ORIGINAL ARTICLE

Diagnostic validation of vertebral heart score machine learning algorithm for canine lateral chest radiographs

J. Solomon, S. Bender, P. Durgempudi, C. Robar, M. Cocchiaro, S. Turner, C. Watson, J. Healy, A. Spake and D. Szlosek

IDEXX Laboratories, Inc., Westbrook, ME, USA

[1]Corresponding author email: donald-szlosek@idexx.com

Objectives: The vertebral heart score is a measurement used to index heart size relative to thoracic vertebra. Vertebral heart score can be a useful tool for identifying and staging heart disease and providing prognostic information. The purpose of this study is to validate the use of a vertebral heart score algorithm compared to manual vertebral heart scoring by three board-certified veterinary cardiologists.

Materials and Methods: A convolutional neural network centred around semantic segmentation of relevant anatomical features was developed to predict heart size and vertebral bodies. These predictions were used to calculate the vertebral heart score. An external validation study consisting of 1200 canine lateral radiographs was randomly selected to match the underlying distribution of vertebral heart scores. Three American College of Veterinary Internal Medicine board-certified cardiologists were enrolled to manually score 400 images each using the traditional Buchanan method. Post-scoring, the cardiologists evaluated the algorithm for misaligned anatomic landmarks and overall image quality.

Results: The 95th percentile absolute difference between the cardiologist vertebral heart score and the algorithm vertebral heart score was 1.05 vertebrae (95% confidence interval: 0.97 to 1.20 vertebrae) with a mean bias of −0.09 vertebrae (95% confidence interval: −0.12 to −0.05 vertebrae). In addition, the model was observed to be well calibrated across the predictive range.

Clinical Significance: We have found the performance of the vertebral heart score algorithm comparable to three board-certified cardiologists. While validation of this vertebral heart score algorithm has shown strong performance compared to veterinarians, further external validation in other clinical settings is warranted before use in those settings.

Journal of Small Animal Practice (2023), 1–7
DOI: 10.1111/jsap.13666

Accepted: 12 July 2023

INTRODUCTION

The vertebral heart score (VHS) is a measurement used to index the heart size relative to the thoracic vertebra and was first described by Buchanan and Bucheler in 1995 and amended slightly for the accommodation of prominent left atria in 2000 (Buchanan & Bücheler 1995, Buchanan 2000). The method involves measuring the long and short axes of the heart from a right lateral radiograph, then measuring the same distances along the spine, starting at the cranial epiphyses of the T4 vertebral body and adding the number of vertebrae together to generate the total score. Since then, a simplified method (Sanchez method) which averages the lengths of the fourth through ninth vertebrae has been shown to have a strong positive correlation with the method originally described by Buchannon and Bücheler (Buchanan & Bücheler 1995, Buchanan 2000, Sánchez *et al.* 2012).

JSAP JOURNAL OF SMALL ANIMAL PRACTICE BSAVA

The VHS is a useful tool for ruling in or out the presence of heart disease in dogs (Guglielmini *et al.* 2009). When echocardiography is not feasible, the VHS can also be used as a substitute for identifying stage B2 degenerative valve disease patients, which is a threshold for initiating cardiac therapy (Ito 2022). Additionally, the absolute VHS and the change in VHS over time have been shown to predict the onset of heart failure in several studies (Boswood *et al.* 2016, 2020). The VHS does have some inherent sources of variability. Two fluoroscopic studies have documented a mean change of around 0.3 to 0.4 vertebrae between systolic and diastolic phases of the cardiac cycle. There can also be a mean change of 0.2 vertebrae between inspiratory and expiratory phases of the respiratory cycle (Olive *et al.* 2015). Finally, human variability studies have shown a mean difference of about 0.4 to 1.0 vertebrae between different readers looking at the same radiographs (Hansson *et al.* 2005).

Recently, the development of computer-aided algorithms for the support of clinical diagnosis in veterinary cardiology has increased (Burti *et al.* 2020, Li *et al.* 2020). Computer-aided clinical decision support increased adherence to clinical guidelines (Taheri Moghadam *et al.* 2021). In addition, human error during routine diagnosis is often unavoidable even for highly trained medical professionals due to human fatigue, inattention and distraction (Alexander 2010, Waite *et al.* 2017). Furthermore, additional sources of variation on VHS measurements can occur based on dog breed, body condition, and cardiac conditions (Puccinelli *et al.* 2021, Baisan & Vulpe 2022, Wiegel *et al.* 2022). The purpose of this study was to assess the performance of the use of a VHS algorithm utilising the simplified Sanchez method compared to manual VHS scoring on 1200 radiographs split between three board-certified veterinary cardiologists using the Buchanan method.

MATERIALS AND METHODS

Study design

A cross sectional, retrospective, cohort design was used in the study. A total of 34,807 canine lateral thoracic radiographs with their respective medical notes were obtained from the IDEXX Vetmedstat™ System and IDEXX Web PACS™ system regardless of health status of the animal. The original VHS scores were extracted from the radiology report of canine chest radiographs to retrospectively evaluate the underlying distribution of VHS scores (Table S1). A total of 1200 lateral radiographs were randomly selected to match the underlying distribution of VHS scores.

Three American College of Veterinary Internal Medicine board-certified cardiologists were enrolled to manually score the 1200 images (400 images each) using the traditional Buchanan method (Buchanan 2000). All cardiologists were masked to the algorithm results. Post-scoring, the cardiologists were unmasked to the results and evaluated the algorithm for misaligned anatomic landmarks (missed cardiac landmarks and missed vertebral landmarks) and the overall quality of images (incomplete visualisation of the entire cardiac silhouette, poor image quality, poor patient positioning).

Algorithm training

In contrast to other machine learning designs for thoracic canine radiograph processing, which rely on image classification or key point prediction, ours centres around semantic segmentation of relevant anatomical features (Burti *et al.* 2020, Zhang *et al.* 2021). Semantic segmentation creates masks for detected objects in the image. These masks are used to then find keypoints in the image to calculate VHS. The model comprises several stages. There is an initial preprocessing of the radiograph where an involving image reorientation by a separate neural net, followed by normalisation and cropping to a standard size. These images are then fed into a convolutional neural net (CNN). The architecture of the CNN is a variant of U-NET, a segmentation network that has become a standard tool in human radiography (Ronneberger *et al.* 2015, Minaee *et al.* 2020, Ulku & Akagunduz 2021). The CNN was trained to predict masks for the cardiac silhouette and intervertebral disc space between the cervical, thoracic, and lumbar vertebrae. Due to the time-intensive process of annotating masks on the vertebrae themselves, an algorithmic differentiation approach on the different types of intervertebral disc space was used. In addition, the CNN was trained to predict masks (*i.e.* regions identified with) for the carina of the trachea and T1 spinous process to serve as landmarks for cardiac silhouette and intervertebral disc space key point identification in the logic layer, respectively.

Second, the mask predictions from the CNN are fed into a logic layer in order to identify key points and compute the VHS (Fig 1). The cardiac silhouette pixel closest to the carina is associated with one of the two points defining the major axis (Fig 1, point labelled A); the other point is defined as the cardiac silhouette pixel furthest away (B), corresponding to the heart apex. In contrast to human-computed VHS approaches, in which the angle between the minor and major axes is often allowed to float, in our logic layer the minor axis is constrained to be orthogonal to the major. The minor axis is subsequently computed by maximising its width across the heart (line segments C and D).

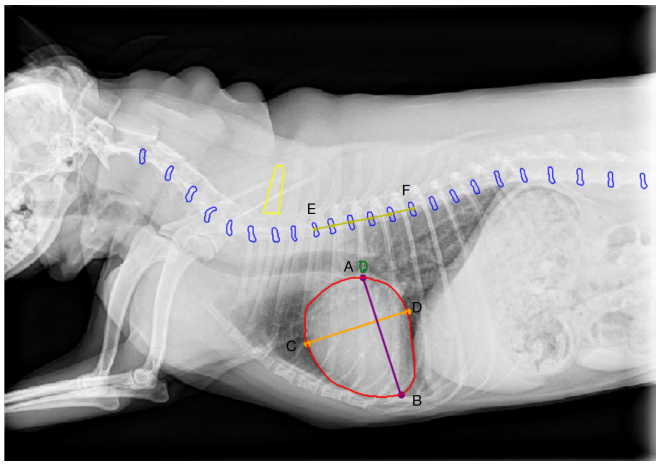

FIG 1. Key point prediction by logic layer, as described in main text. The green, red, blue and yellow contours correspond to the carina, heart, disks and T1 spinous process, respectively, while the purple, yellow and orange lines to the heart major, heart minor and T4-T9 lines, used to compute vertebral heart scores

Finally, the intervertebral disc space to the left of the T1 vertebrae is identified by its position relative to the predicted T1 spinous process mask. The intervertebral disc space immediately to the left T4 and T9 vertebrae are then identified by counting predicted intervertebral disc space masks (line segments E and F, Fig 1), and the distance between these is then used for the VHS computation.

A total of 1559 images, out of a randomly sampled subset of the total available 34,807 images, were determined as having minimal sufficient clarity to identify the relevant anatomical features. Minimal sufficient clarity was defined as an image where the radiologist could diagnose from the image (*i.e.* did not have to retake the image). These radiographs were then annotated as described above to train the CNN. These were split in train, validation and test sets, consisting of 1403, 79 and 77 images, respectively. Mask prediction performance was measured by computing the intersection over union between the predictions and annotated masks for each of the four anatomical classes (cardiac silhouette, carina, intervertebral disc space and T1 spinous process). Training using the Dice loss function was found to generate more accurate mask predictions for small features and mitigated instabilities observed in the learning dynamics, as compared with pixel-wise cross-entropy; for both types of loss functions, we employed class weighting to compensate for the imbalance in the number of pixels associated with the different classes (Jadon 2020). The training process was completed after 42 hours, corresponding to 140 epochs.

Statistical analysis

A simulated power calculation based off requiring a 95th percentile absolute difference less than or equal to 1.0 vertebrae (defined pre-hoc via a panel of two cardiologists) was used to come up with sample size of 1200. Bias was defined as the difference between the cardiologist's VHS score and the algorithm VHS score ($VHS_{cardiologist} - VHS_{algorithm}$). Bland–Altman plots, with mean bias, 95% confidence intervals for the mean bias, limits of agreement (1.96×sd) and 95th percentile of absolute difference were reported. Confidence intervals for quantiles are based off the Nyblom method and confidence intervals for proportions are based off the binomial exact method (Clopper & Pearson 1934, Nyblom 1992). Mean bias and Bland–Altman plots stratified by cardiologist (reader#1-3) and missed anatomical landmarks (cardiac landmarks and vertebral landmarks) are reported. Predicted bias, slope and intercept were calculated using a Passing-Bablock (PB) regression and the 95% confidence interval (CI) was calculated using the bootstrap method (Passing & Bablok 1983, Carpenter & Bithell 2000). Predicted bias was calculated at four medical decision points (VHS scores of 8.0, 9.0, 10.7 and 12.0 vertebrae) using the PB-regression equation and 95% CI using the bootstrap method. Histograms of the difference, squared difference, and absolute difference are presented. A scatter plot of the regression equation along with the line of agreement (intercept=0.0, slope=1.0) was also reported.

To assess model fit, the bootstrap procedure was used to obtain unbiased estimates of a model's future performance using resampling (original sample=1171, bootstrapped samples 2000 repetitions). The bootstrap resampled estimate, optimism and optimism-corrected indices are reported for mean squared error (MSE), g coefficient, intercept and slope (Table S2). Calibration plots of the observed *versus* predicted probability plot were used to validate the accuracy of predictions across the predictive range (Harrell *et al.* 1996, Steyerberg *et al.* 2001, Steyerberg & Harrell 2016, Van Calster *et al.* 2019).

Post-scoring, the cardiologists evaluated the images for misaligned anatomic landmarks and low-quality images (see study design section above). This information is reported as a count and proportion. An additional exploratory analysis recalculating the mean difference 95th percentile absolute difference redone with the misaligned anatomic landmarks and low-quality images removed.

Software

Statistical analysis was done using R version 4.0.4. Graphical analyses and data cleaning were done using *ggplot2* and *tidyverse* (Wickham & RStudio 2017). PB regression, bootstrap confidence intervals and medical decision point analysis were done using the *mcr* package (Ekaterina Manuilova *et al.* 2014). Bootstrap validation and calibration plots were performed using the Hmisc and rms packages (Jr & others 2019, Harrell 2021). Model training and development was done using Python 3.7 and PyTorch 1.5.1.

Ethical approval and informed consent

Only the radiographs and medical record information of client-owned animals were accessed, and no direct patient procedures were performed, therefore prior ethical approval from a committee was not necessary. All radiographs were obtained and submitted to a commercial IDEXX system by a practicing veterinarian during the normal diagnostic workup and monitoring of the patients in their care. All radiographs were obtained with the consent of the pet owner. To ensure privacy, additional demographic information on the pet owner or on the veterinarian who submitted the sample was not collected.

Table 1. Summary statistics of difference between the vertebral heart score of cardiologist and algorithm and regression estimates

Mean bias		95th percentile absolute difference		Passing-Bablock regression			
				Intercept		Slope	
Mean	95% CI	95'tile	95% CI	Intercept	95% CI	Slope	95% CI
−0.09	−0.12 to −0.05	1.05	0.97 to 1.20	−0.42	−0.71 to −0.15	1.05	1.02 to 1.08

CI Confidence interval

RESULTS

A total of 1200 lateral thoracic radiographs images were collected. One case could not be scored by the cardiologist and 28 images were reported as un-scorable by the algorithm through hanging rules. This left 1171 images available for analysis. The median age of the dogs enrolled in this study was 10.6 (interquartile range: 7.3 to 12.6 years). The three most frequent breeds were mixed breed (n=677), Chihuahua (n=65) and terrier (n=38). Two types of digital radiography systems were used in this study, computed radiography (CR, 8.2%, 96/1171) and digital radiography (DR, 91.8%, 1075/1171). The range of the VHS scores measured by the cardiologists was 7.1 to 14.8 vertebrae.

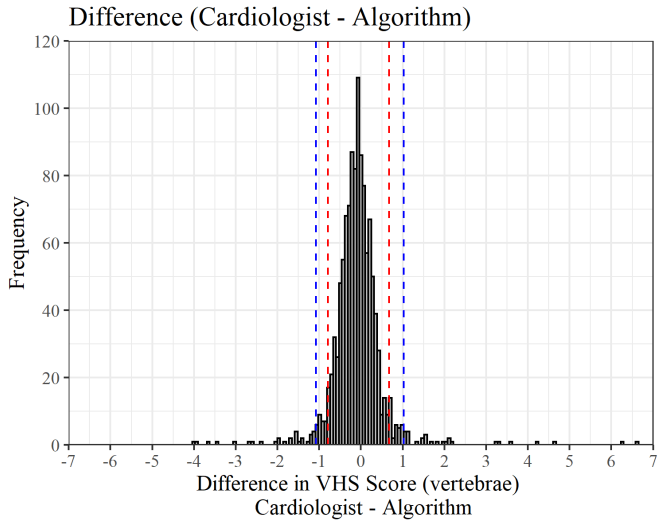

FIG 2. Histograms of the difference between vertebral heart scores (VHS) of cardiologists and algorithm. The red lines indicate the 5th percentile and 95th percentile. The blue dashed lines represent the 2.5th percentile and the 97.5th percentile

The 95th percentile absolute difference between the cardiologist VHS score and the algorithm VHS score was 1.05 vertebrae (95% CI: 0.97 to 1.20 vertebrae, Fig S1A, Table 1) with a mean bias of −0.09 vertebrae (95% CI: −0.12 to −0.05 vertebrae, limit of agreement=1.27, Figs 2 and 3, Table 1). The intercept for the PB regression was −0.42 (95% CI: −0.707 to −0.149) and the slope was 1.05 (95% CI: 1.024 to 1.078, Fig 4, Table 1), suggesting a small bias. At the 8.0 and 9.0 vertebral medical decision points, no bias in the estimates was

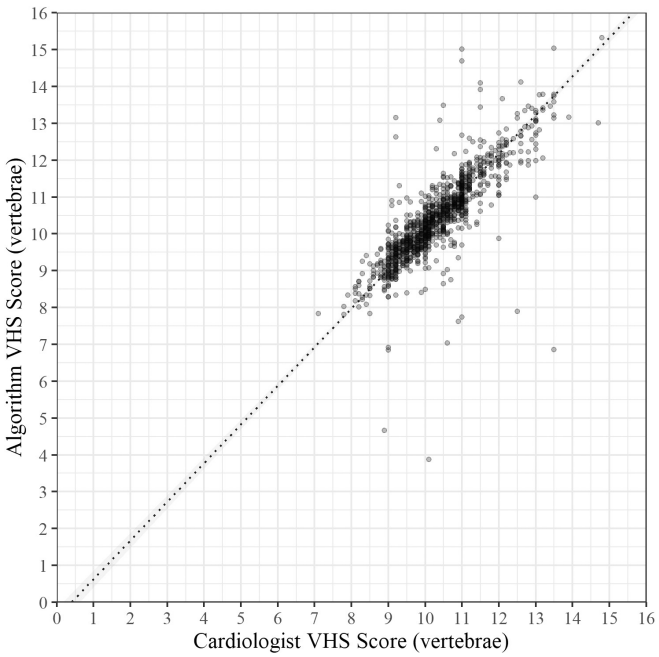

FIG 4. Scatter plot between cardiologist and algorithm vertebral heart scores (VHS) with Passing-Bablock regression. The dashed black lines represent the regression line and the grey shaded regions represent the 95% confidence interval for the regression line

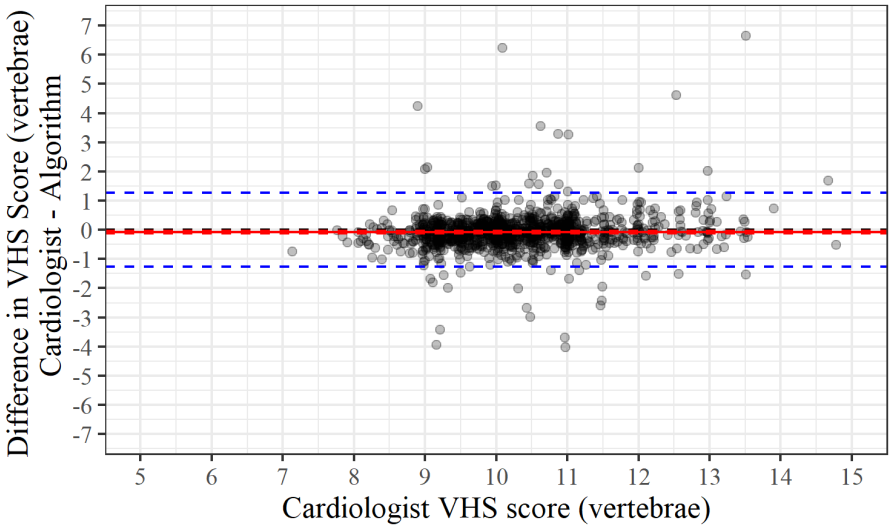

FIG 3. Bland–Altman plot between algorithm and cardiologist vertebral heart scores (VHS). Red solid and dashed lines represent the mean bias and 95% confidence interval around the mean bias. Blue dashed lines represent the limits of agreement as defined by 1.96×sd of the difference in VHS

observed (Table 2). At the 10.7 and 12.0 medical decision points, a less than 0.2 vertebrae bias was observed (Table 2). Although bias was detected, the model was well calibrated across the predictive range (Fig 5, Table S2). The mean bias between the cardiologists and the algorithm were similar when stratified by reader (Table 3).

Following scoring and unmasking, each of the cardiologists reviewed the algorithm's identification of anatomical landmarks and evaluated whether it was possible to score the image appropriately. The algorithm misplaced the vertebral anatomical landmarks 0.9% (11/1171) of the time. The algorithm misplaced the cardiac anatomical landmarks 3.5% (41/1171) of the time. There were 20 radiographs where the cardiologist had difficulty scoring the image for the following reasons: incomplete visualisation of the entire cardiac silhouette (14), poor image quality (five) and poor patient positioning (one). When excluding low-quality images and images where the algorithm missed the vertebral and cardiac landmarks, the overall mean bias decreased (Table 4).

Table 2. Calculated bias for medical decision points based off the Passing-Bablock regression of vertebral heart scores (VHS) from cardiologists and VHS from algorithm

Decision point (vertebrae)	Bais	95% CI
8.0	−0.02	−0.09 to 0.039
9.0	0.03	−0.02 to 0.07
10.7	0.12	0.08 to 0.15
12.0	0.18	0.12 to 0.24

CI Confidence interval

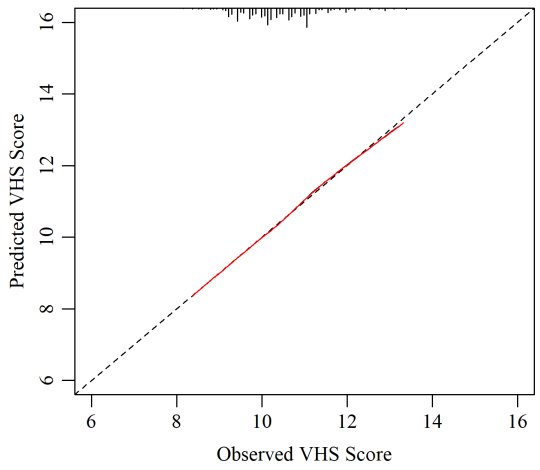

FIG 5. Calibration curve of actual *versus* predicted vertebral heart scores (VHS). The dashed black line represents the Youden equivalency line (slope=1.0, intercept=0.0), the solid red line represents the bias-corrected fit. At the top of the figure is a histogram of the observed VHS values. The bootstrapped model was run with 2000 repetitions with a sample of 1171 and gave a mean absolute error of 0.024

DISCUSSION

Errors in the interpretation of radiographs are frequent (Berlin 2007). Studies in human and veterinary medicine have shown major diagnostic errors in up to 20% of cases (Berner & Graber 2008, Graber 2013, Cohen *et al.* 2023). Recently, the development of computer-aided algorithms for the support of clinical diagnosis in both veterinary and human medicine has increased (Shiraishi *et al.* 2011, El-Dahshan *et al.* 2014, Szlosek & Ferrett 2016, Burti *et al.* 2020, Estrada *et al.* 2021). Computer-aided clinical decision support can increased adherence to clinical guidelines and help mitigate some forms of error (Alexander 2010, Taheri Moghadam *et al.* 2021).

In this validation study, the VHS algorithm was observed to have a mean bias less than 0.1 vertebrae on 1171 radiographs read by one of three veterinary cardiologists. Hansson and colleagues have previously reported that the mean bias across 16 veterinarians of different experience levels reading 50 radiographs from Cavalier King Charles Spaniels was 1.0 vertebrae (Hansson *et al.* 2005). While our study showed much smaller mean bias across multiple breeds, only three cardiologists were used (with a single read of each radiograph) in the validation in comparison to the 16 subjects and a single breed used by Hansson and colleagues (Hansson *et al.* 2005). While the Hansson study was performed on two distinct patient cohorts (normal-sized hearts and those with enlarged hearts) our study pooled both normal and enlarged hearts by randomly selecting subjects that matched the distribution of VHS scores population (as described in the methods section). In addition, as none of the images overlapped among the cardiologists, inter-rater agreement between the three cardiologists was not assessed. Lamb *et al.* have previously reported a comparison of 126 dogs across three veterinary practitioners of varying levels of experience showing a maximum likely difference (defined as mean error ±1.96 sd) of >1.0 vertebrae (Lamb *et al.* 2000). This is comparable to the findings in this study with a mean difference of −0.1 vertebrae and limits of agreement (1.96×sd) of 1.3 vertebrae.

The bias across different medical decision points (VHS scores of 8.0, 9.0, 10.7, 12.0 vertebrae) was not found to be constant. There was a small negative bias at 8.0 vertebrae (with the

Table 3. Summary statistics of difference between the vertebral heart score of cardiologist and algorithm and regression estimates stratified by reader

Reader	Mean bias	95% CI	sd	Limits of agreement	95'tile absolute difference	95% CI
#1	−0.04	−0.11 to 0.03	0.70	1.37	1.12	0.95 to 1.66
#2	−0.16	−0.22 to −0.09	0.63	1.23	1.02	0.81 to 1.15
#3	−0.06	−0.12 to 0.002	0.62	1.21	1.04	0.94 to 1.7

CI Confidence interval

Table 4. Summary statistics of difference between the vertebral heart score of Cardiologist and Algorithm when image quality and algorithm anatomic landmark cases were removed

Removed cases	Mean bias (vertebrae)	95% CI	sd	Limits of agreement	95'tile absolute difference	95% CI
Missed cardiac landmarks	−0.07	−0.10 to −0.04	0.57	1.12	0.99	0.88 to 1.07
Missed vertebral landmarks	−0.09	−0.12 to −0.05	0.65	1.27	1.04	0.96 to 1.16
Low-quality images	−0.10	−0.13 to −0.06	0.63	1.23	1.02	0.94 to 1.13
Low-quality images and missed anatomic landmarks	−0.08	−0.11 to −0.05	0.54	1.05	0.93	0.83 to 1.02

CI Confidence interval

algorithm reporting scores a mean of −0.02 vertebrae lower than cardiologist) which moved to being a positive bias at 12 vertebrae (with the algorithm reporting scores a mean of 0.18 vertebrae higher than cardiologist). The non-constant bias in VHS score prediction can partially be explained by the lack of VHS scores around the 8.0 vertebrae medical decision point, with less than 1 % of all results being below 8.7 vertebrae and less than 5% of results less than 9.0 vertebrae. Even with the non-constant bias across the range of VHS scores, the overall calibration curve showed strong performance with a mean absolute error of 0.02 vertebrae, a bias-corrected slope of 1.00 and a bias-corrected intercept of −0.01. With the slope so close to one, the intercept close to zero and a well-fitted calibration curve, the VHS algorithm showed strong performance. While both Lamb *et al.* and Hansson *et al.* showed direct comparisons of the interobserver variability (through mean error) in VHS was large, Nakayama *et al.* observed smaller differences in indirect measures of variability using the coefficient of variation were shown to be 8.0% (Nakayama *et al.* 2001).

Finally, the algorithm misidentified 0.9% of vertebral landmarks and 3.5% of cardiac landmarks. Although infrequent, this led to bias and was not an acceptable result in practice. When used in clinical practice, the lines placed by the algorithm over the radiograph are designed to be manually repositioned if needed. Manual repositioning of the lines changes their colour as a visual indicator that the tool had been adjusted.

There were some challenges when working with these data. While our goal was to review quality radiographs this study included some images with low contrast and cropping of the cardiac silhouette that lead to the inability of scoring or poor performance. This study was performed on data from a single commercial laboratory, thus the generalisation of this algorithm and results into other medical systems would need additional validation. Caution should be used before applying this algorithm to additional clinical environments as the algorithm was not assessed for its robustness in different populations. It should be noted that while VHS is regularly used to assess the size of the heart, it is not a perfect substitute for echocardiography for the identification of stage B2 heart disease and the assessment of the VHS algorithm detection of heart disease was out of the scope of this study (Ito 2022).

We have found the performance of the VHS algorithm comparable to that of three board-certified cardiologists. While validation of this VHS algorithm has shown strong performance compared to three cardiologists (with a single read of each radiograph) in this clinical setting, further external validation in other clinical environments are warranted before use in those systems.

Author contributions

Jessica Solomon: Data acquisition; analysis and interpretation; article drafting; article revision; final approval. **Scott Bender:** Conception and design (co-lead); article drafting. **Pavan Durgempudi:** Conception and design (co-lead); data acquisition (lead). **Caitline Robar:** Conception and design; article revision; final approval. **Michael Cocchiaro:** Data acquisition; analysis and interpretation. **Sean Turner:** Conception and design. **Chris Watson:** Conception and design. **Joseph Healy:** Conception and design. **Allison Spake:** Data acquisition; analysis and interpretation; final approval. **Donald Szlosek:** Analysis and interpretation (lead); article drafting (lead); article revision (lead); final approval.

Conflict of interest

All authors work at IDEXX Inc. and were given compensation for working on this study.

References

Alexander, K. (2010) Reducing error in radiographic interpretation. *The Canadian Veterinary Journal* **51**, 533-536

Baisan, R. A. & Vulpe, V. (2022) Vertebral heart size and vertebral left atrial size reference ranges in healthy Maltese dogs. *Veterinary Radiology & Ultrasound* **63**, 18-22

Berlin, L. (2007) Radiologic errors and malpractice: a blurry distinction. *AJR. American Journal of Roentgenology* **189**, 517-522

Berner, E. S. & Graber, M. L. (2008) Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine* **121**, S2-S23

Boswood, A., Häggström, J., Gordon, S. G., *et al*. (2016) Effect of Pimobendan in dogs with preclinical myxomatous mitral valve disease and cardiomegaly: the EPIC study – a randomized clinical trial. *Journal of Veterinary Internal Medicine* **30**, 1765-1779

Boswood, A., Gordon, S. G., Häggström, J., *et al*. (2020) Temporal changes in clinical and radiographic variables in dogs with preclinical myxomatous mitral valve disease: the EPIC study. *Journal of Veterinary Internal Medicine* **34**, 1108-1118

Buchanan, J. W. (2000) Vertebral scale system to measure heart size in radiographs. *The Veterinary Clinics of North America. Small Animal Practice* **30**, vii-393

Buchanan, J. W. & Bücheler, J. (1995) Vertebral scale system to measure canine heart size in radiographs. *Journal of the American Veterinary Medical Association* **206**, 194-199

Burti, S., Longhin Osti, V., Zotti, A., *et al*. (2020) Use of deep learning to detect cardiomegaly on thoracic radiographs in dogs. *The Veterinary Journal* **262**, 105505

Carpenter, J. & Bithell, J. (2000) Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine* **19**, 1141-1164

Clopper, C. J. & Pearson, E. S. (1934) The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **26**, 404-413

Cohen, J., Fischetti, A. J. & Daverio, H. (2023) Veterinary radiologic error rate as determined by necropsy. *Veterinary Radiology & Ultrasound* **64**, 573-584

Ekaterina Manuilova, Fabian Model & Andre Schuetzenmeister (2014) mcr: Method Comparison Regression

El-Dahshan, E.-S. A., Mohsen, H. M., Revett, K., *et al*. (2014) Computer-aided diagnosis of human brain tumor through MRI: a survey and a new algorithm. *Expert Systems with Applications* **41**, 5526-5545
Estrada, A. H., Spake, A., Kleman, M. E., *et al*. (2021) Diagnostic accuracy of computer aided electrocardiogram analysis in dogs. *The Journal of Small Animal Practice* **62**, 145-149
Graber, M. L. (2013) The incidence of diagnostic error in medicine. *BMJ Quality & Safety* **22**, ii21-ii27
Guglielmini, C., Diana, A., Pietra, M., *et al*. (2009) Use of the vertebral heart score in coughing dogs with chronic degenerative mitral valve disease. *The Journal of Veterinary Medical Science* **71**, 9-13
Hansson, K., Häggström, J., Kvart, C., *et al*. (2005) Interobserver variability of vertebral heart size measurements in dogs with normal and enlarged hearts. *Veterinary Radiology & Ultrasound* **46**, 122-130
Harrell, F. (2021) rms: Regression Modeling Strategies
Harrell, F. E., Lee, K. L. & Mark, D. B. (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**, 361-387
Ito, D. (2022) Vertebral heart size is associated with cardiac enlargement in Chihuahuas with myxomatous mitral valve disease. *The Canadian Veterinary Journal* **63**, 627-632
Jadon, S. (2020) A survey of loss functions for semantic segmentation. Proceedings of the 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). pp 1–7
Jr, F.E.H. & others, with contributions from C.D. and many (2019) Hmisc: Harrell Miscellaneous. https://cran.r-project.org/web/packages/Hmisc/index.html. Accessed May 5, 2023
Lamb, C. R., Tyler, M., Boswood, A., *et al*. (2000) Assessment of the value of the vertebral heart scale in the radiographic diagnosis of cardiac disease in dogs. *The Veterinary Record* **146**, 687-690
Li, S., Wang, Z., Visser, L. C., *et al*. (2020) Pilot study: application of artificial intelligence for detecting left atrial enlargement on canine thoracic radiographs. *Veterinary Radiology & Ultrasound* **61**, 611-618
Minaee, S., Boykov, Y., Porikli, F., *et al*. (2020) Image segmentation using deep learning: a survey. *ArXiv:2001.05566 [cs]*
Nakayama, H., Nakayama, T. & Hamlin, R. L. (2001) Correlation of cardiac enlargement as assessed by vertebral heart size and echocardiographic and electrocardiographic findings in dogs with evolving cardiomegaly due to rapid ventricular pacing. *Journal of Veterinary Internal Medicine* **15**, 217-221
Nyblom, J. (1992) Note on interpolated order statistics. *Statistics & Probability Letters* **14**, 129-131
Olive, J., Javard, R., Specchi, S., *et al*. (2015) Effect of cardiac and respiratory cycles on vertebral heart score measured on fluoroscopic images of healthy dogs. *Journal of the American Veterinary Medical Association* **246**, 1091-1097
Passing, H. & Bablok (1983) A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, part I. *Journal of Clinical Chemistry and Clinical Biochemistry* **21**, 709-720
Puccinelli, C., Citi, S., Vezzosi, T., *et al*. (2021) A radiographic study of breed-specific vertebral heart score and vertebral left atrial size in chihuahuas. *Veterinary Radiology & Ultrasound* **62**, 20-26
Ronneberger, O., Fischer, P. & Brox, T. (2015) U-net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Lecture Notes in Computer Science. Eds N. Navab, J. Hornegger, W. M. Wells, *et al*. University of Freiburg, Freiburg im Breisgau, Germany. pp 234-241
Sánchez, X., Prandi, D., Badiella, L., *et al*. (2012) A new method of computing the vertebral heart scale by means of direct standardisation. *The Journal of Small Animal Practice* **53**, 641-645
Shiraishi, J., Li, Q., Appelbaum, D., *et al*. (2011) Computer-aided diagnosis and artificial intelligence in clinical imaging. *Seminars in Nuclear Medicine* **41**, 449-462
Steyerberg, E. W. & Harrell, F. E. (2016) Prediction models need appropriate internal, internal-external, and external validation. *Journal of Clinical Epidemiology* **69**, 245-247
Steyerberg, E. W., Harrell, F. E., Borsboom, G. J. J. M., *et al*. (2001) Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology* **54**, 774-781
Szlosek, D. A. & Ferrett, J. (2016) Using machine learning and natural language processing algorithms to automate the evaluation of clinical decision support in electronic medical record systems. *EGEMS (Washington, DC)* **4**, 1222
Taheri Moghadam, S., Sadoughi, F., Velayati, F., *et al*. (2021) The effects of clinical decision support system for prescribing medication on patient outcomes and physician practice performance: a systematic review and meta-analysis. *BMC Medical Informatics and Decision Making* **21**, 98
Ulku, I. & Akagunduz, E. (2021) A survey on deep learning-based architectures for semantic segmentation on 2D images. *ArXiv:1912.10230 [cs]*
Van Calster, B., McLernon, D. J., van Smeden, M., *et al*. (2019) Calibration: the Achilles heel of predictive analytics. *BMC Medicine* **17**, 230
Waite, S., Scott, J., Gale, B., *et al*. (2017) Interpretive error in radiology. *American Journal of Roentgenology* **208**, 739-749
Wickham, H. & RStudio (2017) tidyverse: Easily Install and Load the "Tidyverse"
Wiegel, P. S., Mach, R., Nolte, I., *et al*. (2022) Breed-specific values for vertebral heart score (VHS), vertebral left atrial size (VLAS), and radiographic left atrial dimension (RLAD) in pugs without cardiac disease, and their relationship to Brachycephalic Obstructive Airway Syndrome (BOAS). *PLoS One* **17**, e0274085
Zhang, J., Chen, Z. & Tao, D. (2021) Towards high performance human keypoint detection. *ArXiv:2002.00537 [cs, eess]*

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig S1. (A) Histogram of absolute difference between the VHS scores of the algorithm and cardiologists. The red dashed line indicates the 95th percentile of absolute difference. (B) Histogram of squared difference

Table S1. The original VHS scores were extracted from the radiograph reports to retrospectively evaluate the underlying distribution of VHS scores

Table S2. Calibration and discrimination Indices from internal validation