HHS Public Access

Author manuscript

Toxicol Pathol. Author manuscript; available in PMC 2024 September 20.

Published in final edited form as:

Toxicol Pathol. 2024 July ; 52(5): 258–265. doi:10.1177/01926233241259998.

Inter-Rater and Intra-Rater Agreement in Scoring Severity of Rodent Cardiomyopathy and Relation to Artificial Intelligence–Based Scoring

Thomas J. Steinbach[1], **Debra A. Tokarz**[1], **Caroll A. Co**[2], **Shawn F. Harris**[2], **Sandra J. McBride**[2], **Keith R. Shockley**[3], **Avinash Lokhande**[4], **Gargi Srivastava**[4], **Rajesh Ugalmugle**[4], **Arshad Kazi**[4], **Emily Singletary**[1], **Mark F. Cesta**[3], **Heath C. Thomas**[5], **Vivian S. Chen**[6,7], **Kristen Hobbie**[8], **Torrie A. Crabbs**[1]

[1]Experimental Pathology Laboratories, Inc., Research Triangle Park, North Carolina, USA

[2]Social & Scientific Systems, Inc., Durham, North Carolina, USA

[3]National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA

[4]AIRA Matrix, Mumbai, India

[5]Aclairo Pharmaceutical Development Group, Vienna, Virginia, USA

[6]Charles River Laboratories, Durham, North Carolina, USA

[7]Biogen, Cambridge, Massachusetts, USA

[8]Inotiv, Research Triangle Park, North Carolina, USA

Abstract

We previously developed a computer-assisted image analysis algorithm to detect and quantify the microscopic features of rodent progressive cardiomyopathy (PCM) in rat heart histologic sections and validated the results with a panel of five veterinary toxicologic pathologists using a multinomial logistic model. In this study, we assessed both the inter-rater and intra-rater agreement of the pathologists and compared pathologists' ratings to the artificial intelligence (AI)-predicted scores. Pathologists and the AI algorithm were presented with 500 slides of rodent heart. They quantified the amount of cardiomyopathy in each slide. A total of 200 of these slides were novel to this study, whereas 100 slides were intentionally selected for repetition from the previous study. After a washout period of more than six months, the repeated slides were examined to assess intra-rater agreement among pathologists. We found the intra-rater agreement to be substantial,

Corresponding Author: Thomas J. Steinbach, Experimental Pathology Laboratories, Inc., Research Triangle Park, NC, USA., tsteinbach@epl-inc.com.

Animal Ethics Committee

All data produced for this manuscript was taken from glass slides and whole slide images in an archive. No additional animals were used and so no Animal Ethics Committee approval was required.

Supplemental Material

Supplemental material for this article is available online.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

with weighted Cohen's kappa values ranging from k = 0.64 to 0.80. Intra-rater variability is not a concern for the deterministic AI. The inter-rater agreement across pathologists was moderate (Cohen's kappa k = 0.56). These results demonstrate the utility of AI algorithms as a tool for pathologists to increase sensitivity and specificity for the histopathologic assessment of the heart in toxicology studies.

Keywords

inter-rater agreement; intra-rater agreement; kappa; cardiomyopathy; deep learning; artificial intelligence; rat; Sprague Dawley; computer-assisted image analysis

Introduction

Progressive cardiomyopathy (PCM) is a common background finding in rodents, especially in rats, where incidence and severity typically increase with age.[10] The histologic features of PCM in rodents consist of cardiomyocyte degeneration and/or necrosis, mononuclear inflammatory cell infiltration, fibrosis, and mineralization. Progressive cardiomyopathy first appears as small, discrete foci of myocardial degeneration and/or necrosis accompanied by inflammatory cell infiltration.[4,9,10] Eventually necrosis and inflammation are replaced by varying degrees of fibrosis and to a lesser extent mineralization.[10]

The spectrum of lesions and temporal changes that characterize PCM present a challenge for creating a cohesive approach to diagnosing and grading among pathologists. Because PCM lesions are commonly found in toxicity studies using rats, we differentiate spontaneous lesions from those related to cardiotoxicity, which can mimic the morphologic changes found in PCM. Application of harmonized diagnostic criteria for PCM increases concordance among pathologists in the presence of varying diagnostic criteria, terminology, and severity grading approaches.[10] In this context, any approach to diagnosing and grading PCM should be repeatable, generating the same results using the same team and operating system, as well as reproducible, generating the same study results using a different team and operating system.[16] Further, inter-rater agreement (agreement among raters) and intra-rater agreement (agreement within raters) are critical to understanding the limits that approaches to harmonization such as those suggested by Hailey et al[9] will have in achieving repeatable results. The extent to which pathologists agree with each other (inter-rater) and with themselves (intra-rater) will define the extent to which PCM can be differentiated from test article-related cardiotoxicity.

Inter-rater and intra-rater agreement among pathologists has been well studied in human and to a lesser extent in veterinary pathology in the context of clinical diagnostic medicine.[3,12,17] In general, inter-rater agreement is higher when the confidence of the diagnosis is high[2] and when using a lower resolution scoring system with fewer severity levels for the pathologist to select.[19] Difficult to diagnose lymphomas and sarcomas can reduce inter-rater agreement,[3] as can diagnosis types, such as necroinflammatory changes in the liver.[11] In our experience, inter- and intra-rater agreement have not been examined as closely in the field of toxicologic pathology as in diagnostic pathology.

The use of artificial intelligence (AI) in pathology has been evolving in both human and veterinary medical fields. In addition to the algorithm, we developed to detect cardiomyopathy in Tokarz et al,[20] recently published AI-based analyses of animal tissues include spermatogenic staging assessments in rats, identification and quantification of colitis in mice, retinal atrophy and hepatocellular hypertrophy in rodents, assessment of bone marrow cellularity in non-human primates, and detecting mitotic figures.[22] A survey in June of 2021 by Palazzi et al of veterinary pathologists (mostly within the toxicologic pathology field) indicated that the use of AI was well established and had generally strong support. However, challenges such as regulatory acceptance and scalability remain.[15] In human pathology, AI is being explored in tumor detection and identification, object detection such as mitosis, quantification such as Ki-67 proliferative index, and integration with genomics data.[14] Challenges remain here also, as currently there is only one Food and Drug Administration (FDA)-approved AI-based software solution for the detection of cancer.[18] In this article, we distinguish the deterministic AI algorithm we developed for rodent cardiomyopathy, which we abbreviate as AIA, from broader applications of AI, which we abbreviate as AI.

In this study, we added new data and analysis to the work published in Tokarz et al[20] to investigate the relationship between intra-rater and inter-rater agreement among pathologists with predicted scores from the AIA. We quantify intra-rater agreement by letting the same set of pathologists examine 100 previously seen slides. In addition, we added 200 new whole slide images (WSIs) to validate whether the multinomial logistic regression model developed in Tokarz et al[20] performed similarly in terms of classification accuracy with new data. The analysis for the model validation is included in the Supplemental section. In total, we collected cardiomyopathy severity grading on 500 distinct slides from a panel of 5 pathologists allowing us to get an estimate of inter-rater agreement (Table 1). We hypothesize that intra-rater agreement would be greater than inter-rater agreement and that the repeatability of AI makes it a useful tool for mitigating diagnostic discrepancies among pathologists. To our knowledge, this is the first study to investigate intra-rater agreement using the diagnostic criteria established by Hailey et al[9] and to investigate inter-rater agreement with this criterion on a large data set.

Materials and Methods

Case selection, whole slide imaging, AIA training, network architecture, and implementation details were similar to those performed in Tokarz et al.[20]

Evaluation of Test Cohort

A set of 300 hematoxylin and eosin (H&E)-stained slides containing heart sections meeting the case selection criteria were chosen as a test cohort. The chosen sections represented the spectrum of cardiomyopathy-associated histologic changes and a range of severities. To evaluate intra-rater agreement, 100 out of the 300 slides were chosen from the set of slides previously examined in Tokarz et al.[20] To select these 100 slides, we used hierarchical clustering with Ward linkage[7] to group the original 298 slides (2 slides with missing AIA scores were excluded) into 10 distinct clusters, based on information that was

previously obtained (median severity grade, AIA predicted score, relative share of fibrosis, and percentage of mononuclear cells [MNCs]) in Tokarz et al.[20] Then, 10 slides from each of the 10 clusters were randomly selected to get 100 repeat slides. This stratified selection method ensured even representation of the variety of information found in the 300 slides used in Tokarz et al.[20] An additional 200 slides (the non-repeats) were selected from available control slides in the National Toxicology Program (NTP) archive. These slides were used to validate the statistical model from Tokarz et al[20] (See Supplemental section for results).

The same five veterinary pathologists who participated in Tokarz et al[20] independently evaluated all 300 glass slides, including the 100 which were previously examined. The washout time exceeded 6 months. Pathologists were instructed to assign one of the following diagnoses: no abnormality, PCM, and other. For slides with a PCM diagnosis, they were further instructed to assign a grade of PCM ranging from 1 to 5 based on criteria modified from Hailey et al.[9] Slides with no abnormality were given a grade of 0. No grade was given for slides showing a non-PCM diagnosis. The pathologists focused on the main components of PCM, which consists of necrosis, mononuclear cell infiltrate, fibrosis, and mineralization. The grading criteria were as follows:

Grade 1: Aggregate lesion size <45% field of view (FOV) for 40X objective.

Grade 2: Aggregate lesion size ≥ 45% FOV for 40X objective and <5% of the heart section.

Grade 3: Aggregate lesion size ≥5% but <25% of the heart section.

Grade 4: Aggregate lesion size ≥25% but <50% of the heart section.

Grade 5: Aggregate lesion size ≥50% of the heart section.

Similarly, the AIA assessed the extent of PCM based on the same components of necrosis, mononuclear cell infiltrate, fibrosis, and mineralization. The components were assessed independently and scored as a percent area affected (eg, area of necrosis/total area of the heart). The AIA's overall assessment of PCM was taken as a sum of these individual components to quantify the percentage of the heart showing lesions consistent with progressive cardiomyopathy. The AIA is sensitive in detecting all PCM components and returns a continuous numeric value corresponding to the percent of heart area affected. Based on previous work, the AIA was able to quantify percent area affected ranging from 0.015% to 8.65%.

To confirm the repeatability of the AIA, we compared the AIA output from the 100 repeated slides for repeated runs of the AIA. We reported the mean difference and the Pearson correlation coefficient as a measure of agreement, rather than the Cohen's kappa statistic, because the output of the algorithm is a continuous variable. The pathologists, on the contrary, scored the amount of cardiomyopathy into distinct grades (0–5), so the Cohen's kappa or weighted Cohen's kappa was a more appropriate statistic. Finally, we measured the inter-rater agreement of the pathologists using percent agreement and weighted Cohen's kappa.

Comparison of the pathologist's results was described in terms of inter-rater and intra-rater agreement. Inter-rater and intra-rater agreement can be expressed in terms of percent agreement, Cohen's kappa, or Pearson's correlation.[13] Percent agreement is a direct comparison of the results of the raters involved and a calculation of how many times they agreed. Percent agreement does not consider the possibility that pathologists would score the same grade simply by chance. Cohen's kappa, a type of correlation coefficient, improves on simple percent agreement calculations by accounting for the effect of random matches.[5] Cohen's kappa compares the measured percent agreement to the amount of agreement that would occur by chance and can range between 0 and 1 (although negative values can occur), with a 0 value indicating agreement only by chance and a value of 1 indicating complete agreement. Interpretation of Cohen's kappa is shown in Table 2.

Often in pathology, we are concerned only with "major disagreement" between pathologists. If, as in the case of scoring cardiomyopathy in a range from 0–5, one pathologist scores a grade 2, while another scores a grade 3, we could consider their scores to be in general agreement. If the same pathologist scored a slide a grade 1 and another scored a grade 5, we would consider their scores to be in major disagreement. For these cases, a weighted Cohen's kappa may be appropriate, in which less weight is assigned to agreement as categories are further apart.[5,19]

Statistical Analysis

Statistical analyses on both pathologist intra-rater and interrater agreement were performed in R v4.1.2 using the "irr" package v0.84.1.[8] Agreement was defined in terms of the percentage of slides where the raters agreed on the grade and in terms of the Cohen's kappa statistic to account for chance agreement. For both measures, we calculated agreement two different ways to account for some tolerance in the ratings. We utilized a strict rule where agreement is reached only when two raters gave the exact same grade and a more lenient rule allowing for some margin of error between the raters. For percent agreement, the margin of error was chosen to be within ±1 level, such that agreement is reached when the difference in ratings for any pair of pathologists is at most 1 grade. For Cohen's kappa, a linear weighting scheme was utilized, assuming equal spacing across severity grades and attribution of smaller penalties on near disagreements relative to far disagreements. Pathologist intra-rater reliability was measured on the 100 repeated slides, whereas inter-rater reliability was measured using all 500 unique slides from the previous and current studies. Owing to small counts in the grade 5 severity category, grades 4 and 5 were combined in all analysis as was done in Tokarz et al.[20] Observations where at least 1 out of the 5 raters gave a non-PCM diagnosis were excluded from the analysis (2 out of 500).

Stratified analyses on intra-rater and inter-rater agreement by AIA predicted score were performed by binning the AIA scores into percentiles—quintiles (for intra-rater agreement) or deciles (for inter-rater agreement). Percentiles partition the data into groups (5 for quintiles and 10 for deciles), with each group containing the same number of observations. For intrarater agreement, percent agreement was calculated for each of the 5 raters in each AIA score quintile, where each quintile bin contained 20 slides. For inter-rater agreement, with a larger sample size of 500, AIA scores were binned into deciles, which correspond to

50 slides per AIA decile. Similar statistics were obtained for the 10 rater pairs at each AIA decile.

Results

Rater Assessments

Group consensus was defined as having at least 3 out of 5 raters in agreement on the exact grade severity. Out of 500 slides, 385 (or 77%) reached group consensus.

Intra-rater assessments.—Consistency within raters varied by individual, with percent intra-rater agreement values ranging from 51% to 73%. When the discrepancy between scores was at most 1 grade level apart, the percent agreement across all pathologists increased substantially to at least 90%. Similarly, the unweighted Cohen's kappa ranged from 0.39 (fair) to 0.65 (substantial). Using the weighted Cohen's kappa statistic, substantial consistency across all five pathologists was observed, with Cohen's kappa values ranging from 0.64 to 0.80. Measures of intra-rater agreement for each of the five pathologists are provided in Figure 1. To assess whether intra-rater agreement varied with the degree of cardiomyopathy, we evaluated percent agreement in relation to AIA score. The AIA score quintiles were: [minimum = 0.01%, 0.14%], (0.14%, 0.30%], (0.30%, 0.56%], (0.56%, 1.58%], and (1.58%, maximum = 8.65%]. The average intra-rater percent agreement across AIA score quintiles ranged from 53.8% in the (0.56%–1.58%] category to 67.8% in the (1.58%–8.65%] category. Results are shown in Figure 2.

Repeating the run of the same AIA with the same training set as before, we do not expect to see any difference in results between the two runs as the algorithm is deterministic. This was confirmed, as the percent of heart affected by cardiomyopathy results from the two different runs of the AIA on the same 100 slides showed a mean difference of 0.0027%. This small imprecision is likely attributed to a version change in the platforms (inferencing environment) used in the algorithm.

Inter-rater agreement.—Inter-rater agreement among the five pathologists was assessed using percent agreement and Cohen's kappa. The mean agreement between the ten pairs of raters was 45% (min = 33%, max = 65%). When the tolerance for agreement is relaxed by ±1, allowing for perceived concordance when any rater pair's ratings are different by only 1 grade, the mean agreement increased substantially to 90% (min = 84%, max = 97%). Alternatively, the mean inter-rater agreement using unweighted Cohen's kappa among the ten rater pairs was 0.32 (fair) and ranged from 0.17 (slight) to 0.55 (moderate). When the weighted Cohen's kappa was used, the mean agreement increased to 0.56 (moderate), with values ranging from 0.46 (moderate) to 0.72 (substantial). Figure 3 shows the distribution of agreement values for strict agreement (exact grade) and relaxed agreement allowing for some tolerance in rater grading. The AIA score deciles were [minimum = 0.015%, 0.08%], (0.08%, 0.13%], (0.13%, 0.20%], (0.20%, 0.28%], (0.28%, 0.37%], (0.37%, 0.48%], (0.48%, 0.75%], (0.75%, 1.22%], (1.22%, 2.30%], and (2.30%, maximum = 8.65%]. The percent agreement across AIA score deciles show a U-shape pattern where inter-rater agreement was highest on samples showing the least and most amount of lesion

detected by the AIA and lower agreement in the middle categories. This pattern is consistent when percent agreement was calculated by median grade, as shown in Figure 4.

Discussion

Intra-rater weighted Cohen's kappa values for the five pathologists ranged from 0.64 to 0.80, indicating substantial agreement and consistency of scoring within raters. However, Cohen's kappa values were lower for inter-rater agreement (ranging from 0.46 to 0.72), reflecting differences among pathologists. These results are similar to other reported studies of pathologist agreement in veterinary medicine. Intra-observer agreement was high (correlation coefficient 0.78 to 0.91) and interobserver agreement moderate (kappa 0.43 to 0.60) for assessment of canine soft tissue sarcoma[21] while fair (kappa 0.16 to 0.35) for histologic scoring of canine livers.[11]

As expected, the AIA, evaluated on 100 of the same WSIs from Tokarz et al,[20] gave basically the same results, with the slight observed differences attributed to a change in AIA versions. There were no programming changes to the AIA and no changes to its training set.

The relative importance of inter-rater and intra-rater agreement depends on the area of application. For a pathologist reading an individual study to consistently assess the same grade of cardiomyopathy throughout the study, high values of Cohen's kappa for intra-rater agreement would be required. The more consistent the study pathologist is, the more sensitive the study data would be to detecting differences in heart lesions between control and treated groups. Under the conditions of this study, the pathologists had substantial intra-rater agreement. Still, an AIA like the one we developed may be a tool the study pathologist could use to enhance the sensitivity for histopathologic heart lesions in a particular study by providing continuous quantitative data in a repeatable manner.

Similarly, inter-rater agreement supports interpretation of any multistudy analysis. If different study pathologists are involved in reading different studies of the same compound (or toxicologic agent), poor inter-rater agreement among the pathologists would make it difficult to interpret any findings across studies. Here again, using a single AIA on any set of studies being compared for the level of cardiomyopathy would potentially be more repeatable and may make the comparison easier to interpret.

In toxicologic pathology, the use of AI to assist the pathologist is in its infancy. Currently, there are no FDA-approved algorithms in the assessment of whole slide images in nonclinical safety studies. We believe there are a number of potential applications of AIA in toxicologic pathology. As Tokarz et al[20] and this study have demonstrated, the focused application of AI-assistance has the potential to be a powerful tool to enhance the ability of the toxicologic pathologist to detect toxicologically relevant differences between exposure groups within a study and to compare the results of similar studies more easily. Additional applications may be for the AI to act as a "peer reviewer." The pathologist's scores for cardiomyopathy could be compared to the AI's grades as predicted by the statistical model. Where significant differences exist, the pathologist could then review the AI-annotated slide and reassess the grade, if warranted. Similarly, comparison of pathologist's scoring to AI

scoring could be useful in a residency or post-residency training program. Trainees could assess slides (WSIs or glass) and then compare to the AI scoring. Differences could be examined in detail as the trainee would have access to AI-annotated WSIs and could see where and what feature of cardiomyopathy the AI detected. These are just a few examples of where AI-assisted detection, classification, and comparison of grade of cardiomyopathy may be used in the field of toxicologic pathology.

In the results presented here, and as reported in Tokarz et al,[20] the AIA and manual scoring showed large differences in estimates of absolute percent area affected. Among the 500 slides collected in this study, the highest percentage of lesion detected by the AIA was 8.65%. This may seem low when compared to a manual score of 4/5, corresponding to ≥25% in aggregate lesion size. Trained veterinary pathologists performed visual inspections of the AIA annotations for several slides, including those where pathologists scored high levels of cardiomyopathy (grades 4 or 5), and did not find any evidence of the AIA missing or underestimating the characteristic lesions of cardiomyopathy. We think this difference in absolute percentage can be explained by two factors. First, the AIA is quantifying area affected at a pixel level, whereas a pathologist is generally reviewing the slide at a low magnification (eg, 2× magnification) to roughly estimate the percentage affected, rather than precisely measuring any feature. Second, there are a number of visual and cognitive traps, as described in Aeffner et al,[1] that can affect the pathologist's ability to accurately assess size. We contend that these two factors explain at least some or most of the differences where the AIA scored lower percentages of cardiomyopathy than the pathologists. However, we concede the design of these experiments was insufficient to fully explain this difference and that additional research would be required to understand the factors driving the differences between AIA and pathologist perception when assessing cardiomyopathy.

An important area of future work is evaluation of reproducibility of the AI results across different AI models. Using the same sets of WSI training and test data, independently developed AI models could be assessed to determine whether they yield scores that similarly predict pathologist ratings, or equivalently, have high inter-rater agreement. This type of critical assessment would assist in characterizing the generalizability and reliability of AI algorithms, identify AI models that perform well across data sets, and lead to the development of standardized benchmarks and protocols for AI model evaluation. Rigorous evaluation of AI reproducibility is essential to the development of effective and reliable AI-assisted pathology workflows and practices in toxicologic pathology.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the National Institute of Environmental Health Sciences under contract GS-00F-173CA/75N96022F00055 to Social & Scientific Systems, a DLH Holdings Corp Company; contract HHSN273201500014C to Experimental Pathology Laboratories, Inc; contract HHSN273201500012C to Charles River Laboratories, Inc; and contract HHSN273201500013C to Inotiv, Inc.

References

1. Aeffner F, Wilson K, Martin NT, et al. The gold standard paradox in digital image analysis: manual versus automated scoring as ground truth. Arch Pathol Lab Med. 2017;141(9):1267–1275. doi:10.5858/arpa.2016-0386-RA [PubMed: 28557614]
2. Belluco S, Avallone G, Di Palma S, Rasotto R, Oevermann A. Inter- and intraobserver agreement of canine and feline nervous system tumors. Vet Pathol. 2019;56(3):342–349. doi:10.1177/0300985818824952 [PubMed: 30663521]
3. Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. Am J Med. 2008;121(5):S2–S23. doi:10.1016/j.amjmed.2008.01.001
4. Chanut F, Kimbrough C, Hailey R, et al. Spontaneous cardiomyopathy in young Sprague-Dawley rats. Toxicol Pathol. 2013;41(8):1126–1136. doi:10.1177/0192623313478692 [PubMed: 23475560]
5. Cohen J. A coefficient of agreement for nomial scales. Educ Psychol Meas. 1960;20(1):37–46. doi:10.1177/001316446002000104
6. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull. 1968;70(4):213–220. [PubMed: 19673146]
7. Everitt BS, Landau S, Leese M, Stahl D. Cluster Analysis. 5th ed. John Wiley; 2011. doi:10.1002/9780470977811
8. Gamer M, Lemon J, Fellows I, Singh P. Various coefficients of interrater reliability and agreement. 2022. https://cran.r-project.org/web/packages/irr/irr.pdf
9. Hailey JR, Maleeff BE, Thomas HC, et al. A diagnostic approach for rodent progressive cardiomyopathy and like lesions in toxicology studies up to 28 days in the Sprague Dawley rat (part 2 of 2). Toxicol Pathol. 2017;45(8):1055–1066. doi:10.1177/0192623317743948 [PubMed: 29233079]
10. Jokinen MP, Lieuallen WG, Johnson CL, Dunnick J, Nyska A. Characterization of spontaneous and chemically induced cardiac lesions in rodent model systems: the national toxicology program experience. Cardiovasc Toxicol. 2005;5(2):227–244. doi:10.1385/CT:5:2:227 [PubMed: 16046796]
11. Lidbury JA, Rodrigues Hoffmann A, Ivanek R, et al. Interobserver agreement using histological scoring of the canine liver. J Vet Intern Med. 2017;31(3):778–783. doi:10.1111/jvim.14684 [PubMed: 28295598]
12. Marchevsky AM, Walts AE, Lissenberg-Witte BI, Thunnissen E. Pathologists should probably forget about kappa. Percent agreement, diagnostic specificity and related metrics provide more clinically applicable measures of interobserver variability. Ann Diagn Pathol. 2020;47:151561. doi:10.1016/j.anndiagpath.2020.151561
13. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med. 2012;22:276–282. doi:10.11613/BM.2012.031
14. Meroueh C, Chen ZE. Artificial intelligence in anatomical pathology: building a strong foundation for precision medicine. Hum Pathol. 2023;132:31–38. doi:10.1016/j.humpath.2022.07.008 [PubMed: 35870567]
15. Palazzi X, Barale-Thomas E, Bawa B, et al. Results of the European Society of Toxicologic Pathology Survey on the use of artificial intelligence in toxicologic pathology. Toxicol Pathol. 2023;51(4):216–224. doi:10.1177/01926233231182115 [PubMed: 37732701]
16. Plesser HE. Reproducibility vs. replicability: a brief history of a con-fused terminology. Front Neuroinform. 2018;11:76. doi:10.3389/fninf.2017.00076 [PubMed: 29403370]
17. Pospischil A, Folkers G. How much reproducibility do we need in human and veterinary pathology? Exp Toxicol Pathol. 2015;67(2):77–80. doi:10.1016/j.etp.2014.11.005 [PubMed: 25483119]
18. Shafi S, Parwani AV. Artificial intelligence in diagnostic pathology. Diagn Pathol. 2023;18(1):109. doi:10.1186/s13000-023-01375-z [PubMed: 37784122]
19. Steigen SE, Søland TM, Nginamau ES, et al. Grading of oral squamous cell carcinomas—intra and interrater agreeability: simpler is better? J Oral Pathol Med. 2020;49(7):630–635. doi:10.1111/jop.12990 [PubMed: 31899572]

20. Tokarz DA, Steinbach TJ, Lokhande A, et al. Using artificial intelligence to detect, classify, and objectively score severity of rodent cardiomyopathy. Toxicol Pathol. 2021;49(4):888–896. doi:10.1177/0192623320972614 [PubMed: 33287662]

21. Yap FW, Rasotto R, Priestnall SL, Parsons KJ, Stewart J. Intra- and interobserver agreement in histological assessment of canine soft tissue sarcoma. Vet Comp Oncol. 2017;15(4):1553–1557. doi:10.1111/vco.12300 [PubMed: 28133880]

22. Zuraw A, Aeffner F. Whole-slide imaging, tissue image analysis, and artificial intelligence in veterinary pathology: an updated introduction and review. Vet Pathol. 2022;59(1):6–25. doi:10.1177/03009858211040484 [PubMed: 34521285]

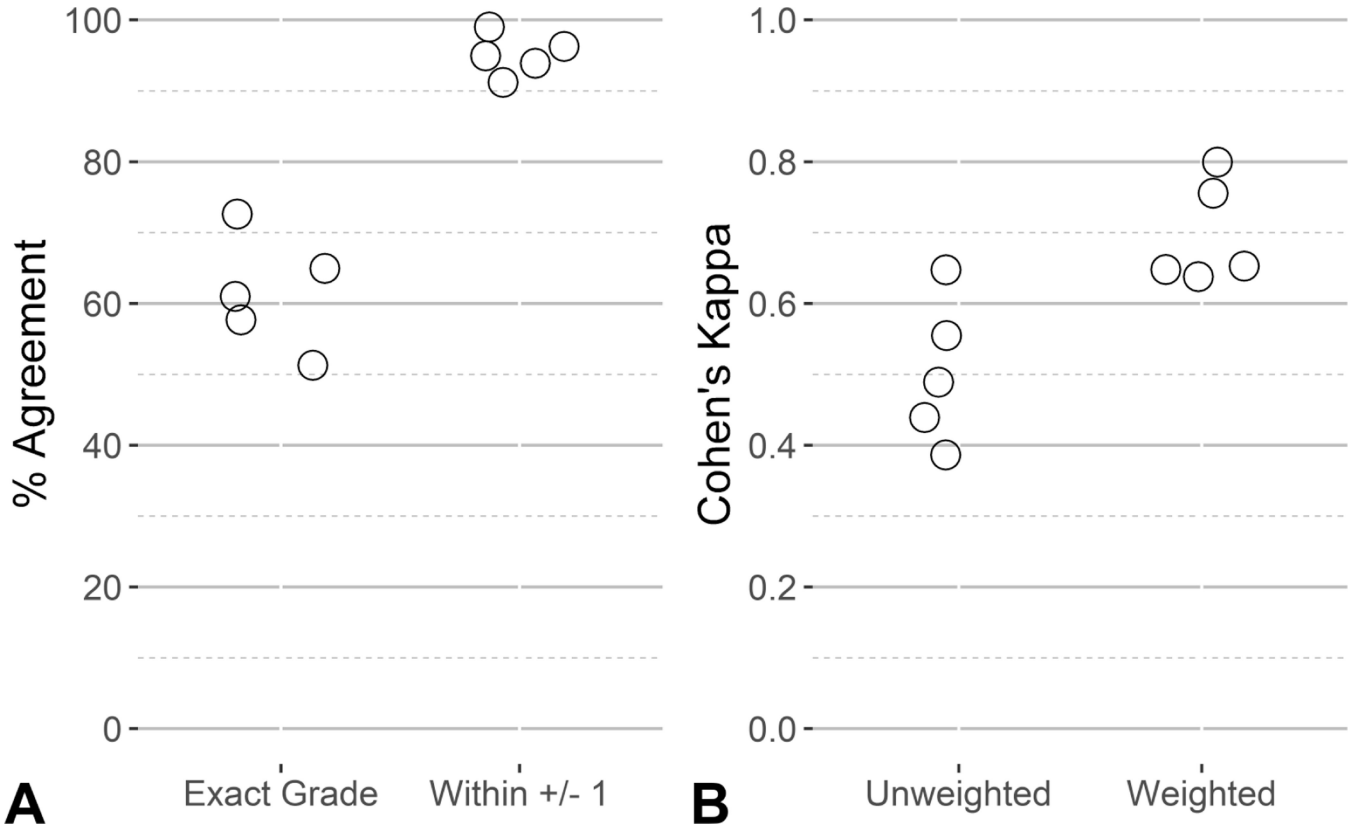

Figure 1.

Distribution of intra-rater reliability for the five veterinary pathologists using data from the 100 repeated slides. The points represent the level of agreement for each pathologist using accuracy (A) and Cohen's kappa (B). In each panel, values are shown for strict agreement (Exact Grade in panel A, Unweighted Cohen's kappa in panel B) and a tolerance for some margin of disagreement (within ± 1 grade in panel A, weighted Cohen's kappa in panel B).

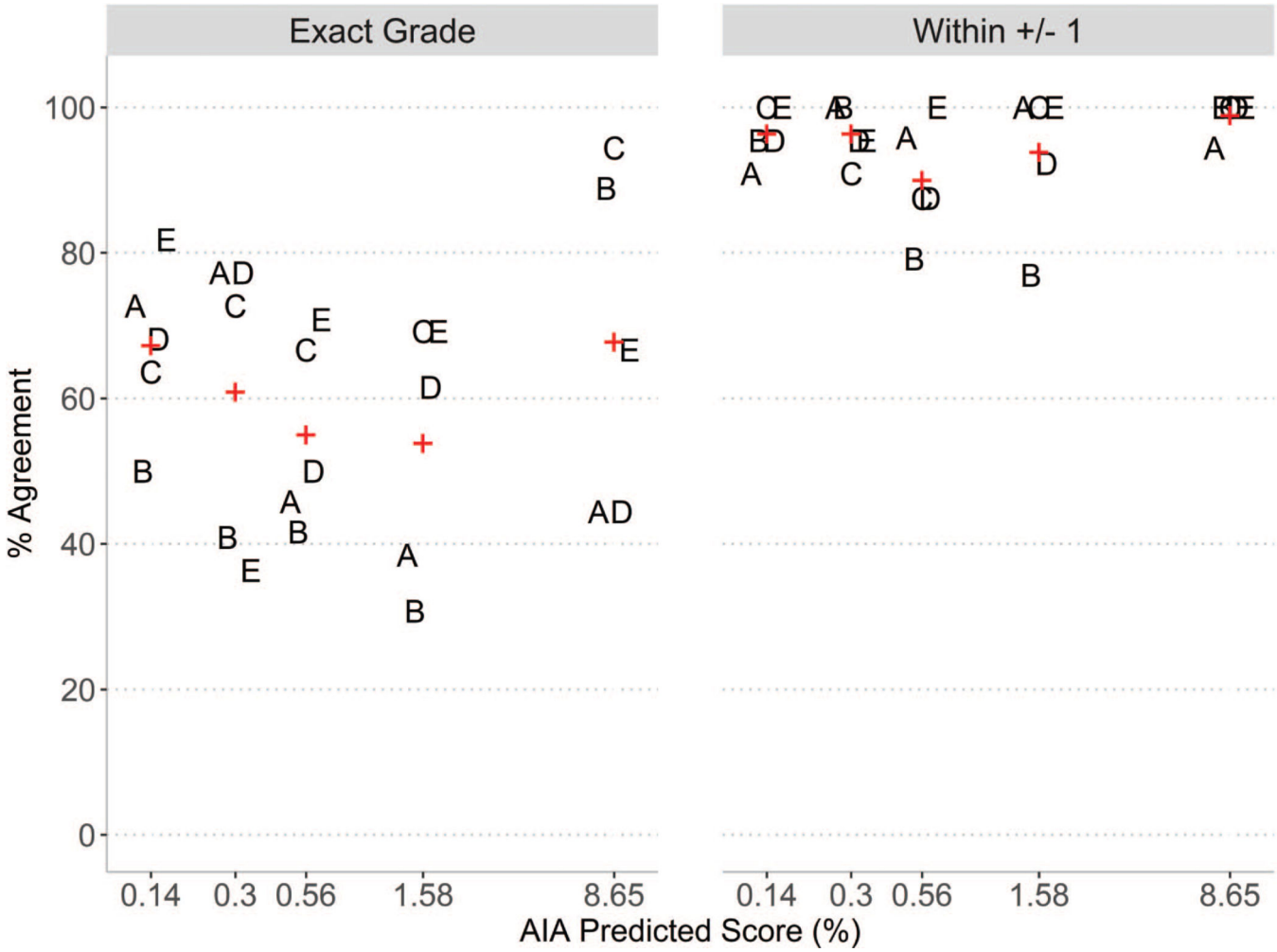

Figure 2.

Intra-rater reliability agreement for the five veterinary pathologists by quintiles of AIA scores using data from the 100 repeated slides. Values for each rater (A, B, C, D, E) are shown for percent agreement for each quintile of AIA predicted score, where the upper bound of each quintile interval is shown on the horizontal axis. Mean values are shown in red +.

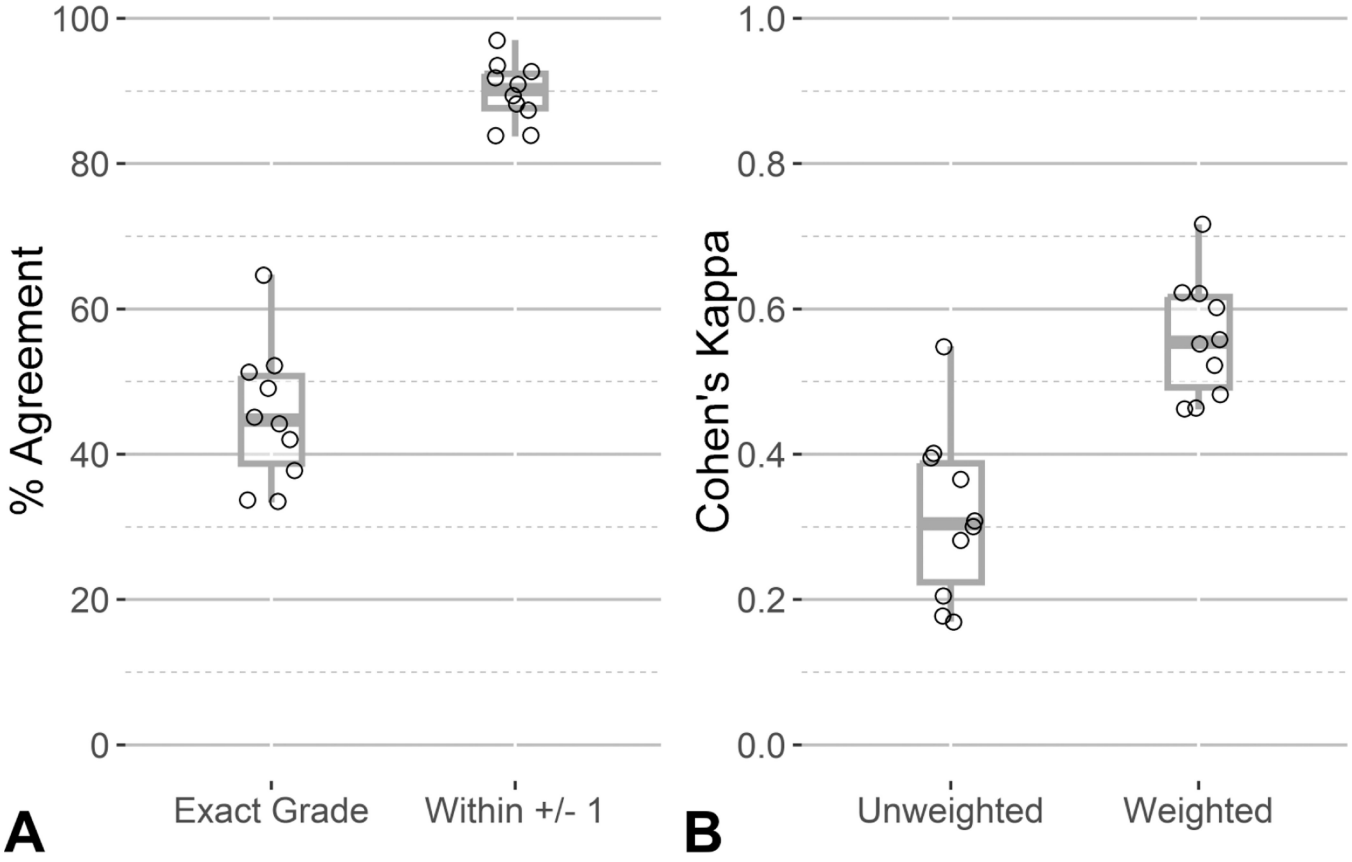

Figure 3.

Distribution of pairwise inter-rater reliability measures. The points represent the level of agreement between each of the ten rater pairs using percent agreement (A) and Cohen's kappa (B). In each panel, values are shown for strict agreement (Exact Grade in panel A, Unweighted Cohen's kappa in panel B) and a tolerance for some margin of disagreement (within ±1 grade in panel A, weighted Cohen's kappa in panel B). Boxplots are overlaid to show the distribution of the data.

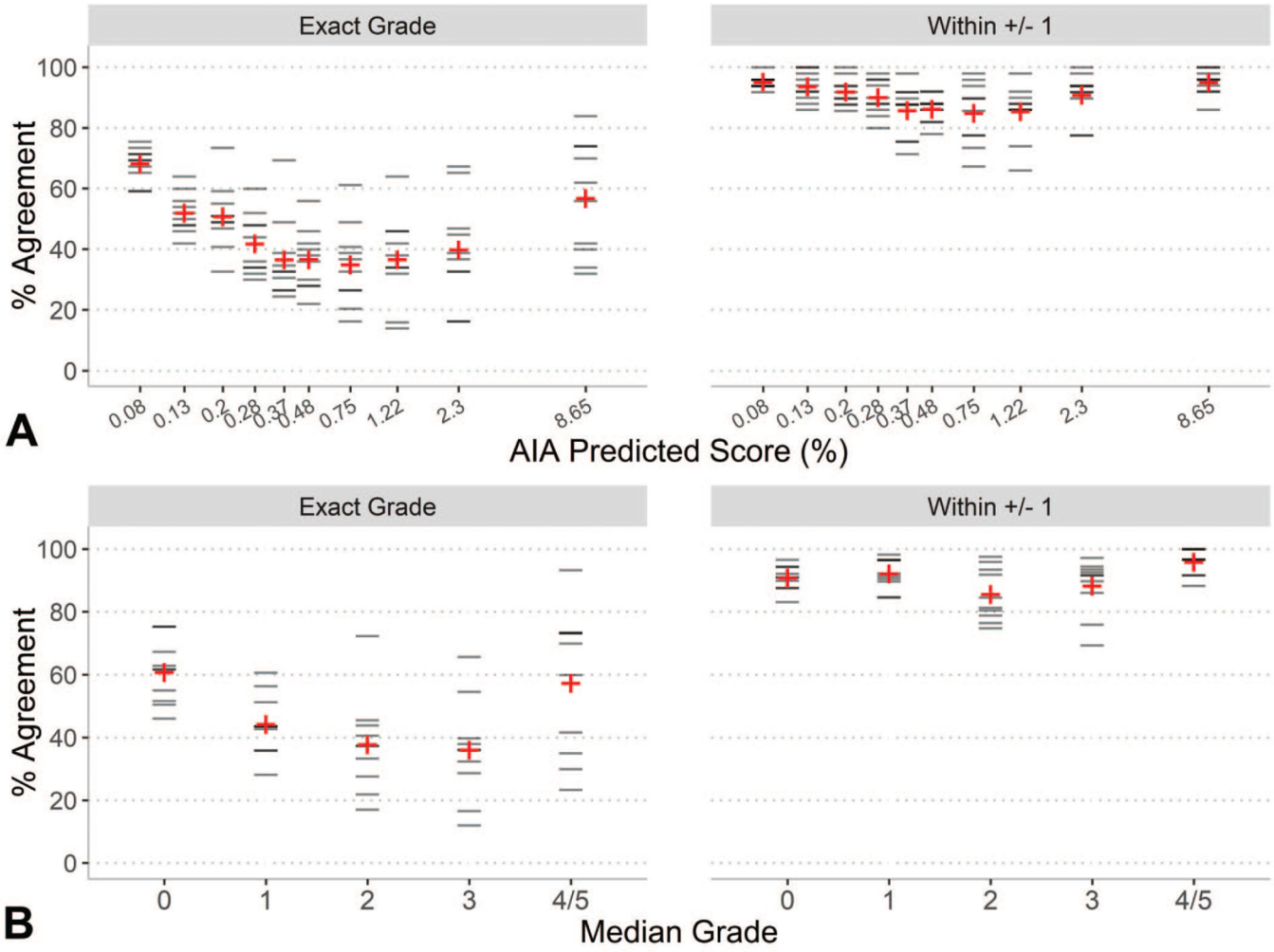

Figure 4.

Distribution of percent agreement across all pairs of raters. Horizontal lines represent the percent agreement between each of the ten rater pairs by deciles of AIA score (A) and by median grade severity (B). Mean values are shown in red +. In panel A, the upper bound of each decile interval is shown on the horizontal axis. In panel B, the median grade across all 5 raters is shown on the horizontal axis, with grades 4 and 5 combined.

Table 1.

Number of whole slide images used in the analysis of intra-rater and inter-rater agreement.

Study	Number of whole slide images	Analysis
Tokarz et al[20]	300 slides	Used a multinomial logistic regression model to relate AIA-predicted scores with the median cardiomyopathy severity score from a panel of 5 pathologists[20]
Current work	Select 100 (from Tokarz et al) via stratified sampling	Intra-rater reliability
	200	Model validation (Supplemental section)
	300 (from Tokarz et al) + 200 = 500 unique slides	Inter-rater reliability

Table 2.

Interpretation of Cohen's kappa.[6]

Kappa	Agreement
<0	Poor
0–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–0.99	Almost perfect