DORDT COLLEGE

Digital Collections @ Dordt

Faculty Work Comprehensive List

8-21-2018

# Common Pitfalls in Analysis of Tissue Scores

David K. Meyerholz
*University of Iowa*

Nathan L. Tintle
*Dordt College*, nathan.tintle@dordt.edu

Amanda P. Beck
*Albert Einstein College of Medicine*

Follow this and additional works at: https://digitalcollections.dordt.edu/faculty_work

 Part of the Pathology Commons

Recommended Citation

Meyerholz, David K.; Tintle, Nathan L.; and Beck, Amanda P., "Common Pitfalls in Analysis of Tissue Scores" (2018). *Faculty Work Comprehensive List*. 1008.
https://digitalcollections.dordt.edu/faculty_work/1008

This Article is brought to you for free and open access by Digital Collections @ Dordt. It has been accepted for inclusion in Faculty Work Comprehensive List by an authorized administrator of Digital Collections @ Dordt. For more information, please contact ingrid.mulder@dordt.edu.

# Common Pitfalls in Analysis of Tissue Scores

**Abstract**
Histopathology remains an important source of descriptive biological data in biomedical research. Recent petitions for enhanced reproducibility in scientific studies have elevated the role of tissue scoring (semiquantitative and quantitative) in research studies. Effective tissue scoring requires appropriate statistical analysis to help validate the group comparisons and give the pathologist confidence in interpreting the data. Each statistical test is typically founded on underlying assumptions regarding the data. If the underlying assumptions of a statistical test do not match the data, then these tests can lead to increased risk of erroneous interpretations of the data. The choice of appropriate statistical test is influenced by the study's experimental design and resultant data (eg, paired vs unpaired, normality, number of groups, etc). Here, we identify 3 common pitfalls in the analysis of tissue scores: shopping for significance, overuse of paired *t*-tests, and misguided analysis of multiple groups. Finally, we encourage pathologists to use the full breadth of resources available to them, such as using statistical software, reading key publications about statistical approaches, and identifying a statistician to serve as a collaborator on the multidisciplinary research team. These collective resources can be helpful in choosing the appropriate statistical test for tissue-scoring data to provide the most valid interpretation for the pathologist.

**Keywords**
grading, lesions, pathology, pitfalls, reproducibility, statistics, scoring, tissues

**Disciplines**
Pathology

This article is available at Digital Collections @ Dordt: https://digitalcollections.dordt.edu/faculty_work/1008

Check for updates

Commentary

# Common Pitfalls in Analysis of Tissue Scores

Veterinary Pathology
1-4
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0300985818794250
journals.sagepub.com/home/vet

**David K. Meyerholz[1], Nathan L. Tintle[2], and Amanda P. Beck[3]**

## Abstract

Histopathology remains an important source of descriptive biological data in biomedical research. Recent petitions for enhanced reproducibility in scientific studies have elevated the role of tissue scoring (semiquantitative and quantitative) in research studies. Effective tissue scoring requires appropriate statistical analysis to help validate the group comparisons and give the pathologist confidence in interpreting the data. Each statistical test is typically founded on underlying assumptions regarding the data. If the underlying assumptions of a statistical test do not match the data, then these tests can lead to increased risk of erroneous interpretations of the data. The choice of appropriate statistical test is influenced by the study's experimental design and resultant data (eg, paired vs unpaired, normality, number of groups, etc). Here, we identify 3 common pitfalls in the analysis of tissue scores: shopping for significance, overuse of paired *t*-tests, and misguided analysis of multiple groups. Finally, we encourage pathologists to use the full breadth of resources available to them, such as using statistical software, reading key publications about statistical approaches, and identifying a statistician to serve as a collaborator on the multidisciplinary research team. These collective resources can be helpful in choosing the appropriate statistical test for tissue-scoring data to provide the most valid interpretation for the pathologist.

## Keywords

grading, lesions, pathology, pitfalls, reproducibility, scoring, statistics, tissues

Cells and tissues are commonly studied in biomedical research to offer biological perspective that can clarify and complement clinical and molecular data. At its fundamental level, histopathological evaluation and description of tissues can be summarized by images in a figure to demonstrate group differences. While morphologic descriptions can serve an important function, these have inherent limitations for distinguishing differences between treatment groups. To combat this, as well as to increase the rigor and repeatability of tissue studies, group changes can be enumerated through semiquantitative and/or quantitative scoring.[3,10] Tissue-scoring data can then be analyzed by appropriate statistical tests to provide a more rigorous level of confidence in the interpretations and conclusions. The aim of this article is to identify 3 common pitfalls of tissue-scoring analysis and offer approaches for the pathologist to avoid these issues.

## Shopping for Significance

For many people entering into biomedical research, there is an immediate and broad exposure to many different approaches and tools for investigational studies. It does not take long to quickly become aware (from reading journal articles and attending lab meetings) that statistical significance is a vital component of most analyses. While true on many levels, this concept can become dangerous when it mistakenly assumes that any form of statistical significance is a good thing. In this frame of mind, selection of a statistical test could become like window shopping to find one that produces the greatest significance (eg, smallest *P*-value). This is a flawed approach.

Statistical tests typically have underlying assumptions about the data that should be met to have confidence in the resulting analysis. If the assumptions for a statistical test are not met, the analysis may be prone to incorrect interpretations. Therefore, the best approach is to select a statistical test that fits the experimental design and data. For instance, one common question is whether the data fulfill the assumptions of parametric (eg, continuous data, normal distribution) or nonparametric (eg, discontinuous data or lack of normal distribution) tests, to help guide the selection of a statistical analysis.

[1]Department of Pathology, University of Iowa Carver College of Medicine, Iowa City, IA, USA
[2]Department of Mathematics, Statistics and Computer Science, Dordt College, Sioux Center, IA, USA
[3]Department of Pathology, Albert Einstein College of Medicine, Bronx, NY, USA

**Corresponding Authors:**
David K. Meyerholz, Department of Pathology, University of Iowa Carver College of Medicine, Iowa City, Iowa 52242, USA.
Email: david-meyerholz@uiowa.edu
Amanda P. Beck, Department of Pathology, Albert Einstein College of Medicine, Bronx, NY 10461, USA.
Email: amanda.beck@einstein.yu.edu

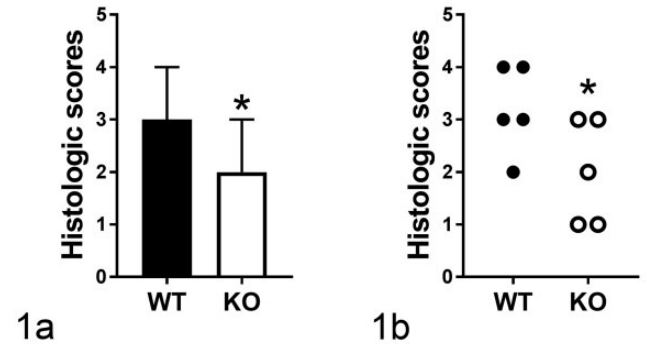


**Figure 1.** Mock ordinal tissue-scoring data comparing wild-type (WT) and knockout (KO) groups in a bar graph (a; bar = median with 95% confidence interval) or as a dot plot (b). *$P = .004$, paired *t*-test. GraphPad Prism Software, v7.03 (GraphPad Software, La Jolla, CA; www.graphpad.com) was used for all statistical analysis in this article.

**Table 1.** Mock Ordinal Scoring Data From Wild-Type and Knockout Animals ($n = 5$/Group) From Figure 1 With Parametric (Paired *t*-Test, Unpaired *t*-Test) and Nonparametric (Mann Whitney *U*-test, Kolmogorov-Smirnov test) Statistical Analyses[a]

| Wild-Type | Knockout |
|---|---|
| 4 | 3 |
| 4 | 3 |
| 3 | 2 |
| 3 | 1 |
| 2 | 1 |

[a] Statistical analyses: paired *t*-test ($P = .004$), unpaired *t*-test ($P = .074$), Mann-Whitney *U*-test ($P = .143$), and Kolmogorov-Smirnov test ($P = .810$).

To show how selection of a statistical test is important for the final analysis, we will review and then deconstruct an experiment comparing 2 groups of data. In this mock example, we show both a bar graph and dot plot of the same ordinal tissue-scoring data between wild-type (WT) and knockout (KO) groups of mice (Fig. 1a, b). If Fig. 1a were the only figure shown, it would be very difficult for the reader/reviewer to know the number of mice, the study design, and the meaning of the error bar (in this case, it is the margin of error). Transparency is one way to avoid errors, and in this case, one can show the data in a dot plot to be more transparent to reviewers and readers (Fig. 1b). Assuming someone were shopping for a statistical test, they might screen several tests (eg, Mann-Whitney *U*-test, Kolmogorov-Smirnov, unpaired *t*-test, paired *t*-test, etc) to find that the paired *t*-test was significant and choose it (Fig. 1; Table 1). However, the data from this experiment are ordinal, not continuous. Ordinal data are the most common type of semiquantitative scoring data in pathology research, and in this case, they were assigned from a grading system that included 5 grades: 0, 1, 2, 3, and 4. When using ordinal data, nonparametric tests are often recommended.[4,7,23]

Table 1 illustrates the results of nonparametric tests on the mock data to demonstrate how easy it is to make errors. Notice that the 4 different tests give 4 very different *P*-values! Had we shopped around and chose the paired *t*-test because it had a significant difference, we would be ignoring the fact that the different tests are making very different assumptions about how the data were generated and, as noted above, the ordinal nature of the data. In reality, the large *P*-values for the Mann-Whitney *U*-test and Kolmogorov-Smirnov test (which are appropriate for these data, whereas the *t*-tests are not) suggest that these data could have easily been generated randomly. For example, assume that the Table 1 data were instead generated through rolling 2 sets of dice ($n = 5$ samples/group), using the right hand for one group (see "WT" data) and using the left hand for the other group (see "KO" data). The values in this mock data set are in line with the kind of results that would happen by rolling dice (chance) versus an actual effect. If we used appropriate nonparametric tests, these mock data would have shown no significant differences in rolling dice between the right and left hand, or rather, that the outcome is in line with being a random event.

What are other ways can we increase our confidence in the data besides using valid scoring systems and appropriate statistical analyses? We could repeat the experiment to show that multiple replicates have similar tendencies and thereby also increase the sample size to strengthen our statistical analysis. Another way we can increase our confidence in the data is to corroborate these data to other biologically relevant data collected from the same animals. For instance, the severity of acute inflammatory lesions in lung tissue between 2 groups of animals might be corroborated by changes in a complete blood count (eg, neutrophilia).[8]

## Overuse of Paired *t*-Tests

The example above highlights a frequent mistake seen in the comparison of 2 groups: that of using paired *t*-tests to compare unpaired treatment groups.[22] If treatment groups are independent of each other, meaning that the samples in one group are not linked in any way to samples in the other group, the data are unpaired (Fig. 2a). Conversely, if the treatment groups are dependent, meaning the samples in one group are linked to the samples in the other group, the data are paired. In pathology studies, common examples of paired data include (1) repeated measures taken on the same animal at different times (Fig. 2b) or (2) two different treatments performed on the same animal (or tissue) with similar endpoints (Fig. 2c). Importantly, data for paired samples are linked and must be analyzed in a related fashion. If an accidental switch in data entry order occurs, it can radically influence the analysis of paired *t*-tests, whereas entry order for unpaired *t*-tests does not matter because the samples are not linked (Table 2).

## Misguided Analysis of Multiple Groups

While many pathology studies are composed of 2 basic test groups (eg, a control and treated group), as in the above examples, other studies are more complex, such as those involving

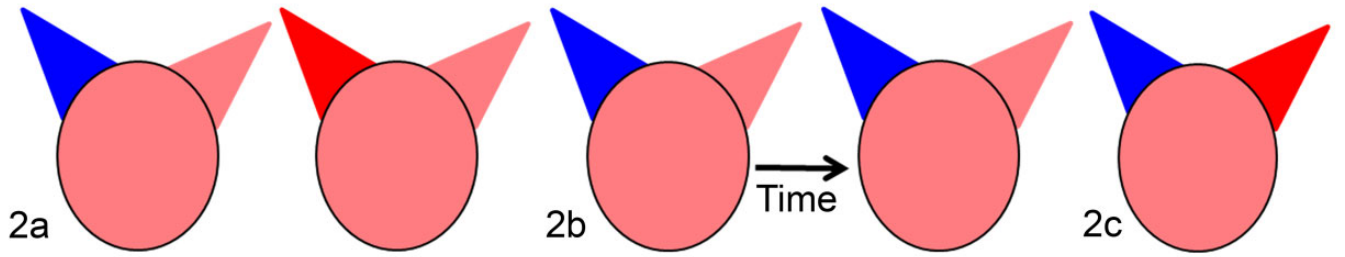


**Figure 2.** Three examples of study designs to evaluate topical compounds (identified as red or blue) applied on the ears of pigs to assess epidermal injury from biopsied skin. (a) In an unpaired design, 2 groups of pigs were compared. Each group of pigs had a unique treatment (red or blue). (b) In a paired design using repeated measures, each animal had the same treatment, but tissue biopsy data were collected at 2 different times. Each paired comparison (ie, early and late time points) was from the same pig. (c) In a paired design not using repeated measures, 2 distinct treatments were applied to the ears of each pig. Each paired comparison (red and blue ears) was from the same pig.

**Table 2.** Mock Quantitative Scoring Data for Cohorts 1 and 2[a]

| (A) Cohort 1[b] | | (B) Cohort 2[c] | |
|---|---|---|---|
| WT | KO | WT | KO |
| 4.1 | 3.7 | 4.1 | **3.2** |
| 3.6 | 3.5 | 3.6 | 3.5 |
| 3.5 | 3.2 | 3.5 | **3.7** |
| 3.1 | 2.6 | 3.1 | 2.6 |
| 2.9 | 2.8 | 2.9 | 2.8 |

Abbreviations: KO, knockout; WT, wild-type.
[a] The only difference between cohorts 1 (A) and 2 (B) are that 2 KO values are switched (see bold values in the far-right column). This difference results in significant changes in paired *t*-tests (assumes linked data), but this assumption is not critical for the unpaired *t*-tests (assumes independent data).
[b] Paired *t*-test ($P = .025$) and unpaired *t*-test ($P = .368$).
[c] Paired *t*-test ($P = .216$) and unpaired *t*-test ($P = .368$).

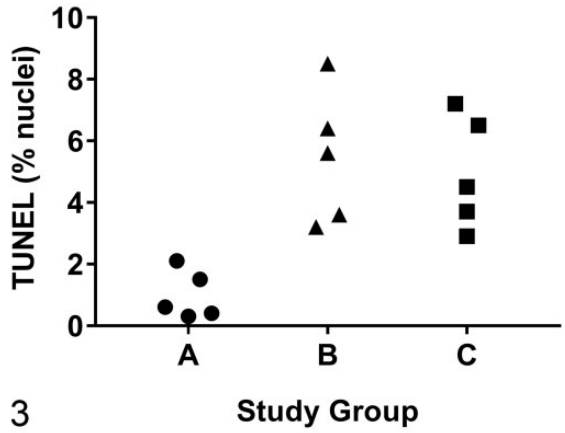


**Figure 3.** Mock example analyzing a treatment applied to 3 independent groups (A–C). The one-way analysis of variance was significant ($P = .0024$), suggesting that the means for the groups were not all equal. Further evaluation of post hoc tests (eg, Tukey's tests) for specific group comparisons showed the following results: A versus B ($P = .0034$), A versus C ($P = .0078$), and B versus C ($P = .8885$).

multiple (3 or more) treatment groups. In these cases, investigators may try to apply a series of *t*-tests to make group-to-group comparisons. This is not valid because *t*-tests are designed for studies that compare 2 groups. The added complexity of having multiple groups requires different types of statistical analysis.

For example, if the data support parametric tests, one approach is to evaluate the groups using analysis of variance (ANOVA); this test initially evaluates whether the mean of each group is the same (Fig. 3). If the ANOVA results in significance, the interpretation from the evidence is that the group means are not all equal. A nonparametric correlate to the one-way ANOVA is the Kruskal-Wallis test. After performing an ANOVA or Kruskal-Wallis test and finding evidence that at least one group mean is different from the rest, post hoc evaluations comparing pairs of groups can be conducted (Table 3).

While we have briefly discussed the most common approaches to analyze tissue scores, these are not to be viewed as dogmatic recommendations, as it is important to recognize that there are several ways to analyze tissue-scoring data. The approaches we have described are commonly used and published in the literature. Sometimes, the data, experiment designs, or questions being asked can increase the permutations

**Table 3.** Examples of Common Statistical Tests for Various Group Comparisons[a]

| Comparison | Parametric | Nonparametric |
|---|---|---|
| Two dependent groups | Paired *t*-test | Wilcoxon matched pairs signed-rank test |
| Two independent groups | Unpaired *t*-test | Mann-Whitney *U*-test, Kolmogorov-Smirnov test |
| Three or more independent groups | One-way analysis of variance test with Tukey's post hoc tests | Kruskal-Wallis test with Dunn's post hoc tests |

[a] Adapted from statistical software design for researchers (GraphPad Prism Software, v7.03, GraphPad Software, La Jolla, CA, www.graphpad.com) and from Kim, 2014.[6]

and complexity of the study, so other statistical approaches might be better used and advised by the statistician.

## Summary

We have highlighted several approaches to avoid slipping into common pitfalls when analyzing tissue scores. Pathologists who perform tissue scoring should have access to fundamental statistical resources, and several are available. In recent years, statistical software platforms have become increasingly more user-friendly and as such are common tools used in biomedical publications.[4,23] Several published resources (eg, books or articles) are also available to learn more about tissue scoring, experimental design, and statistical analysis.[2,6,7,9,11–21,23] Last, but not least, pathologists should secure the professional expertise and collaboration of a statistician as a part of the multidisciplinary team.[1,5,24]

### Declaration of Conflicting Interests

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

### Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

### ORCID iD


David K. Meyerholz http://orcid.org/0000-0003-1552-3253

### References

1. Antonucci TC. Teams do it better! *Res Hum Dev*. 2015;**12**(3–4):342–349.
2. Festing MF. Design and statistical methods in studies using animal models of development. *ILAR J*. 2006;**47**(1):5–14.
3. Gibson-Corley KN, Olivier AK, Meyerholz DK. Principles for valid histopathologic scoring in research. *Vet Pathol*. 2013;**50**(6):1007–1015.
4. Imai DM, Pesapane R, Conroy CJ, et al. Apical elongation of molar teeth in captive microtus voles. *Vet Pathol*. 2018;**55**(4):572–583.
5. Kilkenny C, Parsons N, Kadyszewski E, et al. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *Plos One*. 2009;**4**(11):e7824.
6. Kim HY. Statistical notes for clinical researchers: nonparametric statistical methods: 1. Nonparametric methods for comparing two groups. *Restor Dent Endod*. 2014;**39**(3):235–239.
7. Kim HY. Statistical notes for clinical researchers: nonparametric statistical methods: 2. Nonparametric methods for comparing three or more groups and repeated measures. *Restor Dent Endod*. 2014;**39**(4):329–332.
8. 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 U S A*. 2017;**114**(15):E3119–E3128.
9. Meyerholz DK, Beck AP. Principles and approaches for reproducible scoring of tissue stains in research. *Lab Invest*. 2018;**98**:844–855.
10. Meyerholz DK, Sieren JC, Beck AP, et al. Approaches to evaluate lung inflammation in translational research. *Vet Pathol*. 2018;**55**(1):42–52.
11. Olsen CH. Review of the use of statistics in infection and immunity. *Infect Immun*. 2003;**71**(12):6689–6692.
12. Petrie A, Watson PF. *Statistics for Veterinary and Animal Science*. Oxford, UK: Blackwell Science; 1999.
13. Shott S. Comparing means or distributions. *J Am Vet Med Assoc*. 2011;**238**: 1422–1428.
14. Shott S. Comparing percentages. *J Am Vet Med Assoc*. 2011;**238**(9): 1122–1125.
15. Shott S. Designing studies that answer questions. *J Am Vet Med Assoc*. 2011; **238**(1):55–58.
16. Shott S. Detecting statistical errors in veterinary research. *J Am Vet Med Assoc*. 2011;**238**(3):305–308.
17. Shott S. Relationships between categorical dependent variables and other variables and between waiting times and other variables. *J Am Vet Med Assoc*. 2011; **239**(3):322–327.
18. Shott S. Relationships between more than two variables. *J Am Vet Med Assoc*. 2011;**239**(5):587–593.
19. Shott S. Relationships between two categorical variables and between two noncategorical variables. *J Am Vet Med Assoc*. 2011;**239**(1):70–74.
20. Shott S. Statistics simplified: describing data. *J Am Vet Med Assoc*. 2011; **238**(5):588–591.
21. Shott S. Testing ideas and estimating clinical importance. *J Am Vet Med Assoc*. 2011;**238**(7):871–876.
22. Skaik Y. The bread and butter of statistical analysis "t-test": uses and misuses. *Pak J Med Sci*. 2015;**31**(6):1558–1559.
23. Vrolyk V, Wobeser BK, Al-Dissi AN, et al. Lung inflammation associated with clinical acute necrotizing pancreatitis in dogs. *Vet Pathol*. 2017;**54**(1): 129–140.
24. Zeiss CJ, Ward JM, Allore HG. Designing phenotyping studies for genetically engineered mice. *Vet Pathol*. 2012;**49**(1):24–31.