ILAR Journal, 2019, Vol. 59, No. 1, 13–17

doi: 10.1093/ilar/ily025
Review Article

Fundamental Concepts for Semiquantitative Tissue Scoring in Translational Research

David K. Meyerholz[1] and Amanda P. Beck[2]

[1]Department of Pathology, University of Iowa Carver College of Medicine, Iowa City, Iowa and [2]Department of Pathology, Albert Einstein College of Medicine, Bronx, New York

Address correspondence and reprint requests to David Meyerholz, DVM, PhD, DACVM, DACVP, Department of Pathology, University of Iowa Carver College of Medicine, 1165ML, Iowa City, Iowa, 52242 or email david-meyerholz@uiowa.edu.

Abstract

Failure to reproduce results from some scientific studies has raised awareness of the critical need for reproducibility in translational studies. Macroscopic and microscopic examination is a common approach to determine changes in tissues, but text descriptions and visual images have limitations for group comparisons. Semiquantitative scoring is a way of transforming qualitative tissue data into numerical data that allow more robust group comparisons. Semiquantitative scoring has broad uses in preclinical and clinical studies for evaluation of tissue lesions. Reproducibility can be improved by constraining bias through appropriate experimental design, randomization of tissues, effective use of multidisciplinary collaborations, and valid masking procedures. Scoring can be applied to tissue lesions (eg, size, distribution, characteristics) and also to tissues through evaluation of staining distribution and intensity. Semiquantitative scores should be validated to demonstrate relevance to biological data and to demonstrate observer reproducibility. Statistical analysis should make use of appropriate tests to give robust confidence in the results and interpretations. Following key principles of semiquantitative scoring will not only enhance descriptive tissue evaluation but also improve quality, reproducibility, and rigor of tissue studies.

Key words: bias; clinical; grading; pathology; preclinical; reproducibility; semiquantitative scoring; tissue scores

Introduction and Uses

Tissue evaluation is a common research tool used in basic science and[1–4] toxicological[5–10] and clinical studies.[11–14] Scoring of tissues changes or lesions can aid in assessing model phenotypes, disease pathogenesis, toxicities, and efficacy of therapies.[2,5,12–15] Morphological examination of tissues produces text descriptions and visual images that can be valuable to define initial group-specific differences; however, these observations are qualitative in nature and have limitations for rigorous group comparisons. In general, quantitative and semiquantitative approaches can be applied to tissues to produce scores that enhance the rigor of data. "Quantitative" scores are derived from measuring tissue parameters often using manual techniques or by using specialized software to analyze digital images[3,16,17] and yield a discrete numeric value on a continuous scale (eg, 0.3, 1.25, 4.5, etc.). In contrast, "semiquantitative" scores are assigned by an observer based on predefined morphologic criteria,[3] and these whole number scores are, by definition, less precise than quantitative scores because they approximate relative changes. Semiquantitative scoring can be applied to macroscopic and microscopic tissue changes, allowing generation of robust data that are amenable to statistical analysis and evaluation of experimental groups. The goals of this paper are to introduce investigators to key ideas in reproducible semiquantitative scoring of tissues and guide them in finding additional resources for more detailed discussions and examples. For the remainder of this paper, "scores" and "scoring" will refer, unless otherwise specified, to semiquantitative methods.

Integration of semiquantitative scoring in translational research can be useful in several situations.[3,4,18,19] First, semiquantitative

scoring data are relatively inexpensive, because no software or computational tools are necessarily needed. Second, it can be a quick screening method to produce pilot data for grant applications or guide future research studies. Third, semiquantitative data can enhance the rigor of descriptive text. While annotated images and descriptive text may show apparent differences between groups, semiquantitative scores can provide a comprehensive overview of tissue changes for group comparisons. Lastly, semiquantitative data can be used to guide, corroborate, and validate observations or data obtained from other assays.

Semiquantitative scoring can be used to acquire data in several scientific areas, and fundamentally the core concepts are similar.[3–8] In the preclinical area, which utilizes models (eg, animal, tissue/cell cultures, etc.) of human diseases/conditions, semiquantitative scoring is regularly used to compare experimental groups.[1,2,11,20] In the clinical area, semiquantitative scoring of human tissues (eg, cancers, tissue/cell cultures, etc.) is often used to help define disease diagnosis, pathogenesis, biomarkers, and clinical prognosis.[12–14]

Semiquantitative scoring is also a key component of nonclinical toxicology studies,[5,10] which are performed to support regulatory agency submissions and thus have an inherently different purpose than preclinical investigative studies. Here, the goal is to evaluate the safety of the material being tested (ie, hazard identification and risk assessment) rather than to assess potential treatment efficacy. To support future clinical trials, all toxicity studies must be performed according to guidance documents from various regulatory agencies, such as the Food and Drug Administration. Additionally, the usage of consistent diagnostic terminology for each organ system in rodents and large animals is strongly recommended.[9,21] Collaboration with experienced toxicologists and toxicological pathologists is highly encouraged before investigators plan these types of studies to ensure the current regulatory guidelines are followed. Unless specified, the remainder of the paper will focus on foundational concepts for semiquantitative scoring emphasizing nontoxicologic translational studies (Table 1).

Bias Control

Statistician George Box once stated, "*All models are wrong, but some are useful.*"[22] To apply this quote in the context of translational research, modeling in itself (eg, animal models) is never fully identical to the condition being modeled (eg, human disease). Due to several factors (genetic diversity, comorbidities, etc.), even small cohorts of humans do not fully "model" the human condition. This is, in part, why large and multiple clinical trials are often required to test for efficacy and adverse effects of new therapeutics in humans. In research, studies that model the human condition should be constructed to be as useful and reproducible as possible; one way to do this is to guard against factors that are known to cause bias. In science, bias is a term applied to areas of subjectivity (from overt to subconscious) that can skew data and contribute to lack of scientific reproducibility, an unfortunate reality that has been increasingly recognized.[23–25] There are several ways to constrain bias when scoring tissues, and by using these precepts investigators can acquire more objective data.

Table 1 Fundamental Concepts for Semiquantitative Scoring of Tissues

Principles	Resources/Examples
Bias control	Experimental design[4,27,32]
	Randomization[31,32]
	Expertise[23,32–34,50]
	Masking[3–5,51]
Methods	Lesions (size, shape, number, etc.)[13,20,41,44]
	Stains (incidence; intensity)[3,12,13,20]
	Scoring methods[4,43,45]
Evaluation	Biological validation[3,4,14,48]
	Validation of repeatability[52–54]
	Group comparisons[32,43,55–58]

Experimental Design

A critical step for reproducible science is to establish a strong foundation in sound experimental design.[4,23,26–29] Constraining bias early, at the experimental design stage, avoids downstream "junk in, junk out" problems and issues of "regret" that can lead to adverse and unexpected influences in the quality and analyses of tissues.[4,30] Considerations to address during the experimental planning stage include selection of the appropriate model (eg, species or strain), consideration of the appropriate controls (eg, matching with respect to age, sex, or litter), and calculation of the sufficient sample size needed for statistical significance. It can be helpful to revisit proper techniques for tissue collection as well as the different options available for fixation and storage because tissue handling variables can influence staining quality.[3,4,27,30] Staining techniques can also vary in consistency as a function of stain choice and by staining protocol. For example, the planning phase for a hypothetical experiment involving viral-induced inflammation in the lungs of a mouse should address whether there is sufficient tissue for multiple tests (eg, bronchoalveolar lavage, paraffin, and OCT embedded tissues, PCR, microarray, protein quantification, and viral culture). Novice investigators might make several invalid assumptions (eg, homogenous virus distribution in lungs, bronchoalveolar lavage collection does not affect other analyses, murine lung size will allow for ample tissue sampling, etc.) that can lead to incomplete and/or skewed data.[4] Early consultation with all key collaborators (especially pathologists) at the time of experimental design will ensure all needs are accounted for (eg, appropriate amount and type of tissue allocations) to prevent oversights.

Randomization

Randomization ("heterogenization") is an important tool to prevent the introduction of treatment bias that arises from overly homogenized groups; this situation has been variably coined as litter effect, cage effect, or batch effect.[30–32] The introduction of such bias can sometimes happen in innocuous ways. For example, tissue harvest from a large cohort of animals will likely produce a wide range of times from onset of the experimental day until necropsy. If animal in one treatment group were necropsied early, before starting on the other group, tissue parameters such as liver glycogen stores (especially in fasted animals) could be affected and create artifactual group-specific bias. Randomization of all the groups (animals and their tissues) can mitigate bias introduced by the experimental procedures. Other examples of variables that could render a study nonrandomized include differential housing of subjects (single vs group) or subject/sample processing order. Any variable that is not randomized across treatment groups has the potential to confound the data.

Expertise

Bias may also be introduced into translational research in studies conducted without the support of expertise-specific collaborators to help plan, execute, and appropriately interpret the study.[33,34] Specifically, statistical and pathological analyses are common components in translational studies, but trained statisticians and board-certified pathologists are often omitted from these multidisciplinary teams, leading to data interpretations that are more prone to errors.[22,23,35] For tissue scoring, a designated "observer" must thoroughly examine samples and ascribe scores. Various biomedical personnel (including principal investigators, postdocs, and even students) have been assigned the role of observer to score tissues. This approach, which lacks the expertise of a board-certified pathologist trained in tissue interpretation, has been labeled as do-it-yourself pathology, a practice that has been associated with numerous publications with erroneous interpretations.[4,30,36–39] While observations made by biomedical personnel may be biologically accurate in some cases, it is important to note that tissue examination by nonpathologists (even those who are "scientific experts" for a particular disease) is not recommended. Nonpathologist observers are more prone to making Type I errors (ie, "false positives" often from inadequate consideration of other morphologically similar tissue changes) and Type II errors (ie, "false negatives" often from not recognizing unexpected tissue changes). Inclusion of experienced and board-certified pathologists, who are specially trained to examine and interpret tissue changes as part of the multidisciplinary team, can greatly enhance the quality of tissue evaluation and scoring.

Masking

Semiquantitative scoring depends on the judgment of an "observer," exposing the evaluation to some level of bias. Masking (also known as blinding) is a method to keep the observer from knowing the treatment groups when assigning tissue scores. Experts at every level (even pathologists!) are at risk of having their judgment subliminally influenced by information cues from the study. Masking significantly reduces this possibility. There are several methods to mask observers to the experimental groups, each with advantages and disadvantages that have been previously reviewed.[3,4,40] Briefly, comprehensive masking prevents the observer from knowing any details about the study design, treatments, or grouping of samples at initial examination. This approach may seem unbiased and even useful upon first glance, but in reality can easily lead to false negatives and skewed interpretations. An alternative approach to comprehensive masking is group masking. Here, the study design, treatments, and goals are all transparent to the observer; however, the samples are each assigned into de-identified groups, so that the observer does not know which group had specific treatments. A final example is that of postexamination masking. In this approach, full transparency and access are allowed to all study-related information and slides. This is an important step, especially in new or poorly characterized models, to avoid missing subtle or unexpected treatment-related changes. Once the decision is made to score the tissues, the slides are masked to the observer and scores assigned. Masking should be a standard component that is defined in the methodology of all studies that use semiquantitative scoring. For each of these approaches, the observer should evaluate the scores and tissues after scoring in a nonmasked fashion to give confidence in the scoring system and interpretation of the results.

Methods

One of the major benefits of semiquantitative scoring is the transformation of descriptive (qualitative) observations into numerical data so as to allow statistical group comparisons and enrich data quality. A widely accepted premise for tissue scoring is the exhibition of at least three characteristics: it should be definable, reproducible, and produce meaningful results.[5] In translational studies, scoring is typically performed on tissues to detect treatment group differences. There are 2 major types of tissue changes that are targeted when scoring tissues: lesions and stains (or other labeling techniques). Some studies have used a merged scoring (ie, an average or sum of scores) approach in which multiple parameters are combined to form one final "composite" score, but if this approach is used it should have biological relevance.[3,12,41]

Lesions

A tissue lesion can be defined as an observed morphologic change that differs from control or normal tissue architecture. Lesions can be scored in many ways, such as size, shape, distribution, presence/absence, etc., depending on the expected disease-specific findings or tissue observations. Considerations for selecting the appropriate scoring parameter include a thorough examination of all tissues that catalogs the lesions seen; identification of lesion parameters (size, shaped, etc.) that appear to have chronological or group specific differences; and biological relevance to the pathophysiology of the model.

Stains

Another common approach is to score histochemically or immunohistochemically stained tissues or cells.[3,42] Here, the observer can assess either the distribution (eg, percent of stained cells) or intensity (eg, weak to robust) of the labeled cells.[12] Similar to considerations described for "lesions," selection of a scoring parameter may be dependent on the staining presentation as well as the biology of the model. For example, a virus infection of the lung might warrant evaluation of the distribution of staining, whereas a TP53 marker might require staining intensity as a gauge of activation in benign vs malignant tumors.

Scoring Methods

Several methods of semiquantitative scoring have been discussed in recent reviews, and readers are encouraged to use these for more specific details.[3,4,6,41,43–45] While several types of semiquantitative scoring tests are available, ordinal scoring is by far the most common in translational research and will be further discussed here. Ordinal systems produce hierarchal or progressive numeral scores (also known as "grades" or "tiers") that are reflective of the extent and/or severity of change. A mock example of this is an ordinal scoring method composed of whole numbers from 0 to 4 representing distribution of tissue necrosis in which 0 is normal, 1 is <25% necrosis, 2 is 25% to 50% necrosis, 3 is 51% to 75% necrosis, and 4 is >75% necrosis.

Ordinal scoring systems should follow several key principles for enhanced reproducibility. First, the range of levels is recommended to be about 4 to 5; fewer than this decreases sensitivity to detect group differences and more than this reduces repeatability[3,5,6,43] Second, each progressive level should have well-defined descriptors (such as the percentage of tissue affected, as in the example above). Descriptors that are vague and

subjective, such as 0 is normal, 1 is mild, 2 is moderate, and 3 is severe, should be avoided or include additional information to clearly discern each level. Score descriptors in an ordinal system can be defined by multiple lesion parameters (eg, inflammation, proliferation, necrosis), but in these situations reproducibility can sometimes be limited. Therefore, separating each lesion parameter into its own ordinal scoring system is often preferred. Third, ordinal scores are inherently discontinuous data that are not normally distributed (bell-shaped) and require nonparametric statistical analyses. Data that are normally distributed should be analyzed with parametric analysis (eg, paired or unpaired *t* tests). Many statistics software packages include tests for normality for determining whether a given statistical test will be valid for the dataset. It is not appropriate to use parametric analysis to analyze data derived from ordinal scoring systems.[3,4,46]

Evaluation

Biological Validation

For semiquantitative scoring to have purpose and relevance, it should have validation with biologically relevant data. In this evaluation, semiquantitative scores are tested for a correlation with biologically relevant data in the model.[4,47,48] If a significantly positive or negative correlation exists, then this confirms that the scoring system is relevant to the model. Conversely, if no correlation exists, then one has to question the use and utility of the scoring system for the model.

Validation of Repeatability

Another form of validation is that of repeatability by the observer, both intra-observer (same person scoring the data) and inter-observer (different people scoring the data).[3,4,49] Validation of repeatability gives confidence in the scoring system descriptors as it relates to the model and also gives confidence in its repeatable use by other laboratories.

Group Comparisons

Once the semiquantitative tissue scores are collected, appropriate statistical tests can be applied; these have been reviewed.[4,5,8,45,46] As mentioned above, appropriate expertise such as a statistician collaborator would be advantageous to guide proper statistical analyses of the data. Awareness of the type of data produced by semiquantitative scoring is very important because it guides the type of statistical tests used to give the most compelling interpretations of the study.[46] As alluded to above, ordinal scoring is not parametric in nature, and thus selection of nonparametric tests should be considered.

Summary

Semiquantitative scoring is a simple and relatively inexpensive approach to enhance descriptive/qualitative tissue data. Understanding common applications of semiquantitative scoring and the key concepts for repeatability will enhance scientific studies in translational research.

References

1. Engelberg JA, Giberson RT, Young LJ, Hubbard NE, Cardiff RD. The use of mouse models of breast cancer and quantitative image analysis to evaluate hormone receptor antigenicity after microwave-assisted formalin fixation. *J Histochem Cytochem*. 2014;62(5):319–334.
2. Li K, Wohlford-Lenane CL, Channappanavar R, et al. Mouse-adapted MERS coronavirus causes lethal lung disease in human DPP4 knockin mice. *Proc Natl Acad Sci USA*. 2017;114(15):E3119–E3128.
3. Meyerholz DK, Beck AP. Principles and approaches for reproducible scoring of tissue stains in research. *Lab Invest*. 2018;98(7):844–855.
4. Meyerholz DK, Sieren JC, Beck AP, Flaherty HA. Approaches to evaluate lung inflammation in translational research. *Vet Pathol*. 2018;55(1):42–52.
5. Crissman JW, Goodman DG, Hildebrandt PK, et al. Best practices guideline: toxicologic histopathology. *Toxicol Pathol*. 2004;32(1):126–131.
6. Holland T. Reporting of toxicologic histopathology: contrasting approaches in diagnostic versus experimental practice. *Toxicol Pathol*. 2011;39(2):418–421.
7. Morton D, Kemp RK, Francke-Carroll S, et al. Best practices for reporting pathology interpretations within GLP toxicology studies. *Toxicol Pathol*. 2006;34(6):806–809.
8. Shackelford C, Long G, Wolf J, Okerberg C, Herbert R. Qualitative and quantitative analysis of nonneoplastic lesions in toxicology studies. *Toxicol Pathol*. 2002;30(1):93–96.
9. Mann PC, Vahle J, Keenan CM, et al. International harmonization of toxicologic pathology nomenclature: an overview and review of basic principles. *Toxicol Pathol*. 2012;40(4 Suppl):7S–13S.
10. Schafer KA, Eighmy J, Fikes JD, et al. Use of severity grades to characterize histopathologic changes. *Toxicol Pathol*. 2018;46(3):256–265.
11. Klopfleisch R. Multiparametric and semiquantitative scoring systems for the evaluation of mouse model histopathology—a systematic review. *BMC Vet Res*. 2013;9:123.
12. Ellis MJ, Coop A, Singh B, et al. Letrozole is more effective neoadjuvant endocrine therapy than tamoxifen for ErbB-1- and/or ErbB-2-positive, estrogen receptor-positive primary breast cancer: evidence from a phase III randomized trial. *J Clin Oncol*. 2001;19(18):3808–3816.
13. Meyerholz DK, Lambertz AM, McCray Jr PB. Dipeptidyl peptidase 4 distribution in the human respiratory tract: implications for the Middle East Respiratory Syndrome. *Am J Pathol*. 2016;186(1):78–86.
14. Kleiner DE, Brunt EM, Van Natta M, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology*. 2005;41(6):1313–1321.
15. Haskins DL, Howerth EW, Tuberville TD. Experimentally induced selenosis in yellow-bellied slider turtles (Trachemys scripta scripta). *Vet Pathol*. 2018;55(3):473–477.
16. Aeffner F, Blanchard TW, Keel MK, Williams BH. Whole-slide imaging: the future is here. *Vet Pathol*. 2018;55(4):488–489.
17. Aeffner F, Martin NT, Peljto M, et al. Quantitative assessment of pancreatic cancer precursor lesions in IHC-stained tissue with a tissue image analysis platform. *Lab Invest*. 2016;96(12):1327–1336.
18. Ibberson CB, Parlet CP, Kwiecinski J, Crosby HA, Meyerholz DK, Horswill AR. Hyaluronan modulation impacts Staphylococcus aureus biofilm infection. *Infect Immun*. 2016;84(6):1917–1929.
19. Kaser A, Mairinger T, Vogel W, Tilg H. Infliximab in severe steroid-refractory ulcerative colitis: a pilot study. *Wien Klin Wochenschr*. 2001;113(23–24):930–933.
20. Meyerholz DK, Stoltz DA, Gansemer ND, et al. Lack of cystic fibrosis transmembrane conductance regulator disrupts

fetal airway development in pigs. *Lab Invest*. 2018;98(6): 825–838.
21. Keenan CM, Baker J, Bradley A, et al. International Harmonization of Nomenclature and Diagnostic criteria (INHAND): progress to date and future plans. *Toxicol Pathol*. 2015;43(5):730–732.
22. Laman JD, Kooistra SM, Clausen BE. Reproducibility issues: avoiding pitfalls in animal inflammation models. *Methods Mol Biol*. 2017;1559:1–17.
23. Zeiss CJ, Ward JM, Allore HG. Designing phenotyping studies for genetically engineered mice. *Vet Pathol*. 2012;49(1): 24–31.
24. Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. *Nature*. 2012;483(7391):531–533.
25. Sica GT. Bias in research studies. *Radiology*. 2006;238(3): 780–789.
26. Johnson PD, Besselsen DG. Practical aspects of experimental design in animal research. *ILAR J*. 2002;43(4):202–206.
27. Scudamore CL, Soilleux EJ, Karp NA, et al. Recommendations for minimum information for publication of experimental pathology data: MINPEPA guidelines. *J Pathol*. 2016;238(2):359–367.
28. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol*. 2010;8(6): e1000412.
29. Bahr A, Wolf E. Domestic animal models for biomedical research. *Reprod Domest Anim*. 2012;47(Suppl 4):59–71.
30. Gibson-Corley KN, Hochstedler C, Sturm M, Rogers J, Olivier AK, Meyerholz DK. Successful integration of the histology core laboratory in translational research. *J Histotechnol*. 2012; 35(1):17–21.
31. Richter SH. Systematic heterogenization for better reproducibility in animal experimentation. *Lab Anim (NY)*. 2017; 46(9):343–349.
32. Shaw R, Miller S, Curwen J, Dymond M. Design, analysis and reporting of tumor models. *Lab Anim (NY)*. 2017;46(5): 207–211.
33. Antonucci TC. Teams do it better! *Res Hum Dev*. 2015;12(3–4): 342–349.
34. Meyerholz DK, Piersigilli A. Animal models: software for study design falls short. *Nature*. 2016;532(7598):177.
35. Treuting PM, Snyder JM, Ikeno Y, Schofield PN, Ward JM, Sundberg JP. The vital role of pathology in improving reproducibility and translational relevance of aging studies in rodents. *Vet Pathol*. 2016;53(2):244–249.
36. Ince TA, Ward JM, Valli VE, et al. Do-it-yourself (DIY) pathology. *Nat Biotechnol*. 2008;26(9):978–979.
37. Cardiff RD, Ward JM, Barthold SW. 'One medicine—one pathology': are veterinary and human pathology prepared? *Lab Invest*. 2008;88(1):18–26.
38. Ward JM, Schofield PN, Sundberg JP. Reproducibility of histopathological findings in experimental pathology of the mouse: a sorry tail. *Lab Anim (NY)*. 2017;46(4):146–151.
39. Wolf JC, Wheeler JR. A critical review of histopathological findings associated with endocrine and non-endocrine hepatic toxicity in fish models. *Aquat Toxicol*. 2018;197: 60–78.
40. Holland T, Holland C. Unbiased histological examinations in toxicological experiments (or, the informed leading the blinded examination). *Toxicol Pathol*. 2011;39(4):711–714.
41. Snyder JM, Ward JM, Treuting PM. Cause-of-death analysis in rodent aging studies. *Vet Pathol*. 2016;53(2):233–243.
42. Kienhofer D, Hahn J, Stoof J, et al. Experimental lupus is aggravated in mouse strains with impaired induction of neutrophil extracellular traps. *JCI Insight*. 2017;2(10):pii: 92920.
43. 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(4):569–575.
44. Thoolen B, Maronpot RR, Harada T, et al. Proliferative and nonproliferative lesions of the rat and mouse hepatobiliary system. *Toxicol Pathol*. 2010;38(7 Suppl):5S–81S.
45. Ward JM, Thoolen B. Grading of lesions. *Toxicol Pathol*. 2011; 39(4):745–746.
46. Meyerholz DK, Tintle NL, Beck AP. Common pitfalls in analysis of tissue scores. *Vet Pathol*. 2019;56(1):39–42.
47. Hasebe T, Okada N, Tamura N, et al. p53 expression in tumor stromal fibroblasts is associated with the outcome of patients with invasive ductal carcinoma of the breast. *Cancer Sci*. 2009;100(11):2101–2108.
48. 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(1):38–43.
49. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med*. 2005;37(5):360–363.
50. Adissu HA, Estabel J, Sunter D, et al. Histopathology reveals correlative and unique phenotypes in a high-throughput mouse phenotyping screen. *Dis Model Mech*. 2014;7(5):515–524.
51. Gibson-Corley KN, Olivier AK, Meyerholz DK. Principles for valid histopathologic scoring in research. *Vet Pathol*. 2013;50 (6):1007–1015.
52. Cross SS. Observer accuracy in estimating proportions in images: implications for the semiquantitative assessment of staining reactions and a proposal for a new system. *J Clin Pathol*. 2001;54(5):385–390.
53. Cross SS. Kappa statistics as indicators of quality assurance in histopathology and cytopathology. *J Clin Pathol*. 1996;49 (7):597–599.
54. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174.
55. Green JW, Springer TA, Saulnier AN, Swintek J. Statistical analysis of histopathological endpoints. *Environ Toxicol Chem*. 2014;33(5):1108–1116.
56. Festing MF. Design and statistical methods in studies using animal models of development. *ILAR J*. 2006;47(1):5–14.
57. Festing MF, Altman DG. Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR J*. 2002;43(4):244–258.
58. Festing MF. The design and statistical analysis of animal experiments. *ILAR J*. 2002;43(4):191–193.