DOI: 10.5433/1679-0359.2014v35n5p2513

# Interobserver agreement in interpretation of radiographic pulmonary changes in dogs in relation to radiology training

# Concordância interobservador na interpretação de alterações radiográficas pulmonares em cães e sua correlação com treinamento em radiologia

Tilde Rodrigues Froes[1*]; Allison L Zwingenberger[2]; Amy Sato[3]; Daniela Aparecida Ayres Garcia[4]; Andressa Cristina de Souza[4]; Raquel de Souza Lemos[4]; Wilfried Mai[5]

## Abstract

Interpretation of pulmonary radiographs is one of the most difficult aspects of radiology and interobserver variability is high. The aim of this study was to assess variations in interpretation of pulmonary pathology amongst Brazilian veterinarians with different levels of training and experience, using the interpretation by American board-certified radiologists as a reference. We identified areas where interpretation is particularly challenging. Sixty digital canine thoracic radiographic examinations were interpreted by four groups of three Brazilian observers, each group being defined by different levels of training and experience. The radiographic findings of the 4 groups of observers in the study were compared to a reference interpretation established from the findings of three ACVR board-certified radiologists. The degree of discrepancy for each list between each group and the reference interpretation was assessed according to a three-level scoring system: no discrepancy, minor discrepancy, or major discrepancy. Data was analyzed using a Kappa and Cochran-Mantel-Haenszel tests. Brazilian veterinarians with the most training and experience showed the least interobserver variation and best performance when compared to the reference interpretation, followed by those with practical training, but with little work experience in professional practice. The radiographic patterns that were associated with the highest interobserver variability were the vascular, unstructured interstitial and bronchial patterns. Interobserver major discrepancies occurred in all groups, but is more evident in groups with the least training (44.4%) and the general practitioners (26.7%) group. It can be concluded that training positively influences the accuracy of radiographic interpretation and is recommended to reduce erroneous diagnoses.
**Key words**: Training, interobserver agreement, thoracic radiographs, pulmonary patterns

## Resumo

Uma das maiores dificuldades na interpretação radiográfica em cães está nas alterações pulmonares sendo a variabilidade interobservador alta. O objetivo desse estudo é detectar as variações de interpretação entre radiologistas brasileiros em diferentes graus de treinamento e experiência, utilizando a interpretação de consenso feito por radiologistas americanos certificados pelo Colégio Americano de Radiologia

[1] Profª, Universidade Federal do Paraná, UFPR, Curitiba, PR, Brasil. E-mail: tilde@ufpr.br
[2] Profª, University of California, Davis, CA. E-mail: azwingen@ucdavis.edu
[3] Profª, Tufts University School of Veterinary Medicine, TUFTs, North Grafton, MA. E-mail: amy.sato@tufts.edu
[4] Discentes, Programa de Pós-graduação em Ciências Veterinárias, PPGCV UFPR, Curitiba, PR, Brasil. E-mail: daniaapag@gmail.com; andressact.vet@gmail.com; raquel_ufpr@gmail.com
[5] Prof., Universidade da Pennsylvania, PENNVet, Philadelphia, PA. E-mail: wmai@vet.upenn.edu
* Author for correspondence

Recebido para publicação 16/09/13 Aprovado em 01/04/14

Veterinária. Na tentativa de identificar os desafios e as particularidades dessa interpretação. Sessenta exames radiográficos digitais do tórax de cães foram interpretados por quatro grupos de observadores com diferentes graus de treinamento em leitura de exames radiográficos. O grau de discrepância entre as observadores foram comparados seguindo um escore com três subclassificações: sem discrepância, discrepância leve e maiores discrepâncias. Para análise dos dados os métodos estatísticos utilizados foram o Kappa e Cochran-Mantel-Haenszel. Os veterinários brasileiros com maior grau de treinamento e experiência foram o que apresentaram menores variações de interpretação quando comparado aos dados do consenso, seguidos pelos veterinários menor treinamento e por médicos veterinários práticos da clínica diária sem treinamento especializado em interpretação radiográfica. Os padrões radiográficos que foram associados ao alto grau de discordância foram em sequência: vascular, intersticial não estrutural e padrão bronquial. Discrepâncias subclassificadas como maiores ocorreram em todos os grupos, porém foram bem mais evidentes no grupo com menor grau/tempo de treinamento (44,4%) e práticos (26,7%). Conclui-se que o treinamento apresenta influência positiva na acurácia da interpretação radiográfica pulmonar e é recomendado para reduzir erros de diagnóstico.
**Palavras-chave:** Treinamento, concordância interobservador, radiografia torácica, padrões pulmonares

## Introduction

The thorax is a complex anatomical area and thoracic radiographs, in particular pulmonary structures, are some of the most challenging radiographic studies to interpret (AL ASERI, 2009; LAMB; DU; MAMTIS, 2007; THRALL, 2013a). However, thoracic radiographs are commonly performed in first opinion practice and their interpretation is regularly used to inform clinical decisions (AL ASERI, 2009; LAMB, 2007).

Observer interpretation is one of the potential causes of radiographic diagnostic errors. Variations in interpretation are often related to radiologists' experience and also depend upon the amount of clinical information available prior to radiographic interpretation. It is important to understand the causes of discordance in interpretation to improve diagnostic quality and therefore enhance patient management (TUDOR; FINLAY; TAUB, 1997).

Interobserver variations in veterinary diagnostic imaging have been reported in a number of studies (LAMB; DU; MANTIS, 2007; HAMMOND et al., 2008; DUKES-McEWAN; FRENCJ; COCORAN, 2002). However, there are few studies that emphasize the effects of differences in training (LAMB; DU; MANTIS, 2007; ZWINGENBERGER et al., 2011). A previous study reported that one of the most common mistakes in radiologic interpretation made by veterinary students is over-interpretation (LAMB; DU; MANTIS, 2007), and another showed that there are differences in the interpretations made by practitioners versus a board-certified veterinary radiologist (ZWINGENBERGER et al., 2011).

Organized and rigorous post-graduate training in veterinary radiology, followed by standardized certification examinations exist in the United States and Europe (ACRV, 2011) but lack in other parts of the world. In Brazil, few institutions (around 22 of 192 veterinary schools) offer residency programs in diagnostic imaging (CFMV, 2012). These residencies are undertaken over two years, focus only on radiology and ultrasonography and are not sanctioned by a standardized examination system. The demand for, and interest in, this discipline is rising rapidly in Brazil. Unfortunately, there are too few specialized training centers to address the demand, resulting in the development of controversial training programs, over only a few months or through short imaging courses. The aim of this study was to assess the effect of radiology training and experience in Brazil on interobserver variation in the recognition and interpretation of radiograph pulmonary patterns in dogs.

## Methods

This was a cross-sectional study conducted using teleradiology over 20 weeks, involving

interpretation of canine thoracic radiographs with focus on pulmonary disorders.

Twelve Brazilians observers were selected. They were divided into four groups of three individuals, corresponding to the degree of experience and training in veterinary diagnostic imaging (Table 1). The results of their interpretations were compared to a reference interpretation derived from the interpretation made by three academic radiologists certified by the American College of Veterinary Radiology (ACVR), with several years experience in clinical radiology (XU; MA; HE, 2012; STOUT et al., 2010). This research was approved by Ethics Committee CEUA-SCA from Federal University of Paraná.

**Table 1.** Classification of the four groups of observer according to the years of training and years experience.

| Group | Characteristics of the group | Years of training Median (range) | Years of experience Median (range) |
|---|---|---|---|
| Group 1 (n=3) | Experienced professionals, with at least one year of experience after completion their residency in Brazil | 3-5-7 | 6-9-12 |
| Group 2 (n=3) | Residents in radiology in Brazil | 1-2-3 | 1-2-3 |
| Group 3 (n=3) | Practitioners without supervised practical training in radiology | 0 (N/A) | 3-4-13 (3 –13) |
| Group 4 (n=3) | Veterinary students | 0.5 | 0.2-0.3-0.6 |

* Six months of radiology training, once a week.
**Source:** Elaboration of the authors.

Sixty digital thoracic radiographic examinations with (n=54) and without (n=6) pulmonary pathology were selected. The radiographs were obtained from three university veterinary teaching hospital in North America (University of California-Davis, Tufts University, and University of Pennsylvania) and the variety of cases selected represented classic examples of common conditions encountered in general practice. For each case at least a dorsoventral (n=2) / ventrodorsal (n= 58) projection and a lateral projection were available. In 30 cases, two opposite lateral projections were provided in addition to a dorsoventral or ventrodorsal view (3-view protocol). Radiographs were selected to be of optimal diagnostic quality in terms of exposure and positioning, and were available for review by observers in a digital PDF format obtained from original DICOM images that had been optimally windowed and leveled by the participating ACVR board-certified radiologists. Each set of radiographs was coded from 1 to 60 and then up-loaded onto a virtual platform. Each observer was asked to interpret three sets of radiographs per week over the course of the study until completion of the 60 evaluations. They had no knowledge of the clinical history, signalment or final diagnosis.

Interpretation of the radiographs was guided by a review form on which observers were instructed to follow a systematic sequence of reading, which included: thoracic skeletal structures, intra-thoracic/extra-pulmonary structures (pleural space, cranial, middle [including cardiac silhouette] and caudal mediastinal space, diaphragm, trachea and esophagus), and radiographic alterations of the pulmonary parenchyma. For assessment of the lungs, reviewers were asked to indicate location of changes, and categorize the pulmonary patterns according to previously published radiographic

classifications (SUTER; LORD, 1984). The observers were also asked to provide a prioritized list of differential diagnoses for these changes, only based on radiographic appearance since no clinical information was provided at the time of evaluation.

The interpretations made by the board-certified radiologists were compared and used to create a reference interpretation and diagnosis, used for statistical analysis (KUNDEL; POLANSKY, 2003) Agreement between at least two of the three radiologists was required to generate the reference for each case.

Interpretations of each observer (groups 1 through 4) were then compared to the reference standard. For statistical analysis the interpretation of the observers was divided in two parts, the first regarding pulmonary pattern recognition and the second for the differential diagnosis**.**

Pulmonary pattern recognition was considered *'in agreement'* when the observer correctly identified all the pulmonary lesions (eg '*increased opacity in the right cranial lung lobe'*) and classified correctly all the pulmonary pattern(s) present (eg '*lobar alveolar pattern in the right cranial lung lobe*'). Interpretations were considered *'in disagreement'* when either the pulmonary lesions or the pulmonary pattern(s) were not correctly recognized.

To assess the potential clinical impact of the differences in differential diagnoses between the observers and the reference diagnosis established from the board certified radiologists' interpretation, a scoring system was generated to grade the levels of discrepancy between observers (ABUJUDEH et al., 2010). The scoring system, definitions and examples of the levels of discrepancy are provided in Table 2.

**Table 2.** Level of discrepancy scoring system[14] used to assess differences in differential diagnoses between observers and reference differential diagnoses derived from interpretation by three board-certified radiologists.

| SCORE | DESCRIPTION | DEFINITION | EXAMPLES |
|---|---|---|---|
| **1** | No disagreement in interpretation | No discrepancies found | No discrepancies found |
| **2** | Minor disagreement with no clinical significance | Minimal differences in diagnostic list | Bronchopneumonia vs pneumonia<br>Normal vs age-related interstitial fibrosis |
| **3** | Major disagreement with clinical significance and potential to alter patient treatment plan | Failure to include important differentials and/or include erroneous differentials that could change the clinical recommendations for patients | Absence vs presence of pulmonary metastases.<br>Absence vs presence of pulmonary edema<br>Absence vs presence of pneumonia<br>Pulmonary osteomas vs pulmonary metastases |

**Source:** Elaboration of the authors.

The Cohen's Kappa statistical test was used to measure agreement between observers in the different groups and the reference standard derived from the interpretation made by three board-certified radiologists .This was done first considering all pulmonary patterns together and then breaking down the patterns categories. Kappa values below 0.20 were considered as poor agreement, 0.21 to 0.40 reasonable, 0.41 to 0.60 moderate, 0.61 to 0.80 good and > 0.81 as excellent. These calculation were performed with commercially available software (MedCalc Version 12.4.0., Ostend, Belgium). The Cochran-Mantel-Haenszel test (a variation of ANOVA used for categorical data) was used to see if there were statistically significant differences in discrepancy scores between the four groups of observers, and was performed using commercially available software (SAS Version 9.0, SAS Institute, Cary, NC). A P value of less than 0.05 was considered to indicate a statistically significant difference. All statistical analyses were performed by Professor Laila Talarico Dias.

## Results

Among the 60 thoracic radiographic studies, there was agreement between all three board-certified radiologists in 48/60 cases. In 12/60 exams, one of the three radiologists had some sort of disagreement with the other two. In 5/12 cases the disagreement was about pattern definition, in 1/12 cases there was a difference in differential diagnosis, and in 6/12 the disagreement involved both the pattern recognition and differential diagnosis.

The reference standard, defined from agreement between at least two of the three radiologists, determined that two studies (3.3%) showed no radiographic abnormalities, and four studies (6.6%) contained some radiographic changes that were considered to be of no clinical significance (pulmonary osteomas [two cases]), age- or technique-related subtle interstitial pattern [one case] and age-related broncho-interstitial pattern [one case]). The remaining 54 (90%) examinations were considered abnormal. From the 54 examinations classified as having clinically relevant radiographic abnormalities, 53 (98,2%) had increased pulmonary opacity and one showed a reduction in pulmonary opacity (Table 3).

**Table 3.** Distribution of predominant radiographic pulmonary abnormalities in the reference interpretations derived from assessment by ACVR-certified radiologists (Using Standard Nomenclature For Classification).

| **Radiopacity Changes** | **Number of interpretations** |
|---|---|
| Decrease of Radiopacity | 1 (1.8%)* |
| Increase of Radiopacity | 53 (98.2%) |
| **Predominant Pulmonary patterns identified (increase of radiopacity)** | |
| *Alveolar* | 12 (22.7%) |
| *Bronchial* | 16 (30.2%) |
| *Interstitial unstructured* | 14 (26.5%) |
| *Interstitial nodular/masses* | 9 (16.9%) |
| *Vascular* | 2 (3.7%) |

*bulla focal lesion.
**Source:** Elaboration of the authors.

Of the 54 significantly abnormal exams, 21 had more than one pulmonary radiographic pattern (mixed opacities), while in 33 only one pulmonary pattern was listed. The distribution of the predominant pulmonary patterns in all the abnormal cases is listed in Table 3.

The most common differential diagnoses as per the reference standard were: pulmonary neoplasia (26 cases), bacterial pneumonia (13 cases), bronchitis (11 cases) and fungal pneumonia (10 cases). A summary of all differential diagnoses listed in the reference standard is shown in Table 4.

Interobserver agreement for the recognition of the pulmonary patterns in each group of reviewers is summarized in Table 5. Overall we observed that higher level of training and experience resulted in higher agreement in the classification of pulmonary pattern.

**Table 4.** Absolute number and relative percentage of the differential diagnoses as per the reference interpretation by three board certified radiologists.

| Differential diagnosis | Occurence | Differential diagnosis | Occurence |
|---|---|---|---|
| Abscess | 1 (1.7%) | Fungal pneumonia | 10 (16.7%) |
| Age related changes | 2 (3.3%) | Granuloma | 2 (3.3%) |
| Atypical pneumonia | 2 (3.3%) | Heartworn disease | 2 (3.3%) |
| Bacterial pneumonia | 13 (21.7%) | Hemorrhage | 4 (6.7%) |
| Bronchial Plug | 3 (5%) | Interstitial pneumonia | 3 (5.0%) |
| Bronchitis | 11 (18.3%) | Neoplasia | 26 (43.3%) |
| Eosinophilic infiltrate | 4 (6.7%) | No radiographic changes | 6 (10%) |
| Bronchopneumonia | 8 (13.3%) | Pulmonary osteoma | 4 (6.7%) |
| Cavitary lesions | 1 (1.7%) | Parasitic bronchitis | 2 (3.3%) |
| Cardiogenic edema | 1 (1.7%) | Persistent Ductus Arteriosus | 1 (1.7%) |
| Pulmonary fibrosis | 4 (6.7%) | Thrombo-embolism | 3 (5%) |

Note: the number of occurrence refers to the number of examinations in which this differential diagnosis was listed – examinations may have had more than one possible diagnosis hence the total number of diagnoses is higher than the total number of examinations (n=60).
**Source:** Elaboration of the authors.

**Table 5.** Agreement between observers and reference interpretation derived from assessment by three AVCR Board-Certified Radiologists.

| | Group 1 | | Group 2 | | Group 3 | | Group 4 | |
|---|---|---|---|---|---|---|---|---|
| **Pattern** | ***K*** | ***Agreement*** | ***K*** | ***Agreement*** | ***K*** | ***Agreement*** | ***K*** | ***Agreement*** |
| All patterns | 0.62 | Good | 0.59 | Moderate | 0.43 | Moderate | 0.27 | Fair |
| Alveolar | 0.70 | Good | 0.75 | Good | 0.61 | Good | 0.27 | Fair |
| Bronchial | 0.59 | Moderate | 0.49 | Moderate | 0.28 | Fair | 0.16 | Poor |
| Unstructured interstitial | 0.49 | Moderate | 0.48 | Moderate | 0.41 | Moderate | 0.30 | Fair |
| Interstitial nodular / mass | 0.91 | Excellent | 0.85 | Excellent | 0.58 | Moderate | 0.32 | Fair |
| Vascular | 0.27 | Fair | 0.27 | Fair | 0.27 | Fair | 0.07 | Poor |

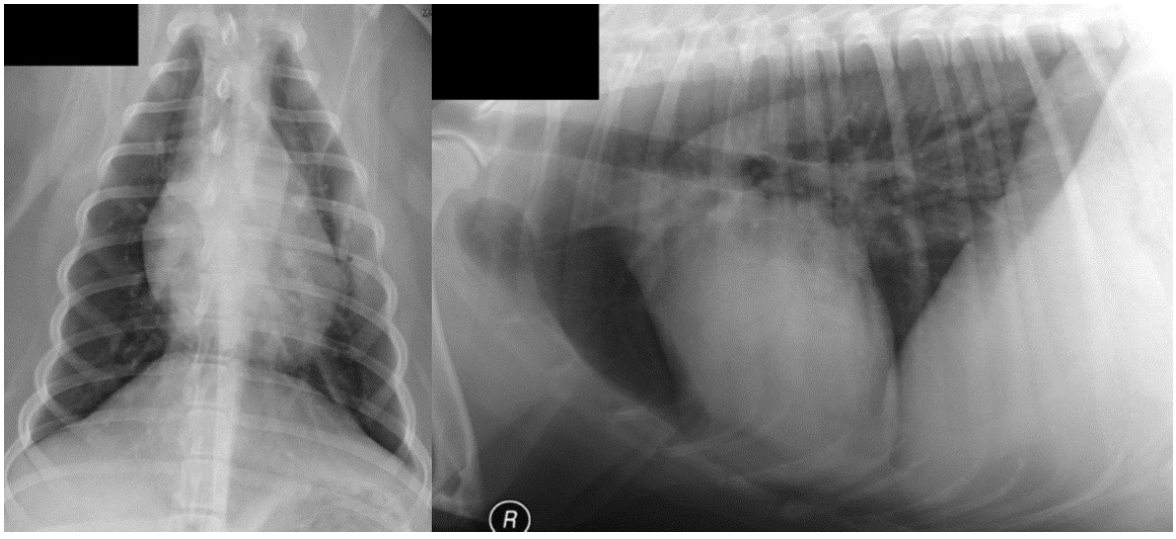
**Source:** Elaboration of the authors.

In this study, the pattern with highest agreement with the reference standard was 'nodular interstitial pattern/mass lesion'. However, even for this easily recognizable pattern, individuals with less training (Groups 3 and 4) showed only fair and poor agreement.

With the exception of Group 4 (students without training or practice) there was generally good agreement for recognition of the alveolar pulmonary pattern. That being said, the case associated with the highest number of erroneous identification across all readers had a faint alveolar pattern superimposed to the cardiac silhouette on the lateral view (Figure 1).

The bronchial, unstructured interstitial and vascular patterns appeared to be more challenging with again lower levels of training associated with higher degrees of mistakes. In particular the vascular pattern (two cases of arterial and two cases of venous enlargement, with diagnoses including persistent ductus arteriosus, mitral insufficiency and heartworm disease) produced the highest degree of disagreement with the reference standard, regardless of the degree of training and experience of the observers. All four cases had some degree of cardiomegaly that was reported by the observers.

**Figure 1.** Thoracic radiographic examination, classified by the ACVR-certified radiologists as an alveolar pattern with most likely differential diagnoses of pneumonia. The alveolar pattern is seen superimposed to the cardiac silhouette ventrally and was overlooked by most of the Brazilian observers.




**Source:** Elaboration of the authors.

Regarding differential diagnoses, we observed that the most trained/experienced Brazilian practitioners (Group 1) made fewer mistakes than less trained practitioners and veterinary students (Group 3 and 4) (Table 6). However even within group 1, there was disagreement with the differentials listed by the board-certified radiologist that could have potential impact on patient management in about 10% of cases (discrepancy score of 3). The percentage of such disagreement increased across groups 2 through 4 (15.0, 26.7, 44.4% respectively).

**Table 6.** Discrepancy scores for differential diagnosis in groups 1-4 after comparative analysis with reference differential diagnoses derived from interpretation by three board-certified radiologists.

| SCORE | DESCRIPTION | GROUP 1 (%) | GROUP 2 (%) | GROUP 3 (%) | GROUP 4 (%) |
|---|---|---|---|---|---|
| 1 | No disagreement in interpretation | 145/180 (80.5%) | 133/180 (73.9%) | 103/180 (57.2%) | 78/180 (43.3%) |
| 2 | Minor disagreement with no clinical significance | 17/180 (9.5%) | 20/180 (11.1%) | 29/180 (16.1%) | 22/180 (12.3%) |
| 3 | Major disagreement with clinical significance and potential to change patient treatment plan | 18/180 (10.0%) | 27/180 (15.0%) | 48/180 (26.7%) | 80/180 (44.4%) |

Note – there were 180 interpretations in each group (60 examinations for 3 different readers).
**Source:** Elaboration of the authors.

In the two cases interpreted by the board-certified radiologists as incidental pulmonary osteomas, one observer out of three in each of groups 1 and 2 listed pulmonary metastasis as a possible differential. Within group 3, two of the three observers suggested pulmonary metastases as a differential diagnosis in one of the two cases, and in group 4, all three observers suggested pulmonary metastases as a differential diagnosis for both examinations – two of these suggesting pulmonary metastases as the sole differential diagnosis.

Out of the five examinations in which most observers made errors in differential diagnosis, three had an unstructured interstitial pattern (with differential diagnoses of lymphoma, solid tumor metastasis or fungal pneumonia in two of them, and of fibrosis, lymphoma or interstitial pneumonia in the other), and two had an alveolar pattern (one with a reference diagnosis of pneumonia and the other of non-cardiogenic edema). In the dog with pneumonia, only one of the 12 Brazilian observers identified the pattern and the differential diagnosis correctly (Figure 1).

The Cochran-Mantel-Haeszel test indicated a significant difference in discrepancy scores between groups ($P<0.0001$) and a strong association between level of training and discrepancy scores ($P<0.0001$) (Table 6).

## Discussion

Radiography is routinely used in the investigation of thoracic diseases in dogs (SUTER; LORD, 1984). However, as this study demonstrates, interpretation of canine thoracic radiographs can be challenging, especially for individuals not trained in radiology and even for individuals with more training and experience. These findings mirror those in similar studies in human medicine (AL ASERI, 2009).

A possible explanation for the difficulty experienced in interpretation of thoracic radiographs is the wide variety of radiographic changes that may be found in the same disease, or the fact that a single radiographic pattern can be associated with a number of diseases (THRALL, 2013b). As recently discussed in the veterinary literature, the traditional radiographic pulmonary patterns may not represent the best way to classify pulmonary diseases (SCRIVANI, 2009), and it is no longer considered reliable or accurate in human medicine (REED, 1997).

Despite the high level of training and experience, board-certified radiologists still disagreed over some cases. Therefore, it is possible that some discrepancies will be associated to simple human error related to cognitive bias (GUNDERMAN, 2009), different emotions or physical tiredness during the interpretation (TUDOR; FINLAY; TAUB, 1997; HERMAN et al., 1975; JOHNSON; KLINE, 2010). This can happen to any observers independent of the degree of training and experience.

Not surprisingly, our study showed that the longer the training period and duration of active experience in reading radiographs, the better the agreement with the reference standard. However, it also showed that even Brazilian veterinarians with extensive experience in radiology did not have a perfect agreement with the board-certified radiologists and that even in that group, major discrepancies (score 3) could exist in terms of differential diagnoses. This places emphasis on the importance to increase the standards of specialization in Brazil through formal training programs and validation of training through a rigorous examination system, perhaps using models that exist elsewhere such as in North America or Europe.

When considering all pulmonary patterns together, in groups 1 and 2 (Brazilian veterinarians with most experience/expertise), there was good agreement for pattern recognition. However even in these groups, interobserver agreement was only moderate for bronchial and unstructured interstitial patterns, and fair for the vascular pattern. Groups with little experience and training in radiology (group 3 and 4) most often showed poor interobserver agreement in pattern recognition and major discrepancies in differential diagnoses that could have potential negative impact on case management and treatment outcome.

Of course, variability in interpretation can also be influenced by other factors than training or experience, such as radiographic quality, and individual radiologist variability (TUDOR; FINLAY; TAUB, 1997; HERMAN et al., 1975; JOHNSON; KLINE, 2010). Even amongst experienced and trained radiologists, there is variation in interpretation and diagnostic performance, as we observed and as has been reported (BREALEY; SCALLY; THOMAS, 2002). Furthermore, although variability serves as an indirect measurement of error, it can also represent a genuine difference in opinion (FITZGERALD, 2001).

Although the Nodular pattern overall showed the best agreement with the reference standard in group 1 and 2, there was only moderate to fair interobserver agreement in groups 3 and 4. Failure to detect metastases may be due to a number of radiological errors: lack of systematic radiographic reading, lack of knowledge of all possible radiographic appearances of metastasis, or early closure due to the identification of other concurrent diseases on the radiographs (THRALL, 2013b). In addition, even when correctly identifying a nodular pattern, false positive diagnoses of metastases were observed in groups 3 and 4 where pulmonary osteomas were mistaken for metastatic nodules. This is probably due to lack of knowledge of the typical appearance of these lesions (small (2-4mm), very opaque and conspicuous nodules) (SCHWARZ; JOHNSON, 2008). Although some errors in pulmonary radiographic interpretation can be compensated for by the attending clinician based on clinical presentation and results of other tests, this is usually not the case with pulmonary metastases, in which radiographic identification guides further therapeutic measures, such as the choice between chemotherapy or surgical removal of the tumor, or decision for euthanasia (HEDLUND, 2007). Such errors may significantly impact the patient treatment plan.

Overall the second best interobserver agreement was seen for alveolar pattern. This is likely because its radiographic appearance is often quite characteristic and well described, including patchy diffuse opacification (‘clouds’), air bronchograms, lobar sign, and silhouette sign (THRALL, 2013a;

SUTER; LORD, 1984). However, even for the alveolar pattern, k value in Group 4 was only 0.27 (fair agreement) underlying the importance of training individuals in reading thoracic radiographs (JEFFREY et al., 2003). Although air bronchograms and lobar signs are common indicators of an alveolar pattern (THRALL, 2013a; LAMB, 2007), sometimes neither will be seen. Air bronchograms may not be seen if alveolar disease in not concentrated adequately around a bronchus for the bronchial lumen become visible. A lobar sign will not be seen if the alveolar disease does not extend to the periphery of the lobe, if adjoining lobes are both affected the same extent, or if lobe junction is not struck parallel by the x-ray beam (THRALL, 2013a) For the inexperienced readers such as those in group 4, the lack of classic Roentgen signs may make the recognition of an alveolar pattern more challenging; this emphasizes the importance of extensive training in high-caseload environment and appropriate mentoring to learn all possible presentations of various pulmonary patterns and conditions. It is interesting to note that the study with most errors was an examination showing an alveolar pattern (suspicious for pneumonia), in which only one of the 12 Brazilians observers identified the pattern and the differential diagnosis correctly. This is surprising, since there was generally good agreement in identification of alveolar disease in all groups. In this case, the errors may be due to the fact that air bronchograms were only faintly visible superimposed to the cardiac silhouette on the lateral view and subtle changes may be missed if systematic assessment is not performed (Figure 1).

Results of the present study showed that unstructured interstitial and bronchial pattern are not easy to identify, even by experienced individuals. The diagnostic importance to accurately differentiate the classic lung patterns has been challenged in a recent review, where the author exposed the fragility of radiographic patterns to represent the actual distribution of histological lesions, and proposed a new approach to radiographic characterization of pulmonary lesions. Many diseases are associated with multiple simultaneous microscopic distributions even when a predominant radiographic pattern (e.g. bronchial) is identified. Conversely, many microscopic distributions can lead to a similar radiographic pattern (for example, increased thickness of the interstitium, partial filling of the alveoli with fluid or cells or partial collapse of the alveoli can all result in a similar diffuse interstitial pattern radiographically). For these reasons, it was suggested that more generic terms are preferable (SCRIVANI, 2009). It is possible that establishing new schemes to teach radiographic interpretation of pulmonary changes would improve overall performance in recognition and interpretation of abnormalities.

It is notable that the vascular pattern was the one with greater disagreement among Brazilian radiologists, regardless of training and experience. In all these cases, cardiomegaly was present and recognized by the observers, which should have clued them in that some vascular condition may be present, yet they failed to recognize vascular enlargement in these cases. The conditions represented in this group (mitral insufficiency, persistent ductus arteriosus and heartworm disease) are all present in the country and therefore not a specific diagnostic challenge. This may result from a generally insufficient training in recognition of this particular Roentgen sign in Brazil.

For this study we combined the radiographic findings of three ACVR board-certified radiologists with several years of experience to produce a 'standard interpretation', a method previously used for interobserver radiographic studies (REED, 1997; HERMAN et al., 1975). ACVR-certified veterinary radiologists were chosen for this purpose due to their extensive and rigorous training validated by examination by a College of specialists. It could be argued that there is a difference between our reference standard and a true gold standard such as histopathological evaluation of the lung. However radiographic interpretation of pulmonary

pattern does not necessarily correlate with the histopathological extent and distribution of disease in the lung, and therefore even histopathology may not be a good gold standard (SCRIVANI, 2009). In addition even with histopathology it would be difficult to accurately match the histopathological changes to focal radiographic abnormalities in the lung, specially when lesions are multifocal and with mixed patterns. A complete consensus between all three radiologists was achieved in 48 of the 60 cases, and agreement between two of the three radiologists was obtained in the remaining 12 cases. Therefore it could be assumed that our reference standard is reasonably solid for the purpose of our study, which mostly aimed at evaluating the interpretation performance of Brazilian veterinarians with various levels of training and experience, as opposed to assess the accuracy of their radiographic diagnosis as compared to actual disease condition.

Despite the relatively low number of radiographic examinations in this study, the majority of pulmonary diseases that affect dogs were included in the reference diagnosis established by the board-certified radiologists. However some of these diseases are less common in Brazil and this could account, to some extent, for some of the discrepancy between differential diagnoses established by Brazilian veterinarians versus the reference diagnoses. For example, the incidence of fungal pneumonia in Brazil is low, and this condition is not thought of as often as it is in the USA. This is in agreement with the findings described in human studies (JEFFREY et al., 2003). Although geographical differences in disease prevalence may explain in part the discrepancies in differential diagnoses, our study design cannot accurately assess this, since no American veterinarians with similar degrees of experience and training as the Brazilian groups were included in the analysis. Even if American veterinarian would be included in the study, there would still be some bias due to the fact that radiology education in veterinary schools in the USA and in Brazil are likely different with regard to many other factors than only disease prevalence.

Another limitation of our study is the fact that the observers had no access to historical data or clinical findings. If they had access to these data, lesion identification and interpretation of the radiographic changes may have improved (ALEXANDER, 2010). Studies in human medicine found that providing the clinical history and additional information may improve an individual's ability to detect an abnormality although not at a statistically significant level. However, occasionally some clinical details can lead the observer to make a false positive diagnoses (TUDOR; FINLAY; TAUB, 1997). In any case, the results of our study may not be generalized to all observers in a clinical setting, since the history is usually available in such setting.

Another issue that may have interfered with the reading of radiographs is the quality of monitors used, since this was not standardized in the study and high-resolution monitors were not available to all readers. However all reviewers, including the board-certified radiologists used consumer-grade monitors for their assessment. The lack of familiarity in reading and interpreting digital radiographic images could also have resulted in poor film interpretation (ZWINGENBERGER et al., 2011), since analog radiography is still common in Brazil and Brazilian radiologists have little to no experience with digital radiography.

Other limitations of our study include the lack of consistency across the 60 thoracic radiographic examinations. For example not all radiographs were obtained in the VD projection. However only two examinations were obtained with DV rather than VD so this variation is considered unlikely to be significant. Another variation is the number of radiographic projections available. Half of the cases had a two-view protocol and the other half a three-view protocol. One may argue that this could have affected the interpretation of the observers, in particular for the least experienced group. We did not specifically analyze this factor, as our main goal

was to evaluate overall interpretation in a sample of cases quite representative of routine, real-life caseload.

We found a statistically significant association between the degree of training and discrepancy scores in radiographic interpretation. The incidence of major discrepancy was high in the groups with lower level of training and experience, and this again emphasizes the importance of more rigorous and structured training, as such errors have high potential impact on the medical or surgical management of these animals.

Our study emphasizes the potential benefit for practitioners and even more specialized veterinarians in Brazil to seek opinion from Board Certified radiologist in selected cases, and also highlights the need to raise the standards for specialty training in Brazil. In addition, to improve students' radiographic skills, educational efforts should be made inside the veterinary schools to teach them how to differentiate between significant and non-significant radiographic findings (LAMB; DU; MANTIS, 2007), besides elaborating a comprehensive list of differential diagnoses for all types of radiographic patterns.

Similar to the findings of previous studies in human medicine (QUEKEL et al., 2001a), this study demonstrated that not only experience but also formal specific training in radiology both improve the ability of an observer to detect and interpret radiographic pulmonary changes, reduce erroneous diagnosis rate and are likely to improve the relevance of differential diagnoses lists. Thus, practical training in reading thoracic radiographs in veterinary medicine is essential. Although training and experience improve interobserver agreement, some variability will still exist, primarily due to individual variations. One option for dealing with this would be double reading of films in veterinary diagnostic centers, to improve sensitivity and specificity, and this has been recommended in some veterinary studies (QUEKEL et al., 2001b). To raise the standards of veterinary radiology in Brazil, a more frequent use of teleradiology to consult with board-certified radiologists in other countries may also be recommended, together with change the structure and raise the standards of the residency training programs offered in Brazil (ALEXANDER et al., 2012).

The degree and type of training of observers influences their interpretation of pulmonary radiographs. Specific training in radiology is therefore essential, in order to reduce incorrect radiographic diagnoses. Structured residency programs are necessary to increase the overall quality of radiographic interpretation by Brazilian specialists.

## Acknowledgments

We thank the following veterinarians and students who participated in interpretation of the thoracic examinations for this study: Gabriela Silva Rodrigues, Rosana Zanatta, Danielle Buch, Eduardo Ayres, Natascha Kellermann Brauer, Simone Cristina Monteiro, André Obladen, Cintia Duquesne, Mayron Tobias da Luz, Juliana Kravetz de Oliveira. And Professor Laila Talarico Dias for statistical Analysis.

## References

ABUJUDEH, H. H.; BOLAND, G. W.; KAEWLAI, R.; RABINER, P.; HALPNER, E. F.; GAZELLE, G. S.; THRALL, J. H. Abdominal and pelvic computer tomography (CT) interpretation: discrepancy rates among experienced radiologists. *European Radiology,* Berlin, v. 20, n. 8, p. 1952-1957, 2010.

ALASERI, Z. Accuracy of chest radiograph interpretation by emergency physicians. *Emergency Radiology,* New York, v. 16, n. 2, p. 111-114, 2009.

ALEXANDER, K.; JOLY, H.; BLOND, L.; D'ANJOU, M.; NADEAU, M. E.; OLIVE, J.; BEAUCHAMP, G. A Comparison of computed tomography, computed radiography, and film-screen radiography for the detection of canine pulmonary nodules. *Veterinary*

*Radiology & Ultrasound*, Oxford, v. 53, n. 3, p. 258-265, 2012.

ALEXANDER, K. Reducing error in radiographic interpretation. *Canadian Veterinary Journal,* Ottawa, v. 51, n. 5, p. 533-536, 2010.

AMERICAN COLLEGE OF VETERINARY RADIOLOGY – ACVR. Membership categories, american college of veterinary radiology. Raleigh, 2011. Available at: <http://www.acvr.org/node/veterinary-professionals/about-acvr/membership-categories>. Accessed at: 1 mar. 2013.

BREALEY, S.; SCALLY, A. J.; THOMAS, N. B. Methodological standards in radiographer plain film reading performance studies. *British Journal of Radiology*, London, v. 75, n. 890, p. 107-113, 2002.

CONSELHO FEDERAL DE MEDICINA VETERINÁRIA – CFMV. Brasil: residência em medicina veterinária, conselho federal de medicina veterinária. Brasilia, 2012. Available at: <http://www.cfmv.gov.br/portal/pagina.php?cod=11>. Accessed at: 1 mar. 2013.

DUKES-McEWAN, J.; FRENCH, A. T.; CORCORAN, B. M. Doppler echocardiography in the dog: Measurement variability and reproducibility. *Veterinary Radiology & Ultrasound,* Oxford, v. 43, n. 2, p. 144-152, 2002.

FITZGERALD, R. Error in radiology. *Clinical Radiology*, Oxford, v. 56, n. 12, p. 938-946, 2001.

GUNDERMAN, R. B. Biases in radiologic reasoing. *American Journal Roetegnology*, Springfield, v. 192, n. 3, p. 561-564, 2009.

HAMMOND, G.; GEMMILL, T.; MELLOR, D.; SULLIVAN, M. Assessment of low-cost teleradiology for grading elbow dysplasia. *Veterinary Radiology & Ultrasound*, Oxford, v. 49, n. 1, p. 20-25, 2008.

HEDLUND, C. Mammary neoplasia. In: FOSSUM, T. *Small animal surgery.* 2. ed. St Louis: Mosby Elsevier, 2007. p. 729-735.

HERMAN, P. G.; GERSON, D. E.; HESSEL, J. S.; WATNICK, M.; BLESSER, B.; OZONOFF, D. Disagreements in chest roentgen interpretation. *Chest.,* Northbrook , v. 68, n. 3, p. 278-282, 1975.

JEFFREY, D. R.; GODDARD, P. R.; CALLAWAY, M. P.; GREENWOOD, R. Chest radiograph interpretation by medical students. *Clinical Radiology*, Oxford,v. 58, n. 6, p. 478-481, 2003.

JOHNSON, J.; KLINE, J. A. Intraobserver and interobserver agreement of the interpretation of pediatric chest radiographs. *Emergency Radiology*, New York, v. 17, n. 4, p. 285-290, 2010.

KUNDEL, H. L.; POLANSKY, M. Measurement of observer agreement. *Radiology*, New York, v. 228, n. 2, p. 303-308, 2003.

LAMB, C. R. The canine and feline lung. In: THRALL, D. E. *Textbook of veterinary diagnostic radiology.* 4. ed. St Louis, Missouri: Elsevier, 2007. p. 591-608.

LAMB, C.; DU, P.; MANTIS, P. Errors in radiographic interpretation made by veterinary students. *Journal Veterinary Medical Education*, Toronto, v. 34, n. 2, p. 157-159, 2007.

QUEKEL, L.; GOIE, R.; KESSELS, A.; VAN ENGELSHOVEN, J. Detection of lung cancer on the chest radiograph: impact of previous films, clinical information,double reading,and dual reading. *Journal Clinical Epidemiology*, New York, v. 54, n. 11, p. 1450-1460, 2001a.

QUEKEL, L.; KESSELS, A.; GOIE, R.; VAN ENGELSHOVEN, J. Detection of lung cancer on the chest radiograph: a study on observer performance. *European Journal Radiology*, Stuttgart, v. 39, n. 2, p. 111-116, 2001b.

REED, J. C. *Chest radiology:* plain film patterns and differential diagnoses. In: _____. *Chest radiology.* St Louis (MO): Mosby-year Book, 1997. p. 185-432.

SCHWARZ, T.; JOHNSON, V. *BSAVA manual of canine and feline thoracic imaging.* New Delhi: Replika Press Pvt., 2008. 395 p.

SCRIVANI, P. V. Non traditional interpretation of lung patterns. *Veterinary Clinics North American: Small Animal Practice*, Philadelphia, v. 39, n. 4, p. 719-732, 2009.

STOUT, J. E.; KOSINSKI, A. S.; HAMILTON, C. D.; GOODMAN, P. C.; MOSHER, A.; MENZIES, D.; SHLUGER, N.; KHAN, A.; JOHNSON, J. L. Effect of improving the quality of radiographic interpretation on the ability to predict pulmonary tuberculosis relapse. *Academic Radiology*, Reston, v. 17, n. 2, p. 157-162, 2010.

SUTER, P.; LORD, P. Normal radiographic anatomy and radiographic examination. In: SUTER, P. *Text atlas of thoracic radiography*: thoracic diseases of the dog and cat. Wettswil: Switzerland, 1984. p. 1-45.

THRALL, D. E. The canine and feline lung. In: _____. *Textbook of veterinary diagnostic radiology.* 5. ed. St Louis, Missouri: Elsevier Saunders, 2013a. p. 608-631.

_____. Principles of radiographic interpretation of the thorax. In: _____. *Textbook of veterinary diagnosotic radiology.* 5. ed. St Louis, Missouri: Elsevier Saunders, 2013b. p. 474-487.

TUDOR, G. R.; FINLAY, D.; TAUB, N. An assesment of interobserver agreement and accuracy when reporting plain radiographs. *Clinical Radiology*, Oxford, v. 52, n. 3, p. 235-238, 1997.

XU, Y.; MA, D.; HE, W. Assessing the use of digital radiography and a real-time interactive pulmonary nodule analysis system for large population lung cancer screening. *European Journal Radiology,* Stuttgart, v. 81, n. 4, p. 451-456, 2012.

ZWINGENBERGER, A.; BOUMA, J.; SAUNDERS, H.; NODINE, C. Expert interpretation compensates for reduced image quality of camera-digitized images referredto radiologists. *Veterinary Radiology & Ultrasound,* Oxford, v. 52, n. 6, p. 591-595, 2011.