Interobserver variation among histopathologic evaluations of intestinal tissues from dogs and cats

Michael D. Willard, DVM, MS, DACVIM; Albert E. Jergens, DVM, MS, DACVIM; Robert B. Duncan, DVM, PhD, DACVP; Michael S. Leib, DVM, MS, DACVIM; Malcolm D. McCracken, DVM, PhD, DACVP; Robert C. DeNovo, DVM, DACVP; Rowland G. Helman, DVM, PhD, DACVP; Margaret R. Slater, DVM, PhD; Jacque L. Harbison, MS

Objective—To determine whether substantial interobserver variation exists among diagnostic pathologists for descriptions of intestinal mucosal cell populations and whether histopathologic descriptions accurately predict when a patient does not have clinically evident intestinal disease.

Design—Comparative survey.

Sample Population—14 histologic slides of duodenal, ileal, or colonic tissue from 10 dogs and 3 cats.

Procedure—Each histologic slide was evaluated independently by 5 pathologists at 4 institutions. Pathologists, who had no knowledge of the tissues’ origin, indicated whether slides were adequate for histologic evaluation and whether the tissue was normal or abnormal. They also identified the main infiltrating cell type in specimens that were considered abnormal, and whether infiltrates were mild, moderate, severe, or neoplastic.

Results—Quality of all slides was considered adequate or superior by at least 4 of the 5 pathologists. For intensity of mucosal cellular infiltrates, there was uniformity of opinion for 1 slide, near-uniformity for 6 slides, and nonuniformity for 7 slides. Five dogs did not have clinical evidence of intestinal disease, yet the pathologists’ descriptions indicated that their intestinal tissue specimens were abnormal.

Conclusions and Clinical Relevance—Substantial interobserver variation was detected. Standardization of histopathologic descriptions of intestinal tissue is necessary for meaningful comparisons with published articles. Clinicians must be cautious about correlating clinical signs and histopathologic descriptions of intestinal biopsy specimens. (J Am Vet Med Assoc 2002;220:1177–1182)

Histopathologic interpretation of intestinal tissues has achieved prominence in the diagnostic evaluation of human, canine, and feline digestive tract disorders, especially since inflammatory bowel disease has become a popular diagnosis. However, concern has been raised about the consistency with which tissue specimens are evaluated and the meaningfulness of evaluation in determining the cause of the patient’s illness. In the words of 1 pathologist, “our ability to interpret endoscopic intestinal biopsy specimens has lagged far behind our ability to obtain them.” Unfortunately, the diagnosis of inflammatory bowel disease remains largely subjective on the basis of evaluation of the cellularity of the lamina propria. If intestinal biopsy is going to be an important and prominent diagnostic tool, there must be assurance of reasonable consistency in results obtained by different pathologists examining the same tissue specimens. Investigators and clinicians will be able to compare their cases with published results only if such consistency is established.

In addition to consistency among pathologists, it is desirable that there be some correlation between the pathologic description of the tissue and the clinical state of the animal. In particular, it would be advantageous to know that one could anticipate receiving a histologic description of normal or near normal when tissues from an animal without clinically important disease were examined. Although many biopsies are performed to define a known disease process, many other biopsies are performed to determine whether a disease exists in a particular tissue.

Four of the authors (MDW, AEJ, MSL, RCD) had become concerned that substantial inconsistencies in interpretation of biopsy specimens appeared to exist among pathologists. They also believed that they had seen instances in which the pathologic description of an intestinal specimen suggested substantial clinical disease, although there was an absence of clinical signs that would be expected in patients with clinically important disease. Therefore, the purpose of the study reported here was to determine whether substantial interobserver variation exists among diagnostic pathologists for descriptions of intestinal mucosal cell populations and whether histopathologic descriptions predict when a patient does not have clinically evident intestinal disease.

Materials and Methods
Histologic slides—Fourteen histologic slides were retrieved from the files of the Department of Veterinary Pathobiology, Texas A&M University. All tissue specimens...
that had been obtained endoscopically were mounted on plastic histology cassette sponges with the mucosal surface upward, prior to fixation in neutral-buffered 10% formalin and routine processing. Tissues were sectioned at 6 µm and stained with H&E. Nine histology slides originated from dogs and cats referred to the Veterinary Teaching Hospital because of various signs of digestive tract disease. In most animals, the intestine was thought to be responsible for the clinical signs, although in some animals, the intestines were considered unlikely to be causing clinical signs. Five slides originated from research dogs without signs of intestinal disease that were used in a prior study of intestinal bacterial populations and mucosal cellular populations. Three slides were from endoscopic biopsies of canine duodenum, 4 were from endoscopic biopsies of canine ileum, 1 was from an endoscopic biopsy of canine ileum, and 1 was from an endoscopic biopsy of canine colon (2 slides were from different sites in the same dog). The slides used in the study were selected by 1 of the authors (MDW) by use of 2 criteria. The first criterion, which was used for all slides, was that the tissue section on the microscope slide appeared to be adequate in size and orientation to allow meaningful evaluation. For 5 slides (ie, slides 9 to 13), the second criterion was that the dogs from which the tissues originated did not have clinical evidence (ie, history, physical examination, laboratory findings, or imaging) of withholding of food, which at that time was considered to represent small intestinal bacterial overgrowth. However, that criterion for diagnosis of small intestinal bacterial overgrowth has come under criticism, and the number of bacteria that were found in these dogs' intestines is now considered by some investigators to be within reference limits.

Selection of pathologists—The 4 diplomates of the American College of Veterinary Internal Medicine (ACVIM) in the specialty of internal medicine involved in the study (MDW, A. E. J., M. S. L., R. C. D.) were clinicians who routinely evaluated patients with signs of digestive tract disease. Each ACVIM diplomate recruited at least 1 diagnostic pathologist who was subjectively considered skilled at small animal intestinal histopathology.

Pathologists—Each pathologist had earned a DVM degree and a PhD degree in pathology. Four of the 5 pathologists were diplomates of the American College of Veterinary Pathology, and all 5 were involved in veterinary diagnostic pathology. In some instances, the pathologist also worked at the same institution as the internist. In order to maintain the anonymity of the pathologists, no ACVIM diplomate knew the name of any of the pathologists recruited by the other ACVIM diplomats, and the pathologists were not given the names of the participating pathologists at the other institutions. Pathologists who participated in this study were given the option of remaining anonymous or of being coauthors on the manuscript. Two pathologists chose to remain anonymous.

Examination and evaluation of slides by the pathologists—The slides were sent out in groups of 2 to 6 slides, and each pathologist received the slides in the same order as the other pathologists. No information regarding signalment, history, physical examination, laboratory findings, or imaging was made available to the pathologists. To accurately compare evaluations among pathologists, a form was developed for recording the pathologists' findings. Each pathologist had to first decide whether the histopathology slide was of inadequate, adequate, or superior quality regarding suitability for diagnostic purposes. Next, the pathologist was asked whether tissue on the slide was normal or abnormal. If the tissue was considered abnormal, the type of infiltrating cell and severity of the mucosal infiltrate were described by use of the terms on the form. The standardized terms (ie, normal, mild, moderate, severe, and neoplastic) that the pathologists were required to use in describing the cellular population of the intestinal mucosa are commonly used by diagnostic pathologists for describing or grading intestinal mucosal cell populations. The term neoplastic was used specifically in reference to lymphosarcoma.

No criteria for these terms were supplied or suggested to the pathologists; instead, pathologists were asked to categorize the cellular infiltrates in these tissues in the same manner they did for clinical cases. After evaluating the slide and describing the cellular infiltrate, the pathologist completed the form, and the slides were returned to the ACVIM diplomate who forwarded the slides to the next ACVIM diplomate (who gave the slides to the pathologist that he had recruited) and sent the evaluation form to the coordinator (MDW).

Evaluation of results—Results were tabulated in chart form. Data regarding infiltrating cell type were analyzed by inspection, and the number of slides in which there was uniformity of opinion (ie, all the pathologists agreed on the primary infiltrating cell type) was compared with the number of slides for which opinions were nonuniform (ie, at least 1 pathologist disagreed about the primary infiltrating cell type). If a slide was considered normal, an infiltrating cell type was not reported by the pathologist. However, when judging uniformity of opinion regarding cell type, normal was not used as a basis for categorizing the pathologists' opinion as nonuniform.

Data regarding intensity of cellular infiltration were analyzed in 2 ways. First, the 5 pathologists' descriptions of the severity of the mucosal infiltrate for each slide were classified by the authors as being uniform, near-uniform, or nonuniform. A classification of uniform required that all pathologists described the mucosal infiltrate using the same term (eg, mild). A classification of near-uniform was given if all pathologists described the mucosal infiltrate using the same term (eg, mild and moderate). A classification of nonuniform was given when at least 1 pathologist's description of the cellular infiltrate was not adjacent to another pathologist's description. These data were also analyzed by use of a statistical measure of agreement.

Statistical analysis—For data regarding severity of cellular infiltration, agreement between each possible pair of pathologists was measured for each of the 14 histology slides. Kappa and weighted K were calculated, and the adjacent grades of normal and mild were considered equivalent classifications. For the weighted K calculation, the computer program assigned weights to the different descriptions of the pathologists. Kappa values < 0 were considered evidence of more disagreement than would occur by chance alone (ie, disagreement), values ≥ 0 but ≤ 0.4 were considered poor agreement, values > 0.4 but ≤ 0.73 were considered fair to good agreement, and values > 0.73 were considered strong agreement.

Kappa values were calculated by use of commercially available software. Kappa is a quantitative measure of agreement that allows objective comparison of the agreement between 1 pair of individuals with that present between other pairs of individuals.
Results

Adequacy of the histology slides for evaluation was assessed as follows: 45 of 70 evaluations were adequate, 22 were superior, 2 were inadequate, and 1 evaluation was missing. Each of the 14 slides was evaluated as being adequate or superior by at least 4 of the pathologists. The 2 slides that were considered inadequate only received this evaluation from 1 pathologist, whereas at least 4 other pathologists considered the same slide to be adequate or superior for examination. Thus, no slide had 2 evaluations of being inadequate for diagnostic purposes.

Uniformity of opinion regarding the principal infiltrating cell type was found for 6 of 14 slides. There was nonuniformity of opinion for 8 of 14 slides.

For severity of the cellular infiltrate, 1 slide yielded uniformity of opinion; this slide was considered by all pathologists to be abnormal and to have a mild infiltrate. For 6 slides there was near-uniformity of opinion. For 7 slides there was nonuniformity of opinion (Fig 1). There was no consistent pattern among pathologists regarding which pathologist characterized cellular infiltrates as more or less severe than the other pathologists did.

Examination of the κ values revealed that pathologist 1 had more disagreement with pathologists 3 and 4 than would be expected from chance alone (unweighted κ = 0.43 and 0.15, respectively; weighted κ = 0.35 and 0.08, respectively). Pathologists 3 and 4 had fair to good agreement (unweighted κ = 0.48; weighted κ = 0.53), whereas the agreement between pathologists 1 and 2 was considered fair to good if the weighted κ was calculated (0.42) but not if an unweighted κ was calculated (0.26). All other comparisons revealed poor agreement, as indicated by κ values ≥ 0 but ≤ 0.40.

There were only 2 slides for which there was uniformity of opinion with regard to the type of infiltrating cell as well as uniformity or near-uniformity of opinion regarding the severity of the mucosal infiltrate. There were only 3 slides for which there was nonuniformity of opinion regarding both the type of infiltrating cell and the severity of the mucosal infiltrate. For the other 9 slides, there was either nonuniformity of opinion regarding the major infiltrating cell type or nonuniformity regarding the severity of the infiltrate, but not both.

For 5 slides, there was discordance between the clinical findings and the pathologists’ descriptions of the mucosal specimens. The tissues on 5 slides were described as abnormal in 22 of 25 evaluations, although these animals had no clinical evidence of intestinal disease. Two specimens were full-thickness ileal specimens, 2 specimens were full-thickness duodenal specimens, and 1 specimen was an endoscopically obtained specimen of colonic mucosa. For 1 of these slides, only 1 pathologist described the mucosal infiltrate as being of moderate intensity; however, for the other 4 slides, at least 2 pathologists described the infiltrate as being abnormal and of at least moderate intensity. Of the remaining 22 descriptions of the intestinal mucosal infiltrate in these slides, 8 indicated that lesions were mild, 9 indicated that lesions were moderate, 3 indicated that lesions were severe, and 2 indicated that lesions were neoplastic.

Discussion

We concluded that not only was there substantial variation in opinion regarding the intensity of intestinal mucosal infiltrates but that this variation was unpredictable. We did not attempt to determine whether pathologists could use a particular set of criteria and be uniform in their descriptions of tissues, because there is no universally accepted set of criteria for evaluating intestinal mucosal tissue specimens. We did not attempt to determine the uniformity or accuracy of pathologists in diagnosing inflammatory bowel disease, because presently it is not clear what constitutes the gold standard for the histologic diagnosis of inflammatory bowel disease; various criteria have been proposed. Instead, we attempted to determine whether there was uniformity or nonuniformity in the way in which different pathologists used the terms normal, mild, moderate, severe, and neoplastic to describe intestinal mucosal cellular infiltrates, because these terms are commonly used by pathologists when describing histopathologic changes in intestinal tissue specimens.

For the study to have maximum clinical applicability, we did not make any suggestions to the pathologists regarding what these terms should mean. The pathologists were to use these terms as they would on any clinical case. The only aspect of the histologic description that was standardized was our request that pathologists use the same terms so that meaningful comparisons among pathologists could be made. We found that there was substantial interobserver varia-

![Figure 1](image_url)

Figure 1—Intensity of mucosal cellular infiltrate in 14 histologic slides of intestinal tissue, as estimated by 5 pathologists (each represented by a separate bar). *Specimens were from clinically normal research dogs. #Specimens had nonuniformity of opinion regarding the intensity of the mucosal cellular infiltrate.
tion among these 5 pathologists, as indicated by simple inspection of the data and use of the $\kappa$ statistic.

We tried to design a study that would maximize the pathologists’ opportunities to be consistent with each other, while still allowing us to determine correlation between clinical appearance and histopathologic descriptions. We attempted to do this in 3 ways. First, we sought to use histology slides that were of adequate diagnostic quality. Although the quality of the histology slide should theoretically make no difference when assessing interobserver variation, we tried to use excellent quality slides. Most of the pathologists described the slides as being of adequate diagnostic quality; in fact, 22 evaluations were superior.

Second, when deciding whether the opinions regarding the infiltrating cell type were uniform or nonuniform, normal was not used as a basis for categorizing the opinions as being nonuniform. In this way, we made it more likely that a slide would be categorized as having a uniform opinion, as opposed to a nonuniform opinion. In retrospect, we should have asked the pathologists to decide on the primary cell type in the mucosa in the slides of normal tissue as well.

Third, calculation of $\kappa$ was done in a manner to provide the pathologists with the opportunity to have as much agreement as possible. For calculation of $\kappa$, we considered the descriptions normal and mild as equivalent, because we arbitrarily decided that it could be difficult to distinguish normal from mild, and many experienced clinicians seemingly consider mild as almost equivalent to normal. An unweighted $\kappa$ was used for information treated as nominal categoric data. When the data were treated as ordinal categoric data, a weighted $\kappa$ was more appropriate. Weighted $\kappa$ takes into account partial agreement by assigning weights to cells on the basis of degree of disagreement; thus, a disagreement between normal-mild and severe would be weighted more heavily than a disagreement between normal-mild and moderate. Therefore, we calculated both unweighted and weighted $\kappa$ to ensure that if there was a lack of agreement, it would not be caused by the form of statistical analysis we chose. There was only 1 instance in which calculating a weighted $\kappa$ resulted in a value that had a different interpretation than the unweighted $\kappa$.

Furthermore, when assigning descriptions to the different values of $\kappa$, we biased the descriptions of our values so that it would be easier to have better agreement. This bias is evident by comparing our descriptions of various values of $\kappa$ to the descriptions used by investigators in human diagnostic gastroenterology. Geboes et al\(^{20}\) stated that values $\leq 0.5$ were considered poor, and values from 0.51 to 0.6 were only considered moderate. If we had used similar descriptors for our values of $\kappa$, 19 of 20 values would have been considered poor, instead of the 17 of 20 that we report.

A second question was whether the histologic description of intestinal tissue from an animal that appears clinically normal would reflect that clinical condition. We asked this question for 2 reasons. First, we believed that with prolonged daily observation, we could state with confidence that a dog consistently appeared clinically normal. A second and perhaps more important reason was that inability to discern that an intestinal tissue specimen was from an animal without clinical signs of disease seemed more important to us than failure to find evidence of disease in a tissue specimen from an obviously diseased patient. Whenever intestinal biopsies are performed on a patient with intestinal disease, it is understood that there is no guarantee that the affected area of the intestinal tract will be sampled. Intestinal diseases may be segmental and of varying severity in different parts of the intestines; one cannot assume simply because there is intestinal disease severe enough to cause clinical signs that performing a biopsy at a given site will procure tissue that is abnormal.\(^{15,22}\) In fact, normal intestinal mucosa can be found in dogs dying of intestinal disease.\(^{15}\) Therefore, failure to find evidence of disease in a specimen of intestinal tissue should not be especially confusing to the experienced clinician.

In contrast, finding changes suggestive of clinically important disease in an intestinal tissue specimen has important implications. Intestinal biopsy is seldom performed unless the intestine is considered a possible cause of the animal’s disease; if there are apparently important changes in an intestinal tissue specimen, these changes are often understood by the clinician to imply the presence of intestinal disease requiring treatment. However, the main clinical sign resulting in biopsy may be weight loss or anorexia, signs which can be caused by disease in many organs besides the intestine. Thus, it is possible that a substantial number of intestinal biopsy specimens submitted to laboratories are from intestines that are not responsible for the patient’s clinical signs. Therefore, it was important to determine whether histopathologic results might suggest intestinal disease in dogs known to not have clinical signs of intestinal disease.

There are 3 aspects of the study that could potentially be criticized. First, we did not give the pathologists any information regarding the animal from which the tissues were obtained. Withholding such data should not have any more effect on the interobserver variation than is encountered clinically. Identical pieces of tissue could be described differently if they were known to have come from different sites (eg, duodenum vs jejunum); however, specimens are often submitted to pathologists with incomplete information regarding site of origin. Thus, our pathologists labored under the same conditions that many diagnostic pathologists must routinely endure.

The second potential criticism is that we used tissue specimens from 3 sites in the digestive tract, specimens were obtained via 2 methods, and specimens came from 2 species. This design was used because we wished to screen for whether substantial interobserver variation or discordance existed in the histopathologic descriptions. If substantial variation was found, then future studies could be designed to determine precisely what caused the variation.

The third potential criticism is that we had a modest number of specimens; nevertheless, we believe that the nonuniformity revealed by our study was real and worthy of publication. Nonuniformity of opinion among pathologists was detected in 7 of the 14 slides,
which is an unexpectedly large number, even considering specimen size. Similarly, the \( \kappa \) values were overwhelmingly consistent with poor agreement (17 of 20 values were \( \leq 0.4 \)). These results suggest that it presently would be difficult or impossible to accurately compare pathology reports of intestinal tissues written by different pathologists, unless perhaps they were known to be using the same grading system. The amount of interobserver variation was particularly concerning for full-thickness surgical biopsy specimen from the ileum of a clinically normal dog, which received 5 descriptions ranging from normal to neoplastic.

The discordance between the abnormal histopathologic descriptions of 5 tissue specimens and the normal clinical appearance of the dogs from which they came (Fig 1) could be attributable to 1 of 2 situations. First, there may be a lack of awareness regarding the range of histologic appearances of intestines from clinically normal animals. Second, the pathologists may be seeing changes in the mucosa that are real but not extensive enough to cause clinical disease. These dogs may have had subclinical disease that caused histologic abnormalities for which they were compensating. This phenomenon has been reported in humans, and there is no reason to believe that it does not occur in dogs and cats. Therefore, finding that the tissues from clinically normal-appearing dogs were described as abnormal could mean that the pathologists were wrong in their descriptions, or it could mean that our knowledge of histologic findings in intestinal tissue in dogs that appear clinically normal is deficient. Regardless, unnecessary and inappropriate treatment could have easily been prescribed for these animals if they had been evaluated for clinical illness. Practicing veterinarians must realize that even finding apparently real lesions in a biopsy specimen does not guarantee that the lesion is responsible for the clinical signs.

The finding that 2 pieces of ileal tissue from clinically normal dogs were described as having neoplastic infiltrates (ie, lymphoma) is especially interesting. Only ileal tissues were so seriously misjudged. Furthermore, of the 5 tissue specimens from clinically normal dogs, only ileal tissues were described as having severe infiltrates. These findings strengthen the argument that the effects of tissue, method of specimen acquisition, and species on interobserver variation should be investigated.

The interobserver variation and discordance seen in this study must be put into proper perspective. Histologic evaluation of the digestive tract is acknowledged to be difficult. Interobserver variation between pathologists has been described in human pathology including gastrointestinal pathology, in which there are sometimes blatan differences in diagnostic criteria. Taken in this light, the uniformity or near-uniformity of results in 7 of 14 slides in our study is encouraging. The purpose of this article was not to be accusatory towards pathologists but rather to point out the need to rigorously define the terms that are used to describe intestinal tissue and attempt to correlate them with clinical status, outcome, or both. Prior work in human pathology indicates that having clearly defined criteria improves \( \kappa \) values substantially (ie, from the range of poor agreement up to good agreement). Diagnosis is best accomplished as a joint effort between clinician and pathologist.

References


