Received: 21 March 2024 | Accepted: 8 May 2024

DOI: 10.1111/rda.14614

Reproduction in Domestic Animals WILEY

SHORT COMMUNICATION

# Development of an equine endometrial histology scoring system to complement the Kenney–Doig scale

I. Martínez-Bartolomé[1] | J. Masot[2] | C. Serres[1]

[1]Animal Medicine and Surgery Department, Veterinary Faculty, UCM, Madrid, Spain

[2]Facultad de Veterinaria, UEX, Cáceres, Spain

**Correspondence**
C. Serres, Animal Medicine and Surgery Department, Veterinary Faculty, UCM, Avda. Puerta de Hierro s/n, Madrid 28040, Spain.
Email: cserres@ucm.es

**Funding information**
Yeguada de Ymas

## Abstract

Kenney–Doig scale is considered the international standard method for classifying uterine biopsies in mares; however, its objectivity has been questioned by various studies. In the present study, we analysed the degree of agreement between two pathologists when assessing the same set of 201 uterine biopsies, obtaining a slight to moderate level of agreement ($\kappa = .34/\kappa_w = .57$). Subsequently, we developed a numerical scale based on the evaluation of histological parameters, including inflammation, fibrosis, glandular density and lymphatic lacunae. Partial scores were summed to obtain a fifth parameter called Summation. The correlation between both scales was demonstrated ($p < .0001$), and their combined use resulted in a notable increase in the degree of agreement between the two pathologists ($\kappa = .53/\kappa_w = .67$).

**KEYWORDS**
endometrial biopsy, equine, Kenney–Doig, pathologist variation, repeatability

## 1 | INTRODUCTION

Infertility and subfertility in mares are commonly caused by uterine pathologies (Tibary & Ruiz, 2018). To assess uterine histologic changes, a categorization (grading) scale proposed by Kenney and Doig (1986) is the international standard. The endometrium is classified into four categories based on the severity of the lesions.

The Kenney–Doig scale, widely used in equine endometrial biopsy evaluation, has been criticized by some authors for its lack of objectivity and limited guidelines (Evans, 1998; Snider et al., 2011). In a recent study, the scale's inter-rater and intra-rater repeatability was assessed, concluding that it exhibits poor repeatability (Westendorf et al., 2021).

In medical histopathology, nominal classification, as described by Kenney and Doig, may not provide sufficient information for the referring clinician. To address this issue, well-defined scoring systems have been developed, which provide more detailed information and reduce variability between pathologists. These ordinal systems are particularly useful in research (Cross, 1998; Gibson-Corley et al., 2013).

The study aims to evaluate the agreement between two pathologists in assessing 201 endometrial biopsies using the Kenney and Doig scale. An ordinal assessment system based on the authors' parameters was developed, and the correlation between both scales was examined. Additionally, the agreement between the pathologists was assessed for 26 histological slides using both scales.

## 2 | MATERIALS AND METHODS

### 2.1 | Animals

The Animal Experimentation Ethics Committee of the Veterinary Teaching Hospital of UCM approved and reviewed this study (Ref. 09/2022). The study used 26 mares aged between 14 and 24 years to ensure a wide range of histological lesion severity (Kabisch et al., 2019).

During four consecutive estrous cycles, 201 biopsies were obtained from both uterine horns to increase variability in histological findings (Muderspach et al., 2024). Slides were stained with haematoxylin–eosin and Masson's Trichrome.

## 2.2 | Histological evaluation procedures

Experiment 1: two independent pathologists (1 and 2) evaluated 201 biopsies according to Kenney–Doig guidelines without considering the history of barrenness. The agreement between their observations was assessed.

Experiment 2: pathologist 1 evaluated the 201 biopsies according to Kenney–Doig guidelines after separately assessing inflammation, fibrosis, glandular density and lymphatic lacunae using a scoring scale. The Summation, which is the total of the partial scores assigned to each parameter (see Table 1), was calculated to evaluate the correlation between the Summation and Kenney–Doig Category.

Experiment 3: pathologist 2 re-evaluated 26 randomly selected slides, considering both the numerical scale and the Kenney–Doig classification. The agreement of their observations with those of pathologist 1 in Experiment 2 was assessed.

## 2.3 | Statistical analysis

Inter-observer agreement was examined using Cohen's kappa statistics. Unweighted kappa values (κ) measure the amount of agreement without accounting for the number of categories pathologists may differ by; weighted kappa ($\kappa_w$) values were used to examine the magnitude of agreement by placing more emphasis on a disagreement of more than one Kenney–Doig category and awarding partial agreement when differing by only one category. Kappa interpretation ranges from poor (<.00) to almost perfect agreement (.81–1.00) (Westendorf et al., 2021).

ANOVA and Tukey's post hoc tests analysed Summation's correlation with Kenney–Doig categories.

The predictive ability of Summation was assessed using the Classification and Regression Trees (CRT), targeting Kenney–Doig Category.

**TABLE 1** Ordinal classification system.

| Parameter | Description | | Score |
|---|---|---|---|
| Inflammation | No inflammation | | 0 |
| | Slight: Isolated inflammatory cells, without forming defined foci (less than 1% of the sample) | | 1 |
| | Moderate: isolated foci (1%–25% of the sample) | | 2 |
| | Moderate to severe: abundant inflammatory cells in large foci (25%–50% of the sample) | | 3 |
| | Severe: diffused inflammatory cells (>50% of the sample) | | 4 |
| Fibrosis | Individual gland branches | No fibrosis | 0 |
| | | Isolated foci, 1–3/LPF | 1 |
| | | Widespread foci, >4/LPF | 2 |
| | Nests | Absent | 0 |
| | | Isolated foci, 1–3/LPF | 1 |
| | | Widespread foci, >4/LPF | 2 |
| | Number periglandular layers | No fibrosis | 0 |
| | | Moderate fibrosis. 1–3 cell layers | 1 |
| | | Severe fibrosis. >4 cell layers | 2 |
| | Total fibrosis | | |
| Glandular density | Normal | | 0 |
| | Focally decreased | | 1 |
| | Multifocally decreased | | 2 |
| | Focally absent glands | | 3 |
| | Multifocally absent glands | | 4 |
| Lymphatic Lacunae | No | | 0 |
| | Yes | | 1 |
| Summation | | | |

*Note*: Lesion frequency is assessed at 100× magnification.
Abbreviation: LPF, low power field.

Biopsies were grouped by age: middle-age (14–17 years) and old (18–24 years) to examine age impacts.

## 3 | RESULTS

### 3.1 | Experiment 1

Observation of the same set of 201 uterine biopsies by two pathologists revealed differences in classification according to the Kenney–Doig scale (Table 2).

Out of the 201 biopsies, 120 coincided with the assigned histological category. The Cohen's kappa coefficient was calculated, resulting in .34 (slight agreement) for unweighted and .57 (moderate agreement) for weighted. The agreement was lower in the middle-aged group ($\kappa = .17$, $\kappa_w = .31$) than in the old mares group ($\kappa = .42$, $\kappa_w = .67$).

### 3.2 | Experiment 2

Each sample was classified and assigned a numerical value for the Summation and a Kenney–Doig category. Statistical analysis revealed significant differences ($p < .0001$) in the mean Summation values for each Kenney–Doig category. Table 3 displays the mean Summation values for each histological category.

The predictive numerical ranges for the Kenney–Doig categories are as follows: Category I [1–3], Category IIA [4–6], Category IIB [7–8], and Category III [9–11]. The overall accuracy rate of the Summation in predicting the Kenney–Doig category was 83.3%.

### 3.3 | Experiment 3

Of the 26 biopsies selected randomly, nine belonged to middle-aged mares and 17 belonged to old mares. Nineteen of the 26 biopsies received the same histological category from both pathologists, with no disagreements between non-adjacent Kenney–Doig categories in any samples. The agreement was moderate (unweighted kappa coefficient = .53) and good (weighted kappa coefficient = .67).

## 4 | DISCUSSION

In a study comparing 63 uterine biopsies, Westendorf et al. (2021) found only slight agreement among eight pathologists ($\kappa = .19$). Our study showed slightly higher agreement ($\kappa = .34$). Like Westendorf et al., we used the weighted Cohen's kappa coefficient ($\kappa_w = .57$), which was higher. This indicates an improvement in agreement when considering the magnitude of disagreement. Most disagreements occurred between adjacent Kenney–Doig histological categories.

Another finding we observed was the high frequency of biopsies assigned to middle categories, also described by Westendorf et al. (2021). Pathologists tend to avoid extreme categories when using nominal classification systems (Cross, 1998; Kiupel et al., 2011; Northrup et al., 2005). The Kenney–Doig scale (Kenney & Doig, 1986) associates each category with a wide range of prognoses regarding the mare's ability to carry a pregnancy to term. Even a one-degree disagreement can cause a 40% variation in the estimated probability. Categories IIA and IIB gravitate around 50%, which gives the pathologist greater confidence in the clinical implications of their study. The tendency to avoid extremes is influenced by vaguely defined classification guidelines, especially in middle categories that easily

**TABLE 2** Frequency distributions of Kenney–Doig categories assigned by two pathologists with kappa coefficients.

| | All slides $n = 201$ | | | | 14–17 years slides, $n = 70$ | | | | 18–24 years slides, $n = 131$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Category | Pathologist 1 | | Pathologist 2 | | Pathologist 1 | | Pathologist 2 | | Pathologist 1 | | Pathologist 2 | |
| I | 3.50% | | 6.50% | | 1.4% | | 1.4% | | 4.6% | | 9.2% | |
| IIA | 44.30% | 87.6% | 67.20% | 87.1% | 42.9% | 90.0% | 78.6% | 98.6% | 45.0% | 86.3% | 60.3% | 80.9% |
| IIB | 43.30% | | 19.90% | | 47.1% | | 20.0% | | 41.2% | | 20.6% | |
| III | 9.0% | | 6.50% | | 8.6% | | 0.0% | | 9.2% | | 9.9% | |
| $\kappa$ | .34 | | | | .17 | | | | .42 | | | |
| $\kappa_w$ | .57 | | | | .31 | | | | .67 | | | |

Abbreviation: *n*, number of biopsies.

**TABLE 3** Mean values obtained for the Summation within each histological category.

| | All samples, $n = 201$ | | 14–17 years, $n = 70$ | | 18–24 years, $n = 131$ | |
|---|---|---|---|---|---|---|
| Category | *n* | Mean summation ± SD | *n* | Mean summation | *n* | Mean summation |
| I | 7 | 2.57 ± 0.98 | 1 | 2 | 6 | 2.66 |
| IIA | 89 | 5.11 ± 0.87 | 30 | 5 | 59 | 5.25 |
| IIB | 87 | 7.22 ± 0.92 | 33 | 7.3 | 54 | 7.28 |
| III | 18 | 9.06 ± 1.26 | 6 | 9.16 | 12 | 8.92 |

Abbreviations: *n*, number of biopsies; SD, standard deviation.

overlap (Westendorf et al., 2021), as can be deduced from the lower degree of agreement observed in the middle-aged biopsy group linked to a higher frequency of middle categories (IIA and IIB).

The study's observations support the idea that the Kenney–Doig scale has significant limitations in providing an objective assessment of endometrial histopathology.

Based on our literature review, the scale used in this study is currently the only numerical not morphometric method for evaluating endometrial tissue in mares. The scale involves considering parameters described by Kenney and Doig, each assigned its own scoring system. These scoring methods are known as 'splitters', which offer higher repeatability and sensitivity to specific parameters or changes (Gibson-Corley et al., 2013). According to Shackelford et al. (2002), research indicates that the detection and repeatability may be optimized by using four to five levels of scoring.

The study defined the number of divisions within each parameter based on the specific weight assigned by Kenney–Doig in their scale. The aim was for each parameter's contribution to the Summation to be proportional to the severity of the lesions. The study demonstrated the correlation between both scales, and the Summation value showed a high predictive capacity for Kenney–Doig categories. Age did not affect the mean summation values in the histological categories; however, these results should be interpreted with caution due to the few samples in categories I and III.

The combined use of the new scale as a complementary tool to the Kenney–Doig scale resulted in an increased degree of agreement between pathologists. According to McHugh (2012), kappa values below .60 provide little confidence in the evaluated method, although not standardized. In our study, we achieved results close to this value, which is notably better than those described in the literature (Westendorf et al., 2021). Further studies with a larger number of biopsies are needed to confirm these findings.

The numerical scale could be particularly useful in studies assessing sequential changes in endometrium histopathology or in the evaluation of paired biopsies in mares over 17 years old, where degenerative changes may not be uniform (Muderspach et al., 2024).

## 5 | CONCLUSIONS

Based on the obtained data, using the numerical scale in conjunction with the Kenney–Doig scale improved its objectivity. To validate these observations, a larger number of pathologists and samples will be necessary.

### AUTHOR CONTRIBUTIONS

I. Martínez-Bartolomé and C. Serres conceived the study, participated in its design, and wrote the manuscript. I. Martínez-Bartolomé carried out sample collection, biopsy classification, and data interpretation. J. Masot carried out the anatomopathological processing and biopsy classification. All authors read and approved the final manuscript.

### ACKNOWLEDGEMENTS

Special thanks for Yeguada de Ymas for all the support and Pedro Cuesta for his statistical assistance.

### CONFLICT OF INTEREST STATEMENT

None of the authors have any conflict of interest to declare.

### DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

### ORCID

*I. Martínez-Bartolomé* https://orcid.org/0009-0007-3024-6553

### REFERENCES

Cross, S. S. (1998). Grading and scoring in histopathology. *Histopathology*, *33*, 99–106.

Evans, T. J. (1998). Morphometric analysis of endometrial periglandular fibrosis in mares. *American Journal of Veterinary Research*, *59*, 1209–1214.

Gibson-Corley, K. N., Olivier, A. K., & Meyerholz, D. K. (2013). Principles for valid histopathologic scoring in research. *Veterinary Pathology*, *50*, 1007–1015.

Kabisch, J., Klose, K., & Schoon, H.-A. (2019). Endometrial biopsies of old mares – What to expect?! *Pferdeheilkunde Equine Medicine*, *35*, 211–219.

Kenney, R. M., & Doig, P. A. (1986). Equine endometrial biopsy. *Current Therapy in Theriogenology*, *2*, 723–729.

Kiupel, M., Webster, J. D., Bailey, K. L., Best, S., DeLay, J., Detrisac, C. J., Fitzgerald, S. D., Gamble, D., Ginn, P. E., Goldschmidt, M. H., Hendrick, M. J., Howerth, E. W., Janovitz, E. B., Langohr, I., Lenz, S. D., Lipscomb, T. P., Miller, M. A., Misdorp, W., Moroff, S., … Miller, R. (2011). Proposal of a 2-tier histologic grading system for canine cutaneous mast cell tumors to more accurately predict biological behavior. *Veterinary Pathology*, *48*, 147–155.

McHugh, M. L. (2012). Lessons in biostatistics interrater reliability: The kappa statistic. *Biochemia Medica*, *22*, 276–282.

Muderspach, N. D., Troedsson, M. H. T., Ferreira-Dias, G., Agerholm, J. S., & Christoffersen, M. (2024). Distribution of degenerative changes in the equine endometrium as observed in a single versus two biopsies. *Theriogenology*, *213*, 52–58.

Northrup, N. C., Howerth, E. W., Harmon, B. G., Brown, C. A., Carmicheal, K. P., Garcia, A. P., Latimer, K. S., Munday, J. S., Rakich, P. M., Richey, L. J., Stedman, N. L., & Gieger, T. L. (2005). Variation among pathologists in the histologic grading of canine cutaneous mast cell tumors with uniform use of a single grading reference. *Journal of Veterinary Diagnostic Investigation*, *17*, 561–564.

Shackelford, C., Long, G., Wolf, J., Okerberg, C., & Herbert, R. (2002). Qualitative and quantitative analysis of nonneoplastic lesions in toxicology studies. *Toxicologic Pathology*, *30*, 93–96.

Snider, T. A., Sepoy, C., & Holyoak, G. R. (2011). Equine endometrial biopsy reviewed: Observation, interpretation, and application of histopathologic data. *Theriogenology*, *75*, 1567–1581.

Tibary, A., & Ruiz, A. (2018). Uterine disorders in the mare: diagnostic approach, treatment and prevention. *Spermova*, *8*, 1–24.

Westendorf, J., Wobeser, B., & Epp, T. (2021). IIB or not IIB, part 2: Assessing inter-rater and intra-rater repeatability of the Kenney–Doig scale in equine endometrial biopsy evaluation. *Journal of Veterinary Diagnostic Investigation*, *34*, 215–225.

**How to cite this article:** Martínez-Bartolomé, I., Masot, J., & Serres, C. (2024). Development of an equine endometrial histology scoring system to complement the Kenney–Doig scale. *Reproduction in Domestic Animals*, *59(Suppl. 3)*, e14614. https://doi.org/10.1111/rda.14614