Published in final edited form as:
Vet Pathol. 2013 November ; 50(6): . doi:10.1177/0300985813485099.

Principles for valid histopathologic scoring in research

Katherine N. Gibson-Corley, **Alicia K. Olivier**, and **David K. Meyerholz**
Division of Comparative Pathology, Department of Pathology, University of Iowa, Carver College of Medicine, Iowa City, IA

Abstract

Histopathologic scoring is a tool by which semi-quantitative data can be obtained from tissues. Initially, a thorough understanding of the experimental design, study objectives and methods are required to allow the pathologist to appropriately examine tissues and develop lesion scoring approaches. Many principles go into the development of a scoring system such as tissue examination, lesion identification, scoring definitions and consistency in interpretation. Masking (a.k.a. “blinding”) of the pathologist to experimental groups is often necessary to constrain bias and multiple mechanisms are available. Development of a tissue scoring system requires appreciation of the attributes and limitations of the data (e.g. nominal, ordinal, interval and ratio data) to be evaluated. Incidence, ordinal and rank methods of tissue scoring are demonstrated along with key principles for statistical analyses and reporting. Validation of a scoring system occurs through two principal measures: 1) validation of repeatability and 2) validation of tissue pathobiology. Understanding key principles of tissue scoring can help in the development and/or optimization of scoring systems so as to consistently yield meaningful and valid scoring data.

Keywords

Grading; Histopathology; Lesions; Ordinal; Semi-quantitative; Scoring; Validation

INTRODUCTION

Through the course of investigation, research laboratories often submit of tissues to histopathology cores for tissue processing and examination by a pathologist.[11,27,48] Pathologists provide morphologic assessment of these tissues including examination for group-specific differences. Many times, there is a need for more rigorous evaluation of the tissue either to prove a group difference or substantiate the observations of the initial examination.

Scoring (a.k.a. “grading”) is a tool that can be used to derive data from biologic systems (e.g. tissues) for analysis and group comparisons. Scoring can be applied at different levels of tissue examination including antemortem imaging techniques[6,35,54], postmortem macroscopic examination[18,36,68] and histopathologic examination.[17,39,46,69] Crissman and colleagues suggested that a scoring system should exhibit three fundamental characteristics: (1) it should be definable, (2) it should be reproducible and (3) it should produce meaningful results.[12] This paper reviews key principles for the development of scoring systems so that the pathologist has the best opportunity to meet these key principles. Importantly, these fundamental principles of scoring tissues are applicable to most organs, tissues, and models systems.

Corresponding author: David K. Meyerholz, DVM, PhD, Carver College of Medicine, University of Iowa, 500 Newton Road, 1165ML, Iowa City, Iowa 52242, Phone: (319)353-4589, Fax: (319)335-8453, david-meyerholz@uiowa.edu.

METHODS

This paper describes key principles for the development of semiquantitative scoring systems via histologic examination. Even so, these same concepts can be useful for development of semiquantitative scoring systems in other research contexts such as commercial immunohistochemistry kits, serologic assays or applications of specialized software packages. Notably, it is beyond the scope of this paper to address principles for quantitative techniques and applications.

All experimental data in the figures and tables of this paper were created to demonstrate important principles associated with scoring. Experimental data were constructed to replicate situations that are commonly encountered by comparative pathologists in academia and emphasis was selectively placed on histopathology-based examples. Importantly, these examples of scoring methods were simplified in scope and complexity for ease of understanding the basic concepts. Representative analyses were made for each example scoring method, but these should not be taken as exclusive statistical options. All statistical analyses and graphs were made using Prism software (GraphPad Software, Inc. La Jolla, CA).

Perspective

Sound methodology in histopathologic scoring is important to detect biologic differences in treatment groups. Importantly, it does not compensate for poor experimental design or improperly sampled tissues that occur “upstream”. Many papers have been submitted to journals (but not necessarily published) in which the sampling and histopathologic scoring approaches were robust in nature, but the experimental designs were markedly flawed. In these cases, even when statistically significant data could be generated by the authors it was without context and lacked validity for proper interpretation. A simple proverb states “junk in, junk out”. Experimental background should be sought out for projects where tissues are submitted for pathologist examination. Proper perspective begins early and many objectives need to be considered. As described below, developing a sound experimental design, understanding the purpose of the study and considering how best to sample the appropriate tissues are all important features of perspective.

Experimental Design—Experimental planning and design are necessary for the development of a sound scientific study and understanding these methods are essential for context of proper data interpretation.[3,4,12,57,75] Species, strain, sex, age, appropriate controls, method/type of genetic manipulation, microbial status of colony, tissue handling and treatments (type dose, route, duration, etc.) all play a role in the evaluation and eventual interpretation of the data. Ancillary data such as clinical chemistries, imaging and/or clinical behavior can further give relevant insights for effective tissue evaluation. For example, if hepatocellular-specific enzymes were elevated in a treatment group then targeted sampling and examination of the liver would be valuable.

Study Objectives—Understanding the study objectives is useful in effective tissue examination and development of a meaningful scoring system. For example, a murine study of *Pseudomonas aeruginosa* infection may demonstrate antemortem group differences in the extent of neutrophilic lung inflammation based on routine examination.[40] A scoring system may be readily applied to corroborate this observation, which would be sufficient for many studies. However, if the study’s objective was to determine if neutrophil transmigration into the lungs was defective, then a scoring system that focuses on neutrophil transmigration might be developed, if possible, to more meaningfully demonstrate this mechanistic change.

Tissue Sampling—Sampling of tissues can greatly influence the diagnostic or treatment-related results of a study.[5,7,37] For example, in some strains of mice islet numbers can vary widely between pancreatic lobes,[32] therefore consistent tissue collection should be performed for optimal islet assessment. In academia, tissues are sometimes collected by the collaborator lab and stained slides submitted to the pathologist for examination. Awareness of the collection method as well as the level of consistency in sampling and sectioning helps to assure that unintentional bias is prevented.[8]

Principles for scoring

To determine an appropriate histologic scoring system for any tissue, key principles should be considered. Although this list is not exhaustive, these considerations will help to develop a useful scoring method.

Masking—An important goal for any experimental study is to constrain biases that can skew the final data and conclusions.[59] Bias can be introduced into any stage of the experimental project.[49,62] "Masking" (a.k.a. blinding) of the pathologist to experimental groups/treatments is a means of preventing bias from entering into the examination and scoring of tissues. Lack of masking can lead to unintentional observational bias that can often exaggerate treatment effects.[15,56] Different levels of masking for the pathologist can be implemented (Table 1), but consideration of the study goals as well as the limitations of the masking method need to be discussed before examination.

Examination—A thorough examination of all tissues/slides provides a context for scoring tissue lesions. For example, a lesion common to all groups could be indicative of a "background" lesion and scoring of this lesion parameter could be of little meaning to the study. But sometimes in the context of a research study, subtle differences in the frequency or severity of the "background" lesion may be indicative of a mechanistic change related to treatment and can be further assessed.[58] A review of the study objectives and the relevant literature may predict differences in specific lesion parameter, that could then be examined and scored to provide context for the current model.

Lesion parameters—What types of lesions can be studied by a scoring system? If lesions are identifiable in tissues, then these can often be applied into a scoring system (Table 2). Some lesions may be detectable in any tissue (e.g. cellular inflammation), whereas other lesion parameters may be specific for the organ/tissue (e.g. cholestasis in liver) being scored. While it is not feasible to concisely review all lesion parameters for all tissues, numerous approaches to scoring for specific organs or models can often be found in a targeted literature search.

Scoring definitions—Scoring systems often segregate samples into defined categories. It is useful to have clear language both characterizing and setting boundaries for each category.[58,67] Exclusive use of vague terms, such as "mild", "moderate", or "severe" in ordinal scoring can reduce interobserver repeatability and may even compromise intraobserver repeatability over time. Whenever possible, specific terminology including the use of percent of tissue affected can enhance the repeatability as well as sensitivity of the system.

Interpretation consistency—"Diagnostic drift" is a situation when the assignment of scores may vary slightly in consistency through the scoring process. This can happen in situations where there are a large number of samples; multiple pathologists examine subsets of tissues; slides are examined over a long period time; or when category characteristics/boundaries are poorly defined.[13] In research settings, it is most useful to have one

pathologist score the slides in a reasonable period of time, if applicable, to provide for additional consistency.[12,13] Of course, this approach is not always possible and review (by the same or a secondary pathologist) at the conclusion of the study may be warranted especially for more arduous studies.

Examples of Scoring Approaches

Types of data measures—Many years ago, Stevens wrote a paper describing four key types of measurement scales used in research: nominal, ordinal, interval and ratio[65] (Table 3). Generally speaking, nominal and ordinal scales produce qualitative data, whereas interval and ratio scales produce quantitative data. Qualitative data is that which approximates or characterizes something as opposed to quantitative data which measures something. For instance, biologic data that are acquired from morphometry have a ratio scale with a true zero point and produce quantitative data; relevant examples include length (e.g. acinus diameter) or area (e.g. acinus area). In contrast, nominal and ordinal scales, which are commonly used in scoring systems, produce qualitative data, thus any scoring is consider "semi-quantitative" in nature. Understanding the types of data as well as their constraints helps in their analysis.

There are multiple approaches to score tissues and common scoring methods for pathologists are highlighted below. For simplicity, these methods have been generally assigned into three groups for enhanced understanding and application. The reader would be advised that for additional information, other resources may be useful.[13,31,58,74]

Incidence method—This approach records the case incidence of a lesion (i.e. those affected) in an experimental cohort.[31,64] Similar types of scoring methods include binomial scoring (presence or absence of lesion) and percent affected. Lesions are defined by categories (i.e. nominal data) and recorded in a contingency table. For example, the trachea can be examined for the presence or absence of inflammation in submucosal glands (Table 4). These nominal data can be reported as a contingency table (Table 4) or shown as a graph for publication (Figure 1).

Ordinal method—The ordinal method is commonly used by many pathologists for lesion scoring and important principles for the method are discussed below.

This method assigns data into defined categorical groups that are arranged in an "ordered" progression in lesion severity.[65] For example, a scoring system can be based on the estimated percentage of the tracheal wall which is affected by a lesion; in this case a score (0–4) may be assigned (Table 5). The most common approach to ordinal scoring is to assign a summary score for each animal based on the tissue examination. An example of this can be seen in Table 6 where tracheal inflammation and hyperplasia are scored.

Another method found in the literature is to count several fields of tissue (e.g. ten random 400x fields) for each animal, each field scored and a mean (i.e. average) score assigned for the whole tissue of that animal. The problem with this approach is that the *mean* represents a measure of central tendency that is only appropriate for interval and ratio data. For ordinal data, the *median* is the most appropriate measure for central tendency. This statistical axiom is not without some controversy and it is not within the scope of this paper to resolve it.

Scoring approaches vary between pathologists. Many times, a tissue will have multiple lesions that can be assigned scored. Dependent on their approach to these situations, pathologists have been described as either "lumpers" or "splitters".[74] "Lumpers" use multiple parameters or anatomic sites to define each ordinal level. For example, multiple separate renal lesions associated with acute tubular injury are grouped together to give a

single scoring system (Table 7). On the other hand, “splitters” separate each parameter or anatomic site for scoring purposes. As opposed to the “lumpers”, “splitters” assign each specific renal lesion was assigned its own appropriate scoring system (Table 8). “Lumper” methods can be more efficient for the pathologist saving labor and time when groups have overt differences; however, “splitter” methods are more sensitive to parameter specific or sequential changes that may occur in a model and also have more repeatability.[13]

When modifying or developing a new ordinal scoring system, it is useful to evaluate the variance of lesion severity in all samples so as to “fit” the scoring system into the range of lesions. For example, if an infection model is studied at day 2 post-inoculation, the range of lesions may be entirely different than those previously studied at day 6 post-inoculation. If this adjustment is not done then the scoring system may be so skewed as to be ineffective for assessment of group differences at the different time point.

The number of score categories within the ordinal method has potential implications for the study and this ranges from as few as three to as many as ten or more per system.[29,33,42,58] A small number of score categories (e.g. three) can reduce the sensitivity of the scoring system so that more animal numbers (or more severe group differences) are required to detect a real biologic difference between groups. Alternatively, a large number of ordinal scores may cause difficulty in score assignment as there is often less obvious distinction between categories. This means that a scoring system with a large number of categories is prone to have reduced repeatability. It has been suggested that ~4–5 score levels may be an optimal range to maximize detection and repeatability.[58,67]

Ordinal scores are most commonly derived from direct evaluation of tissues with assignment of scores by the observer; however, transformation of quantitative data to ordinal scores has been described and is another source of ordinal scores.[19,40] Transformation of data can be a useful tool to constrain sample variance that is often found in animal-based research.

Rank method—The rank (“ordering”) method is not commonly used by pathologists, but it is simplistic in application.[31,64] This method is remarkably similar to what pathologists do (subconsciously) in their routine tissue evaluations. Samples from the treatment groups are combined and then ranked from most severe to least severe (or vice versa) and the rank number for each sample is used for analysis (Figure 2). While the ranked method is conceptually straightforward in application, it may be more labor intensive with a larger sample numbers.[74]

Statistics

Key components of statistical analysis are important in any research project. A biostatistician should collaborate with researchers for routine planning of experimental design through analyses of their data.[21] Access to a user-friendly statistical software package can also be useful for routine analyses and synthesis of graphs for publication. The use of scored data may be discipline dependent as scoring and its analyses are common for pathologists at academic and medical institutions, but recent INHAND recommendations suggest that toxicologic pathologists should rely on their morphologic interpretation preferentially over statistical inference of scoring.[45]

Choosing the appropriate statistical test is an important component for every experimental data set. Statistical tests have “assumptions” on which they function, and if an assumption is not applicable for the data being examined, then the validity (and interpretation) of the statistical approach may be in question. For example, ordinal data do not meet the assumption of a normal (Gaussian) distribution. Parametric analyses (e.g. Student’s T-test) should not be used to analyze ordinal data, but rather nonparametric analyses (e.g. Mann

Whitney test) should be considered.[59,61] Misuse of statistical analysis in research is recognized[59,66] and accordingly it is not uncommon for ordinal scoring data to be analyzed by inappropriate parametric tests (e.g. Student's T-tests). Increasingly, these types of inappropriate statistical analyses are being identified at submission of peer-reviewed papers causing mandatory statistical revision or manuscript rejection. For a broader perspective on statistical analysis of data, the reader is encouraged to examine these resources.[20–22,30,31,41,59–61]

Validation

Scoring methods should be designed to be reproducible as well as a meaningful analysis of data, i.e. a valid scoring system. But how does one know that a scoring method is valid? Validation mechanisms have been used in many tissue-specific scoring systems.[36,50,70,73] Validation can be summarized as two basic approaches: that of validating observer repeatability and that of validating tissue pathobiology.

Validation in repeatability—Recent reports in have highlighted the importance of repeatability in research.[1,52] For instance, Begley and Ellis attempted to repeat the work of fifty three major "landmark" papers, but were successful in only 11% (6 of 53) of the cases.[1] Similarly, recognition of the need to accurately reproduce experimental methods has caused some journals to expand their word limits for material and methods sections (*Nature Cell Biology*, 2009;11:667). Repeatability in pathology methods (including scoring) is a relevant and important consideration in experimental design as well as reporting of data.

One approach to validate scoring systems has been to assess its repeatability through evaluation of intra- and inter-observer correlation.[14,24,43,72] This evaluation is often reported by a kappa value (value of zero to one) that is calculated from observer agreements (Table 9). Validation using this method only assesses the repeatability of the method, but should not be confused with validation of tissue pathobiology as described below.

Validation of tissue pathobiology—Another approach to validate a scoring system is to analyze the relationship between the scores and relevant parameters of disease severity, i.e. pathobiology.[2,10,34,55] This relationship is defined through correlation (e.g. Spearman correlation for nonparametric data), which produces a value from −1.0 to 1.0. For example, comparison of tissue scores to relevant pathobiology data (e.g. clinical score, body weight, complete blood counts, etc.) would ideally demonstrate a strong positive correlation (Figure 3). Its interpretation is similar to that of the kappa - the closer to zero the lower the correlation. If it is a negative value then the scoring system has a negative correlation to pathobiology which would seem unsuitable (if not even "backwards") for many situations. If the scoring system does not have a strong, positive relationship to disease pathobiology, there may be reason to question its value in the respective model.

Each validation method is mutually exclusive in its scope. For example, when evaluating interobserver correlation, a high kappa value gives confidence in the scoring method's repeatability. That said, it does not give any credibility to the scoring method's representation of tissue pathobiology, and the contrary is true as well.

Conclusions

Scoring tissue lesions can be a useful tool for evaluating research tissues and corroborating morphologic findings. Following key principles can guide the pathologist to develop useful and valid scoring system that is both repeatable and meaningful for the project.

Acknowledgments

We would like to thank the Department of Pathology (University of Iowa) for generous support. We acknowledge generous financial support from the NIH (HL091842, HL051670, DK054759, DK091211) and US Veteran's Administration (Center for the Prevention and Treatment of Visual Loss).

References

1. Begley CG, Ellis LM. Raise standards for preclinical cancer research. Nature. 2012; 483:531–533. [PubMed: 22460880]
2. Bleich A, Mahler M, Most C, Leiter EH, Liebler–Tenorio E, Elson CO, Hedrich HJ, Schlegelberger B, Sundberg JP. Refined histopathologic scoring system improves power to detect colitis QTL in mice. Mamm Genome. 2004; 15:865–871. [PubMed: 15672590]
3. Brayton CF, Justice M, Montgomery CA. Evaluating Mutant Mice: Anatomic Pathology. Vet Pathol. 2001; 38:1–19. [PubMed: 11199155]
4. Brayton CF, Treuting PM, Ward JM. Pathobiology of Aging Mice and GEM: Background Strains and Experimental Design. Vet Pathol. 2012; 49:85–105. [PubMed: 22215684]
5. Brisson AR, Matsui D, Rieder MJ, Fraser DD. Translational Research in Pediatrics: Tissue Sampling and Biobanking. Pediatr. 2012; 129:1–10.
6. Brown AT, Ou X, James LP, Jambhekar K, Pandey T, McCullough S, Chaudhuri S, Borrelli MJ. Correlation of MRI findings to histology of acetaminophen toxicity in the mouse. Magn Reson Imaging. 2012; 30:283–9. [PubMed: 22055850]
7. Bucci, TJ. Basic techniques. In: Haschek, WM.; Rousseaux, CG.; Wallig, MA., editors. Handbook of Toxicologic Pathology. 2. Vol. 1. San Diego, CA: Academic Press; 2002. p. 171-85.
8. Burkhardt JE, Pandher K, Solter PF, Troth SP, Boyce RW, Zabka TS, Ennulat D. Recommendations for the evaluation of pathology data in nonclinical safety biomarker qualification studies. Toxicol Pathol. 2011; 39:1129–37. [PubMed: 22116771]
9. Cake MA, Smith MM, Young AA, Smith SM, Ghosh P, Read RA. Synovial pathology in an ovine model of osteoarthritis: effect of intraarticular hyaluronan (Hyalgan). Clin Exp Rheumatol. 2008; 26:561–7. [PubMed: 18799085]
10. Cheng Z, Dhall D, Zhao L, Wang HL, Doherty TM, Bresee C, Frykman PK. Murine model of Hirschsprung-associated enterocolitis. I: Phenotypic characterization with development of a histopathologic grading system. J Pediatr Surg. 2010; 45:475–482. [PubMed: 20223308]
11. Crawford JM, Tykocinski ML. Pathology as the enabler of human research. Lab Invest. 2005; 85:1058–64. [PubMed: 16056243]
12. Crissman JW, Goodman DG, Hildebrandt PK, Maronpot RR, Prater DA, Riley JH, Seaman WJ, Thakeet DC. Best practices guideline: toxicologic histopathology. Toxicol Pathol. 2004; 32:126–131. [PubMed: 14713558]
13. Cross SS. Grading and scoring in histopathology. Histopathol. 1998; 33:99–106.
14. Cross SS. Kappa statistics as indicators of quality assurance in histopathology and cytopathology. J Clin Pathol. 1996; 49:597–599. [PubMed: 8813964]
15. Day SJ, Altman DG. Statistics notes: blinding in clinical trials and other studies. Br Med J. 2000; 321:504. [PubMed: 10948038]
16. De Cock HE, Forman MA, Farver TB, Marks SL. Prevalence and histopathologic characteristics of pancreatitis in cats. Vet Pathol. 2007; 44:39–49. [PubMed: 17197622]
17. Eaton KA, Danon SJ, Krakowka S, Weisbrode SE. A Reproducible Scoring System for Quantification of Histologic Lesions of Inflammatory Disease in Mouse Gastric Epithelium. Comp Med. 2007; 57:57–65. [PubMed: 17348292]
18. Ettreiki C, Gadonna-Widehem P, Mangin I, Coëffier M, Delayre-Orthez C, Anton PM. Juvenile ferric iron prevents microbiota dysbiosis and colitis in adult rodents. World J Gastroenterol. 2012; 18:2619–29. [PubMed: 22690070]
19. Ferguson AR, Hook MA, Garcia G, Bresnahan JC, Beattie MS, Grau JW. A simple post hoc transformation that improves the metric properties of the BBB scale for rats with moderate to severe spinal cord injury. J Neurotrauma. 2004; 21:1601–13. [PubMed: 15684652]

20. Festing MFW. Design and Statistical Methods in Studies Using Animal Models of Development. ILAR J. 2006; 47:5–14. [PubMed: 16391426]
21. Festing MFW, Altman DG. Guidelines for the Design and Statistical Analysis of Experiments Using Laboratory Animals. ILAR J. 2002; 43:244–258. [PubMed: 12391400]
22. Gad, SC. Statistics and experimental design for toxicologists and pharmacologists. Boca Raton, FL, USA: CRC Press; 2006.
23. Gauger PC, Vincent AL, Loving CL, Henningson JN, Lager KM, Janke BH, Kehrli ME Jr, Roth JA. Kinetics of lung lesion development and pro-inflammatory cytokine response in pigs with vaccine-associated enhanced respiratory disease induced by challenge with pandemic) A/H1N1 influenza virus. Vet Pathol. 2012; 49:900–12. [PubMed: 22461226]
24. Germolec DR, Nyska A, Kashon M, Kuper CF, Portier C, Kommineni C, Johnson KA, Luster MI. Extended Histopathology in Immunotoxicity Testing: Interlaboratory Validation Studies. Toxicol Sci. 2004; 78:107–115. [PubMed: 14691208]
25. Gerwin N, Bendele AM, Glasson S, Carlson CS. The OARSI histopathology initiative - recommendations for histological assessments of osteoarthritis in the rat. Osteoarthritis Cartilage. 2010; 18 (Suppl 3):S24–34. [PubMed: 20864021]
26. Gholamiandehkordi AR, Timbermont L, Lanckriet A, Van Den Broeck W, Pedersen K, Dewulf J, Pasmans F, Haesebrouck F, Ducatelle R, Van Immerseel F. Quantification of gut lesions in a subclinical necrotic enteritis model. Avian Pathol. 2007; 36:375–82. [PubMed: 17899461]
27. Gibson-Corley KN, Hochstedler C, Sturm M, Rodgers J, Olivier AK, Meyerholz DK. Successful integration of the histology core laboratory in translational research. J Histotechnol. 2012; 35:17–21. [PubMed: 22904581]
28. Goddard CJ, Smith A, Hoyland JA, Baird P, McMahon RF, Freemont AJ, Shomaf M, Haboubi NY, Warnes TW. Localisation and semiquantitative assessment of hepatic procollagen mRNA in primary biliary cirrhosis. Gut. 1998; 43:433–40. [PubMed: 9863492]
29. Grafe MR, Woodworth KN, Noppens K, Regino Perez-Polo J. Long-term histological outcome after post-hypoxic treatment with 100% or 40% oxygen in a model of perinatal hypoxic-ischemic brain injury. Int J Dev Neurosci. 2008; 26:119–124. [PubMed: 17964109]
30. Holland T. The comparative power of the discriminant methods used in toxicological pathology. Toxicol Pathol. 2005; 33:490–494. [PubMed: 16036867]
31. Holland T, Holland C. Analysis of Unbiased Histopathology Data from Rodent Toxicity Studies (or, Are These Groups Different Enough to Ascribe It to Treatment?). Toxicol Pathol. 2011; 39:569–575. [PubMed: 21558466]
32. Hörnblad A, Cheddad A, Ahlgren U. An improved protocol for optical projection tomography imaging reveals lobular heterogeneities in pancreatic islet and β-cell mass distribution. Islets. 2011; 3:204–8. [PubMed: 21633198]
33. Hübner RH, Gitter W, El Mokhtari NE, Mathiak M, Both M, Bolte H, Freitag-Wolf S, Bewig B. Standardized quantification of pulmonary fibrosis in histological samples. Biotechniques. 2008; 44:507–11. 514–7. [PubMed: 18476815]
34. Isobe K, Adachi K, Hayashi S, Ito T, Miyoshi A, Kato A, Suzuki M. Spontaneous Glomerular and Tubulointerstitial Lesions in Common Marmosets (*Callithrix jacchus*). Vet Pathol. 2012; 49:839–845. [PubMed: 22156228]
35. Jacobs PC, Prokop M, Oen AL, van der Graaf Y, Grobbee DE, Mali WP. Semiquantitative assessment of cardiovascular disease markers in multislice computed tomography of the chest: interobserver and intraobserver agreements. J Comput Assist Tomogr. 2010; 34:279–84. [PubMed: 20351521]
36. Jassal MS, Nedeltchev GG, Osborne J, Bishai WR. A modified scoring system to describe gross pathology in the rabbit model of tuberculosis. BMC Microbiol. 2011; 11:49. [PubMed: 21375756]
37. Kayser K, Schultz H, Goldmann T, Görtler J, Kayser G, Vollmer E. Theory of sampling and its application in tissue based diagnosis. Diagn Pathol. 2009; 4:6. [PubMed: 19220904]
38. Kitching AR, Katerelos M, Mudge SJ, Tipping PG, Power DA, Holdsworth SR. Interleukin-10 inhibits experimental mesangial proliferative glomerulonephritis. Clin Exp Immunol. 2002; 128:36–43. [PubMed: 11982588]

39. Kleiner DE, Brunt EM, Natta MV, Behling C, Contos MJ, Cummings OW, Ferrell LD, Liu Y, Torbenson MS, Unalp-Arida A, Yeh M, McCullough AJ, Sanyal AJ. for the Nonalcoholic Steatohepatitis Clinical Research Network. Design and Validation of a Histological Scoring System for Nonalcoholic Fatty Liver Disease. Hepatol. 2005; 41:1313–1321.
40. Klesney-Tait J, Keck K, Li X, Gilfillan S, Otero K, Baruah S, Meyerholz DK, Varga SM, Knudson CJ, Moninger TO, Moreland J, Zabner J, Colonna M. Transepithelial migration of neutrophils into the lung requires TREM-1. J Clin Invest. 2013; 123:138–149. [PubMed: 23241959]
41. Kohlmann, T.; Moock, J. How to analyze your data. In: Stengal, D.; Bhandari, M.; Hanson, B., editors. Handbooks: Statistics and Data Management. Vol. Chapter 5. AO Publishing; Switzerland: 2009. p. 93-110.
42. Lafemina MJ, Sheldon RA, Ferriero DM. Acute hypoxia-ischemia results in hydrogen peroxide accumulation in neonatal but not adult mouse brain. Pediatr Res. 2006; 59:680–3. [PubMed: 16627881]
43. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977; 33:159–174. [PubMed: 843571]
44. Langlois RA, Meyerholz DK, Coleman RA, Cook RT, Waldschmidt TJ, Legge KL. Oseltamivir treatment prevents the increased influenza virus disease severity and lethality occurring in chronic ethanol consuming mice. Alcohol Clin Exp Res. 2010; 34:1425–31. [PubMed: 20497135]
45. Mann PC, Vahle J, Keenan CM, Baker JF, Bradley AE, Goodman DG, Harada T, Herbert R, Kaufmann W, Kellner R, Nolte T, Rittinghausen S, Tanaka T. International harmonization of toxicologic pathology nomenclature: an overview and review of basic principles. Toxicol Pathol. 2012; 40:7S–13S. [PubMed: 22637736]
46. Miao EA, Leaf IA, Treuting PM, Mao DP, Dors M, Sarkar A, Warren SE, Wewers MD, Aderem A. Caspase-1-induced pyroptosis is an innate immune effector mechanism against intracellular bacteria. Nat Immunol. 2010; 11:1136–42. [PubMed: 21057511]
47. Murthy S, Adamcakova-Dodd A, Perry SS, Tephly LA, Keller RM, Metwali N, Meyerholz DK, Wang Y, Glogauer M, Thorne PS, Carter AB. Modulation of reactive oxygen species by Rac1 or catalase prevents asbestos-induced pulmonary fibrosis. Am J Physiol Lung Cell Mol Physiol. 2009; 297:L846–55. [PubMed: 19684199]
48. Olivier AK, Naumann P, Goeken A, Hochstedler C, Sturm M, Rodgers JR, Gibson-Corley KN, Meyerholz DK. Genetically modified species in research: opportunities and challenges for the histology core laboratory. J Histotechnol. 2012; 35:63–67. [PubMed: 22904582]
49. Pannucci CJ, Wilkins EG. Identifying and Avoiding Bias in Research. Plast Reconstr Surg. 2010; 126:619–625. [PubMed: 20679844]
50. Pearson RG, Kurien T, Shu KSS, Scammell BE. Histopathology grading systems for characterisation of human knee osteoarthritis e reproducibility, variability, reliability, correlation, and validity. Osteoarth Cartil. 2011; 19:324–331.
51. Pinson DM, Schoeb TR, Lindsey JR, Davis JK. Evaluation by Scoring and Computerized Morphometry of Lesions of Early Mycoplasma pulmonis Infection and Ammonia Exposure in F344/N Rats. Vet Pathol. 1986; 23:550–555. [PubMed: 3776012]
52. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov. 2011; 10:712. [PubMed: 21892149]
53. Pritzker KP, Gay S, Jimenez SA, Ostergaard K, Pelletier JP, Revell PA, Salter D, van den Berg WB. Osteoarthritis cartilage histopathology: grading and staging. Osteoarthritis Cartilage. 2006; 14:13–29. [PubMed: 16242352]
54. Rollins KE, Meyerholz DK, Johnson GD, Capparella AP, Loew SS. A Forensic Investigation Into the Etiology of Bat Mortality at a Wind Farm: Barotrauma or Traumatic Injury? Vet Pathol. 2012; 49:362–371. [PubMed: 22291071]
55. Scheinin T, Butler DM, Salway F, Scallon B, Feldmann M. Validation of the interleukin-10 knockout mouse model of colitis: antitumour necrosis factor- antibodies suppress the progression of colitis. Clin Exp Immunol. 2003; 133:38–43. [PubMed: 12823276]
56. Schulz KF, Chalmers I, Hayes R, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA. 1995; 273:408–12. [PubMed: 7823387]

57. Sellers RS. The Gene or Not the Gene--That Is the Question: Understanding the Genetically Engineered Mouse Phenotype. Vet Pathol. 2012; 49:5–15. [PubMed: 21971987]
58. Shackelford C, Long G, Wolf J, Okerberg C, Herbert R. Qualitative and Quantitative Analysis of Nonneoplastic Lesions in Toxicology Studies. Toxicol Pathol. 2002; 30:93–96. [PubMed: 11890482]
59. Shott S. Statistics simplified: Designing studies that answer questions. J Am Vet Med Assoc. 2011; 238:55–58. [PubMed: 21194321]
60. Shott S. Statistics simplified: Detecting statistical errors in veterinary research. J Am Vet Med Assoc. 2011; 237:305–308. [PubMed: 21281212]
61. Shott S. Statistics simplified: Wrapping it all up. J Am Vet Med Assoc. 2011; 239:362–371.
62. Sica GT. Bias in research studies. Radiology. 2006; 238:780–9. [PubMed: 16505391]
63. Snider TA, Confer AW, Payton ME. Pulmonary Histopathology of Cytauxzoon felis Infections in the Cat. Vet Pathol. 2010; 47:698–702. [PubMed: 20442419]
64. Steel, RGD.; Torrie, JH.; Dickey, DA. Nonparametric statistics. In: Steel, RGD.; Torrie, JH.; Dickey, DA., editors. Principles and procedures of statistics: a biomedical approach. 3. Boston, MA, USA: WCB McGraw-Hill; 1997. p. 563-588.
65. Stevens SS. On the Theory of Scales of Measurement. Science. 1946; 103:677–680.
66. Strasak AM, Zaman Q, Pfeiffer KP, Göbel G, Ulmer H. Statistical errors in medical research – a review of common pitfalls. Swiss Med Wkly. 2007; 137:44–49. [PubMed: 17299669]
67. Thoolen B, Maronpot RR, Harada T, Nyska A, Rousseaux C, Nolte T, Malarkey DE, Kaufmann W, Küttler K, Deschl U, Nakae D, Gregson R, Vinlove MP, Brix AE, Singh B, Belpoggi F, Ward JM. Proliferative and nonproliferative lesions of the rat and mouse hepatobiliary system. Toxicol Pathol. 2010; 38:5S–81S. [PubMed: 21191096]
68. Timbermont L, Lanckriet A, Dewulf J, Nollet N, Schwarzer K, Haesebrouck F, Ducatelle R, Van Immerseel F. Control of Clostridium perfringens-induced necrotic enteritis in broilers by target-released butyric acid, fatty acids and essential oils. Avian Pathol. 2010; 39:117–21. [PubMed: 20390546]
69. Torrence AE, Brabb T, Viney JL, Bielefeldt-Ohmann H, Treuting P, Seamons A, Drivdahl R, Zeng W, Maggio-Price L. Serum biomarkers in a mouse model of bacterial-induced inflammatory bowel disease. Inflamm Bowel Dis. 2008; 14:480–90. [PubMed: 18095317]
70. Vascellari M, Giantin M, Capello K, Carminato A, Morello EM, Vercelli A, Granato A, Buracco P, Dacasto M, Mutinelli F. Expression of Ki67, BCL-2, and COX-2 in Canine Cutaneous Mast Cell Tumors: Association With Grading and Prognosis. Vet Pathol. 201210.1177/0300985812447829
71. Venturi C, Sempoux C, Bueno J, Ferreres Pinas JC, Bourdeaux C, Abarca- Quinones J, Rahier J, Reding R. Novel histologic scoring system for long-term allograft fibrosis after liver transplantation in children. Am J Transplant. 2012; 12:2986–96. [PubMed: 22882699]
72. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005; 37:360–3. [PubMed: 15883903]
73. Wachtel MS, Shome G, Sutherland M, McGlone JJ. Derivation and validation of murine histologic alterations resembling asthma, with two proposed histologic grade parameters. BMC Immunol. 2009; 10:58. [PubMed: 19878549]
74. Ward JM, Thoolen B. Grading of lesions. Toxicol Pathol. 2011; 39:745–746. [PubMed: 21666104]
75. Zeiss CJ, Ward JM, Allore HG. Designing Phenotyping Studies for Genetically Engineered Mice. Vet Pathol. 2012; 49:24–31. [PubMed: 21930803]

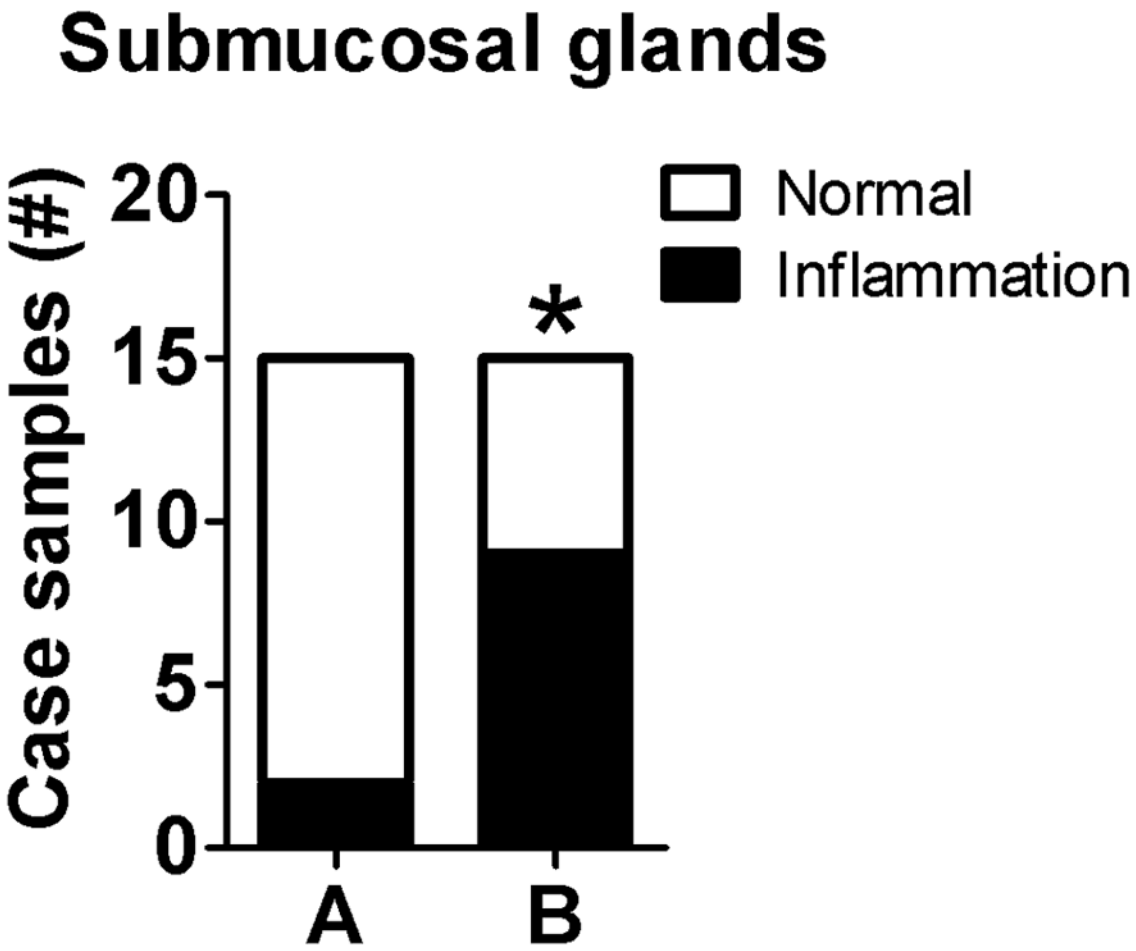

Fig. 1.
Example of the incidence method. Example graph for reporting incidence data from Table 4.
**P*=0.02, Fisher's exact test.

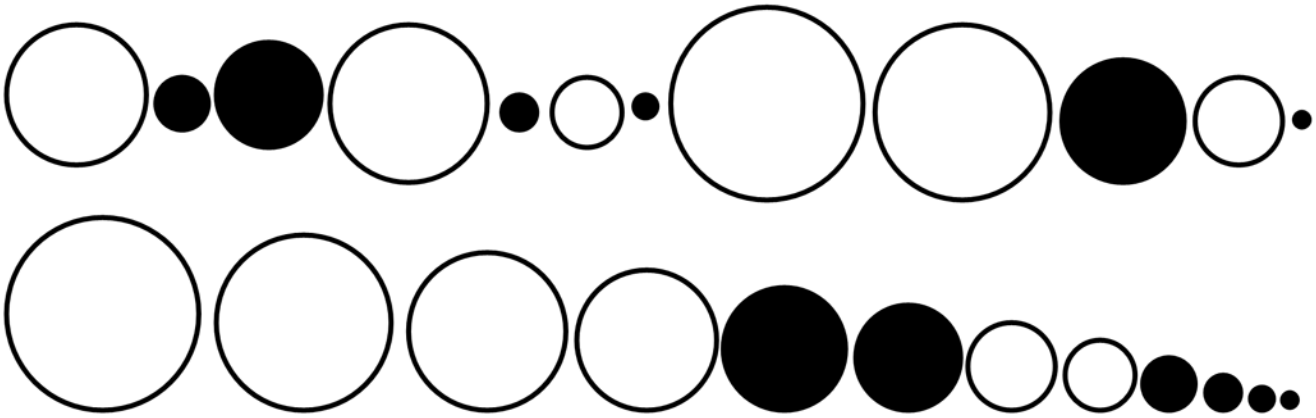

Fig. 2.
Example of the rank method. Samples (circles) from Group A (white circles) and Group B (black circles) are combined (top row) for examination. The samples are then ranked in order of lesion severity (represented by circle diameter, bottom row). The rank numbers for Group A (1,2,3,4,7,8) and Group B (5,6,9,10,11,12) are then analyzed. *P*=0.03, Mann-Whitney test.

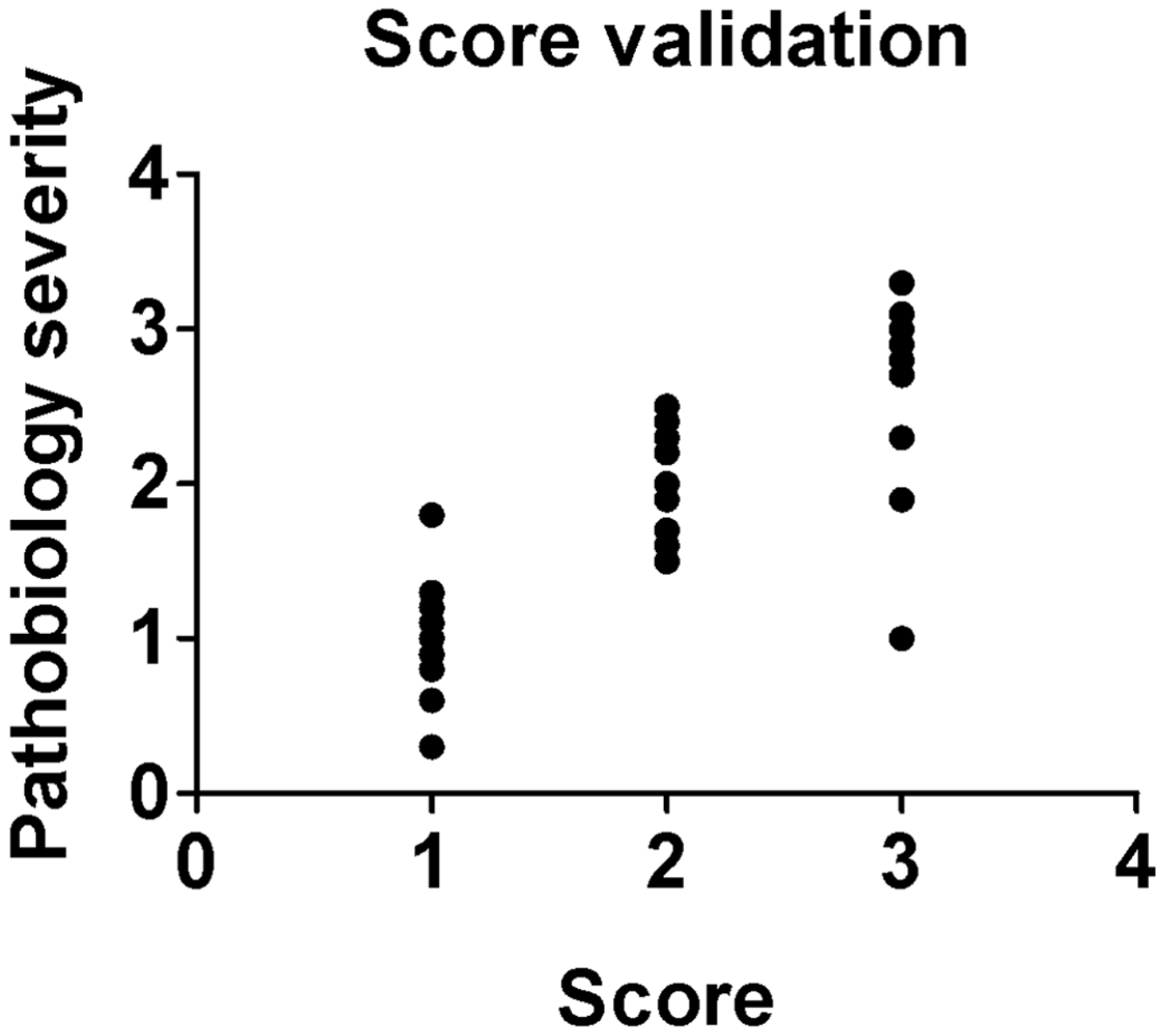

Fig. 3.
Validation of pathobiology. Tissue scores (x axis) are graphed out in comparison to relevant pathobiology parameters (y axis) to see if there is a relationship (i.e. correlation (r = 0.80, *P* = 0.001, Spearman correlation). Since the r value is positive and close to "1", this would indicate a strong correlation of the scoring method with tissue pathobiology.

Table 1

Common methods of masking tissues for histopathologic examination.

Method	Description	Comments
Comprehensive	Individual samples are labeled without reference to treatment group (e.g. 1, 2, 3, 4, 5, etc.) and minimal background information (perspective) is given.	Pro: Bias is comprehensively constrained Con: Pathologist labor may be increased in examination, while sensitivity to subtle study-specific lesions may decrease[12]
Grouped	Samples are coded by groups (e.g. A1, A2, … A10; B1, B2, ….B10); relevant background material including study design and objectives are disclosed to pathologist.	Pro: Pathologist is masked to group treatments, but is aware of tissue grouping and background information. Con: Overt group differences can functionally unmask the pathologist and if performing ordinal scoring may warrant comprehensive masking.
Post examination masking	Full disclosure of experimental design and objectives with unmasked initial evaluation; masking and randomization of samples are done prior to scoring	Pro: Offers full disclosure to the pathologist for examination and scoring development. Con: Pathologists may recall group assignments of samples with small n/group which makes masking ineffective.

Table 2

Examples of tissues and techniques in which histopathologic scoring has been reported.

Pancreas[16]	Cystic degeneration Fat necrosis Fibrosis Lymphoid inflammation Neutrophilic inflammation
Liver[39,71]	Cell injury Nuclear and cytoplasmic features Inflammation Fibrosis Steatosis
Respiratory[23,33,44,47,51,54,63]	Bronchitis/bronchiolitis Edema Epithelial thickening Epithelial degeneration/necrosis Fibrosis Interstitial pneumonia Lymphoid inflammation Metaplasia Neutrophilic inflammation
Spleen[46]	Bacteria Necrosis Neutrophils influx Thrombosis
Orthopedic[9,25,53]	Cartilage calcification Cartilage Fibrosis Osteoarthritis Osteophytes degeneration Subchondral bone damage Synovial hyperplasia Synovial inflammation Vascularity
Digestive tract[10,17,26]	Enterocolitis Epithelial erosion Gut lumen contents Gastric neutrophils Gastritis Gastric metaplasia Hemorrhage Vascular congestion Villous fusion
Brain[29,42]	Hypoxic injury Infarction
Immunohistochemistry[23,38]	Staining distribution
In situ hybridization[28]	Staining distribution

Table 3

Types and examples of data measurements in research (Adapted from Steven 1946)[65]

Types	Definition	Example(s)
Nominal	Samples assigned to a category without reference to severity gradations.	"Binary" - presence or absence of a lesion (+/−) "Categorical" - lesions assigned to a non-ordered category (carcinoma, sarcoma)
Ordinal	Samples assigned to a category showing an ordered progression in severity	0 - normal 1 - mild 2 - moderate 3 - severe
Interval	Samples quantified on a scale between two extremes and with an arbitrary zero value. Samples can be compared based on differences in value, but not using multiplication or division.	Celsius scale of 0–100° based on freezing and boiling points of water.
Ratio	Samples quantified on a scale with a true zero value. Samples can be compared through differences or multiplication/division.	Most morphometry data (e.g. length, area, etc.) produces quantitative values.

Table 4

Scoring of trachea submucosal glands for the presence of cellular inflammation[a].

Group	Normal	Inflammation	% inflammation
A	13	2	13.3%
B	6	9	60.0%

[a]Sections of trachea with submucosal glands from each animal in group A (n=15) and B (n=15) were examined and designated as within normal limits or with cellular inflammation.

Table 5

Example of ordinal scores based on distribution of tracheal lesions.

Score	Trachea (% wall affected)
0	No change
1	<25%
2	26–50%
3	51–75%
4	76–100%

Table 6

Trachea inflammation and hyperplasia scores from treatment groups A and B.[a]

	Group A		Group B	
Animal	**Inflammation**	**Hyperplasia**	**Inflammation**	**Hyperplasia**
1	1	1	3	2
2	0	0	2	1
3	1	0	3	2
4	1	1	2	1
5	0	0	1	1
6	2	1	1	1
7	1	0	2	2
8	1	1	2	1
9	1	0	2	1
10	1	0	1	0
Median	**1**	**0**	**2**[b]	**1**[c]

[a] Scoring was performed for each parameter based on Table 5.

[b] Group A vs. B, P=0.006, Mann Whitney test

[c] Group A vs. B, P=0.011, Mann Whitney test

Table 7

Example of a scoring system that combines lesion parameters to define each category.

Score	Kidney scoring for acute tubular injury
1	Isolated tubular ectasia, rare sloughed cells in tubular lumens, inflammation absent to minimal
2	Multifocal tubular ectasia, patchy sloughed cells in tubular lumens, rare to multifocal interstitial inflammation
3	Coalescing to diffuse tubular ectasia, diffuse sloughed and necrotic cells obstructing tubular lumens, multifocal to diffuse inflammation

Table 8

Example of a scoring system that takes parameters from Table 7 and separates each into its own scoring system.

Score	Ectasia	Necrosis	Inflammation
1	Rare (<5%)	Rare (<5%)	Rare (<5%)
2	Multifocal (6–40%)	Multifocal (6–40%)	Multifocal (6–40%)
3	Coalescing (41–80%)	Coalescing (41–80%)	Coalescing (41–80%)
4	Diffuse (>80%)	Diffuse (>80%)	Diffuse (>80%)

Table 9

Interobserver agreement (observer A and B) for classification of hepatocellular carcinoma (HCC) and hepatocellular adenoma (HCA) from liver tumor samples (n=100).[a]

	HCC - B	HCA - B
HCC - A	**39**	10
HCA - A	6	**45**

[a] kappa value was calculated as (HCC + HCA agreements)/total assessments. kappa = (39+45)/100= 0.84. The kappa score indicates there is a strong agreement between observer A and B in classifying these liver tumors.