. J., Herman, P. G. & Swensson, R. G. (1978) Improving performance by multiple interpretations of chest radiographs: effectiveness and cost. *Radiology* **127**, 589–594

Hopstaken, R. M., Witbraad, T., van Engelshoven, J. M. & Dinant, G. J. (2004) Inter-observer variation in the interpretation of chest radiographs for pneumonia in community-acquired lower respiratory tract infections. *Clinical Radiology* **59**, 743–752

Iinuma, G., Ushio, K., Ishikawa, T., Nawano, S., Sekiguchi, R. & Satake, M. (2000) Diagnosis of gastric cancers: comparison of conventional radiography and digital radiography with a 4 million-pixel charge-coupled device. *Radiology* **214**, 497–502

Kido, S., Ikezoe, J., Takeuchi, N., Kondoh, H., Johkoh, T., Kohno, N., Tomiyama, N., Yamagami, H., Naito, H., Arisawa, J., Morimoto, S. & Kozuka, T. (1994) Interpretation of subtle interstitial lung abnormalities: conventional versus film-digitized radiography. *Radiology* **192**, 171–176

Kittleson, M. D. & Kienle, R. D. (1998a) Myxomatous atrioventricular valvular degeneration. In: Small Animal Cardiovascular Medicine. Mosby, St Louis, MO, USA. pp 297–318

Kittleson, M. D. & Kienle, R. D. (1998b) Radiology of the cardiovascular system. In: Small animal cardiovascular medicine. Mosby, St Louis, MO, USA. pp 47–71

Kvart, C., Häggström, J., Pedersen, H. D., Hansson, K., Eriksson, A., Järvinen, A. K., Tidholm, A., Bsenko, K., Ahlgren, E., Ilves, M., Åblad, B., Falk, T., Bjerkfas, E., Gundler, S., Lord, P., Wegeland, G., Adolfsson, E. & Corfitzen, J. (2002) Efficacy of enalapril for prevention of congestive heart failure in dogs with myxomatous valve disease and asymptomatic mitral regurgitation. *Journal of Veterinary Internal Medicine* **16**, 80–88

Lamb, C. R. (2007a) Statistical briefing: sensitivity and specificity. *Veterinary Radiology and Ultrasound* **48**, 189

Lamb, C. R. (2007b) Statistical briefing: estimating the probability of disease. *Veterinary Radiology and Ultrasound* **48**, 297–298

Lamb, C. R. (2007c) Statistical briefing: spPInS and SnNOuts. *Veterinary Radiology and Ultrasound* **48**, 486–487

Lamb, C. R., Tyler, M., Boswood, A., Skelly, B. J. & Cain, M. (2000) Assessment of the value of the vertebral heart scale in the radiographic diagnosis of cardiac disease in dogs. *Veterinary Record* **146**, 687–690

Langlotz, C. P. (2003) Fundamental measures of diagnostic examination performance: usefulness for clinical decision making and research. *Radiology* **228**, 3–9

Lord, P. & Suter, P. (1999) Radiology. In: Textbook of Canine and Feline Cardiology. Eds P. R. Fox, D. Sisson and N. S. Moise. W. B. Saunders, Philadelphia, PA, USA. pp 107–129

MacDonald, K. A., Kittleson, M. D., Munro, C. & Kass, P. (2003) Brain natriuretic peptide concentration in dogs with heart disease and congestive heart failure. *Journal of Veterinary Internal Medicine* **17**, 172–177

Manninen, H., Remes, J., Partanen, K., Tynkkynen, P., Mykkanen, L., Laakso, M., Soimakallio, S. & Pyorala, K. (1991) Evaluation of heart size and pulmonary vasculature. Conventional chest roentgenography and image intensifier photofluorography compared. *Acta Radiologica* **32**, 226–231

Metz, C. E. (1978) Basic principles of ROC analysis. *Seminars in Nuclear Medicine* **8**, 283–298

Metz, C. E. (2006) Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems. *Journal of American College of Radiology* **3**, 413–422

Monnier-Cholley, L., Carrat, F., Cholley, B. P., Tubiana, J. M. & Arrive, L. (2004) Detection of lung cancer on radiographs: receiver operating characteristic analyses of radiologists', pulmonologists', and anesthesiologists' performance. *Radiology* **233**, 799–805

Norgaard, H., Gjorup, T., Brems-Dalgaard, E., Hartelius, H. & Brun, B. (1990) Interobserver variation in the detection of pulmonary venous hypertension in chest radiographs. *European Journal of Radiology* **11**, 203–206

Obuchowski, N. A. (2000) Sample size tables for receiver operating characteristic studies. *American Journal of Roentgenology* **175**, 603–608

Obuchowski, N. A. (2003) Receiver operating characteristic curves and their use in radiology. *Radiology* **229**, 3–8

Obuchowski, N. A. & Zepp, R. C. (1996) Simple steps for improving multiple-reader studies in radiology. *American Journal of Roentgenology* **166**, 517–521

Prosek, R., Sisson, D. D., Oyama, M. A. & Solter, P. F. (2007) Distinguishing cardiac and noncardiac dyspnea in 48 dogs using plasma atrial natriuretic factor, B-type natriuretic factor, endothelin, and cardiac troponin-I. *Journal of Veterinary Internal Medicine* **21**, 238–242

Quekel, L. G., Goei, R., Kessels, A. G. & van Engelshoven, J. M. (2001a) Detection of lung cancer on the chest radiograph: impact of previous films, clinical information, double reading, and dual reading. *Journal of Clinical Epidemiology* **54**, 1146–1150

Quekel, L. G., Kessels, A. G., Goei, R. & van Engelshoven, J. M. (2001b) Detection of lung cancer on the chest radiograph: a study on observer performance. *European Journal of Radiology* **39**, 111–116

Rockette, H. E., Gur, D., Cooperstein, L. A., Obuchowski, N. A., King, J. L., Fuhrman, C. R., Tabor, E. K. & Metz, C. E. (1990) Effect of two rating formats in multi-disease ROC study of chest images. *Investigative Radiology* **25**, 225–229

Rockette, H. E., King, J. L., Medina, J. L., Eisen, H. B., Brown, M. L. & Gur, D. (1995) Imaging systems evaluation: effect of subtle cases on the design and analysis of receiver operating characteristic studies. *American Journal of Roentgenology* **165**, 679–683

Ruehl, W. M. & Thrall, D. E. (1981) The effect of dorsal versus ventral recumbency on the radiographic appearance of the canine thorax. *Veterinary Radiology and Ultrasound* **22**, 10–17

Satou, G. M., Lacro, R. V., Chung, T., Gauvreau, K. & Jenkins, K. J. (2001) Heart size on chest x-ray as a predictor of cardiac enlargement by echocardiography in children. *Pediatric Cardiology* **22**, 218–222

Sisson, D., Kvart, C. & Darke, P. (1999) Acquired valvular heart disease in dogs and cats. In: Textbook of Canine and Feline Cardiology. Eds P. R. Fox, D. Sisson and N. S. Moise. W. B. Saunders, Philadelphia, PA, USA. pp 536–565

Snashall, P. D., Keyes, S. J., Morgan, B. M., McAnulty, R. J., Mitchell-Heggs, P. F., McLvor, J. M. & Howlett, K. A. (1981) The radiographic detection of acute pulmonary oedema. A comparison of radiographic appearances, densitometry and lung water in dogs. *British Journal of Radiology* **54**, 277–288

Taylor, S. A., Charman, S. C., Lefere, P., McFarland, E. G., Paulson, E. K., Yee, J., Aslam, R., Barlow, J. M., Gupta, A., Kim, D. H., Miller, C. M. & Halligan, S. (2008) CT colonography: investigation of the optimum reader paradigm by using computer-aided detection software. *Radiology* **246**, 463–471

Thompson, W. M., Kilani, R. K., Smith, B. B., Thomas, J., Jaffe, T. A., Delong, D. M. & Paulson, E. K. (2007) Accuracy of abdominal radiography in acute small-bowel obstruction: does reviewer experience matter? *American Journal of Roentgenology* **188**, W233–W238

Corrigendum after online publication

KH, JH, CK and PL declare no conflicts of interest.