Interobserver reproducibility of histological grading of canine simple mammary carcinomas

Santos, M; Correia-Gomes, C; Santos, A; de Matos, A; Dias-Pereira, P; Lopes, C

Published in:
Journal of Comparative Pathology

DOI:
10.1016/j.jcpa.2015.04.005

Print publication: 01/01/2015

Document Version
Peer reviewed version

NEOPLASTIC DISEASE

Short Title: Grading Canine Simple Mammary Carcinomas

Interobserver Reproducibility of Histological Grading of Canine Simple Mammary Carcinomas

M. Santos[*], C. Correia-Gomes[†], A. Santos[‡], A. de Matos[§, ¶], P. Dias-Pereira[‖] and C. Lopes[‖]

[*]Department of Microscopy, Instituto Ciências Biomédicas Abel Salazar, University of Porto, Porto, Portugal, [†]Epidemiology Research Unit, Future Farming Systems, Scotland's Rural College, UK, [‡]Faculty of Veterinary Medicine, Lusófona University of Humanities and Technologies, Lisbon, [§]Department of Veterinary Clinics, Instituto Ciências Biomédicas Abel Salazar, University of Porto, Porto, [¶]Animal Science and Study Central, Food and Agrarian Sciences and Technologies Institute, University of Porto, Porto and [‖]Department of Pathology and Molecular Immunology, Instituto Ciências Biomédicas Abel Salazar, University of Porto, Porto, Portugal

Correspondence to: M. Santos (e-mail: mssantos@icbas.up.pt).

Summary

Histological grading of canine mammary carcinomas (CMCs) has been performed using an adaptation of the human Nottingham method. The histological grade could be a prognostic factor in CMC; however, no data are available concerning interobserver variability in grading. In this study we analyzed the interobserver reproducibility between three observers when assigning individual parameter scores and grade to 46 CMCs. The influence of tumour size and vascular invasion and/or lymph node metastases on the odds of grading disagreement was also evaluated. The mean kappa values were 0.71, 0.51, 0.69 and 0.70 for tubule formation, nuclear pleomorphism, mitotic counts and grade, respectively. There was moderate to good agreement in scoring parameters and tumour grading, with nuclear pleomorphism being least reproducible. These findings are similar to those of human studies. The odds of grading disagreement increased with tumour size, but decreased with the presence of vascular invasion and/or lymph node metastases. Individual scoring differences were moderated by reaching a consensus between two observers.

Keywords: canine mammary tumours; grade; reproducibility

The Nottingham histological grade (NHG), the standard method for scoring human breast tumours, has been adapted for grading canine mammary carcinomas (CMCs) (Karayannopoulou *et al*., 2005; Peña *et al*., 2013). The NHG is composed of the sum of scores assigned to three morphological features (i.e. tubule formation, nuclear pleomorphism and mitotic count), each taking a value of one to three points (Elston and

Ellis, 1991). A total score ≤5 points , 6–7 points or 8–9 points denotes grades I, II and III carcinomas, respectively (Elston and Ellis, 1991).

Although histological grading is used widely in the evaluation of CMCs, there are no studies focusing on the reproducibility of grading. This contrasts markedly with human pathology, where the reproducibility of grading methods, including the NHG, has been debated for years (Stenkvist *et al.*, 1979; Frierson *et al.*, 1995; Robbins *et al.*, 1995; Dalton *et al.*, 2000; Meyer *et al.*, 2005). Furthermore, the measurement of interobserver variability in veterinary oncology is considered critical to validate prognostic markers (Webster *et al*., 2011).

In human studies, the agreement between observers has been estimated in three different ways: percentage of equal judgments, Cohen's kappa statistics (κ) and Spearman's correlation coefficient (Stenkvist *et al.*, 1979; Longacre *et al.*, 2006). Each of these statistical methods has limitations and pitfalls; reporting all three may provide a better reproducibility assessment (Stenkvist *et al.*, 1979). The agreement percentage is intrinsically dependent on the number and frequency of the classifying categories (Stenkvist *et al.*, 1979). The κ statistic implies the assumption that categories have the same width and the so-called 'κ paradox' may occur, namely when the frequencies of categories are clearly unbalanced (Sim and Wright, 2005). In those cases, the proportion of agreement may be high, but the κ value could be low and an interpretation based solely on the κ value would lead to erroneous conclusions (Sim and Wright, 2005). Still, some controversy exists regarding the best κ value (i.e. weighted or unweighted) to be used in breast cancer grade reproducibility studies (Chowdhury *et al.*, 2007). The unweighted κ value (κ_u) considers all types of disagreements as equal, independently of their magnitude (Sim and Wright, 2005). In contrast, the weighted κ value (κ_w)

emphasizes large differences between ratings in ordinal scales (Sim and Wright, 2005). It should be noted that recent guidelines in veterinary oncology recommend the use of κ_w statistics (Webster *et al*., 2011).

The aim of this study was to determine the interobserver agreement in grading simple CMCs and in scoring each grading parameter, using the NHG. Additionally, the influence of clinicopathological parameters (i.e. tumour size, vascular invasion and/or lymph node and tumor progression) on the odds of grading disagreement was estimated.

Pathology archives from the Instituto Ciências Biomédicas Abel Salazar, University of Porto, Portugal were accessed to retrospectively select 46 spontaneously arising simple CMCs that had been removed surgically. The selection of cases and their histological study were blinded with respect to clinical data. For 30 cases follow-up data were collected prospectively over 2 years following the protocol detailed in Santos *et al.* (2013). Owners gave informed consent for both surgery and follow-up.

The histological diagnosis was reviewed by two observers to confirm that all cases fulfilled the criteria for simple carcinomas (i.e. tumours composed of luminal epithelial cells) (Misdorp *et al*., 1999; Goldschmidt *et al*., 2011). For each case, tumour size (i.e. largest diameter) and histological evidence of vascular invasion and/or regional lymph node metastases were recorded. Slides containing the largest cross section were used for grading. Three observers from the same institution (a MSc veterinary pathologist with 10 years of experience, a PhD veterinary pathologist with more than 15 years of experience, both with special interest in canine mammary pathology, and an emeritus Professor of human pathology with more than 40 years of experience) graded all of the tumors independently, using the NHG (Elston and Ellis, 1991; Karayannopoulou *et al*., 2005). Briefly, tubule formation was scored as 1, 2 or 3 when

more than 75%, 10–75% or <10% of neoplastic cells, respectively, were arranged in structures exhibiting an obvious lumen. Nuclear pleomorphism was scored as follows: score 1 denoted a slight increase in variability of nuclear size and shape, compared with normal surrounding epithelial cells; score 2 denoted moderate variation in nuclear size and shape; score 3 denoted marked variation in nuclear size and shape, with very large and bizarre forms. Mitotic figures were counted in 10 high-power fields (×400) and scored using the cut-offs defined by the field diameter of the microscope (field diameter of 0.55 mm; field area of 0.238 mm^2; score 1, ≤8 mitotic figures; score 2, 9–17 mitotic figures; score 3, ≥18 mitotic figures); thus assuring equivalence with assessments made by Elston and Ellis (Elston and Ellis, 1998; Karayannopoulou *et al*., 2005). The selection of the high-power fields for mitotic counts was performed independently by each observer in the most mitotically active parts of the tumor (Elston and Ellis, 1991; Peña *et al*., 2013). Cases with scoring discrepancies between the veterinary pathologists were reviewed using a multihead microscope, in order to obtain a consensus. The consensus grade and its components were also compared with the grade assigned by the medical pathologist.

The interobserver variability was measured by estimating the percentage of equal assessments. The κ_u and κ_w statistics were used to assess the paired interobserver agreement for histological grading and for parameter scoring. A value of $\kappa > 0.8$ is considered to indicate almost perfect agreement, while $0.6 < \kappa \leq 0.8$ and $0.4 < \kappa \leq 0.6$ values indicate good and moderate agreements, respectively. In contrast, $\kappa < 0.4$ is considered a poor agreement (Vieira and Garret, 2005). The interobserver variability in total score assigned (values 3 to 9) was also estimated as a correlation coefficient (Spearman's rank correlation coefficient). For these tests $P < 0.05$ was considered significant. Logistic regression was used to assess the influence of clinicopathological

parameters on the odds of grading disagreement. For this analysis, P <0.1 was considered significant. All analyses were performed using R free software (R Core Team) using packages psych (Revelle, 2014) and Hmisc (Harrell, 2014).

In this series of 46 simple CMCs, mean (standard deviation) tumour size was 3.3 (3.1) cm. At the time of diagnosis, 33% (15/46) of the tumours showed vascular invasion and/or lymph node metastases. During the follow-up period, 27% (8/30) of dogs developed progression-related events (i.e. recurrences or distant metastases). Grade I tumours were relatively uncommon, representing 11–20% of cases depending on the observer (Table 1).

Overall, there was an agreement percentage for tumour grading of 52%. For tubule formation, nuclear pleomorphism and mitotic counts the agreement was 61%, 50% and 54%, respectively. The agreement of the sum of scores was 24%. The interobserver variability, measured as the percentage of concordance and κ values in paired comparisons, is illustrated in Table 2. The tumor grade κ_w varied from 0.59 to 0.80 (mean κ_w of all pairwise comparisons was 0.70). For tubule formation, nuclear pleomorphism and mitotic counts, the mean κ_w of all pairwise comparisons was 0.71, 0.51, and 0.69, respectively. Higher agreement values were obtained for some of the paired comparisons: (1) consensus and observer 3 (medical pathologist) for tubule formation, nuclear pleomorphism and mitotic count; and (2) observer 2 versus observer 3 for overall tumour grade (Table 2). In general, the highest agreement between observers was seen for evaluation of tubule formation, closely followed by the mitotic count (Table 2). The agreement for nuclear pleomorphism in all pairwise comparisons was moderate. In all instances, the Spearman's correlation coefficient for the overall score was higher than 0.70 (P <0.001).

Cases with complete agreement between the three observers for tumour grade are illustrated in Figs.1 and 2. When disagreement existed, the pathologists always clustered their opinions around two adjacent grades and the difference in score of each parameter and the sum of scores were ±1, in the majority of cases. As the disagreement usually corresponded to adjacent scores, the κ_w was invariably higher that κ_u (Sim and Wright, 2005).

The odds of disagreement when scoring parameters increased with tumour size: each centimetre increase in diameter accounted for 1.4 times higher odds of disagreement (P = 0.065). In contrast, the odds of disagreement decreased by a factor of 0.03 when vascular invasion and/or regional lymph node metastases were detected at diagnosis (P = 0.08). The level of disagreement was similar in tumours with and without progression during the follow-up period.

In the last decade, the NHG has been adapted to CMC grading; however, its use requires adjustment for veterinary pathology (Matos *et al*. 2012; Mills *et al*., 2015). In this first study of grade reproducibility, we focused on simple CMCs since they are considered most similar to the common forms of human breast carcinoma. This subtype of tumour is suitable for comparing grading assessment by veterinary and medical pathologists, which was one goal of this study. Moreover, as simple carcinomas are associated with a poorer prognosis when compared with complex and mixed carcinomas (Misdorp *et al*., 1999), it is critical, in prognostic terms, to be aware of interobserver reproducibility in grading this particular tumour subgroup.

The reproducibility observed in this study (κ_w = 0.70 for the overall grade) is in close agreement with the human field (κ_w of 0.30–0.70) (Meyer *et al*., 2005; Rakha *et al*., 2010). Furthermore, the higher reproducibility value in scoring tubule formation (0.71), followed by mitotic count (0.69) and finally nuclear pleomorphism (0.51) is

similar to the majority of human breast cancer studies (reviewed by Meyer *et al.*, 2005; Rakha *et al*., 2010). In grading CMCs, consensus seems to be least common with scoring of nuclear pleomorphism. In the human literature, various reasons have been proposed to justify this, including the qualitative nature of the scoring method and the heterogeneity of the nuclear features within a tumour (Meyer *et al.*, 2005; Longacre *et al.*, 2006; Adams *et al.*, 2009). Moreover, we recently demonstrated that CMCs that scored 1 and 2 have similar mean nuclear volumes (Santos *et al*., 2014). Additionally, the use of the normal surrounding parenchyma as a reference may jeopardize the reproducibility of nuclear pleomorphism in CMCs, since the parenchyma often presents variability in nuclear features, depending on the stage of the oestrous cycle (Santos *et al*., 2010).

The second poorest agreement was seen for mitotic count, probably due to the selection of areas for counting mitotic figures (Meyer *et al*., 2005; Longacre *et al.*, 2006). In large tumours, the high number of slides can be an additional bias, which could explain increased odds of grading disagreement with increasing size. To decrease bias, some studies in human breast pathology have assigned designated counting areas on the slides of each tumour (Tsuda *et al*., 2000). In the present study, there was no attempt to guide observers to any particular slide or tumour area. Even if this led to some of the interobserver variation, it represents more accurately the procedures of pathologists during their routine diagnostic activity (Longacre *et al*., 2006).

In this study, Spearman's correlation coefficient for the total combined score was relatively high, indicating that when an observer attributed a high score to a tumour, it was likely that the other observer would also attribute a high score.

The levels of agreement in grading parameters showed a tendency to increase when consensus between two observers was reached. This suggests that efforts

to obtain a grading consensus are an effective way to compensate for potential individual bias in scoring. In human medicine it has been postulated that two or three pathologists should suffice to reach a valid consensus (Dalton *et al*., 2000).

In conclusion, to our knowledge, this is the first study addressing interobserver agreement in grading CMCs. The grading method presented a level of reproducibility similar to that reported for human breast carcinomas. Future intra- and interdepartmental studies with a panel of observers and different subtypes of CMC are warranted to fully ascertain the reliability of the method.

Acknowledgments

We thank M. Henry for language editing.

Conflict of Interest Statement

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

References

Adams AL, Chhieng DC, Bell WC, Winokur T, Hameed O (2009) Histologic grading of invasive lobular carcinoma: does use of a 2-tiered nuclear grading system improve interobserver variability? *Annals of Diagnostic Pathology*, **13**, 223-225.

Chowdhury N, Pai MR, Lobo FD, Kini H, Varghese R (2007) Impact of an increase in grading categories and double reporting on the reliability of breast cancer grade. *APMIS:acta pathologica, microbiologica, et immunologica Scandinavica*, **115**, 360-366.

Dalton LW, Pinder SE, Elston CE, Ellis IO, Page DL *et al.* (2000) Histologic grading of breast cancer: linkage of patient outcome with level of pathologist agreement. *Modern Pathology*, **13**, 730-735.

Elston CW, Ellis IO (1991) Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, **19**, 403-410.

Elston CW, Ellis IO (1998) Assessment of histological grade. In: *Rosen´s Breast Pathology*, 1st Edit., PP Rosen, Ed., Lippincott-Raven, Philadelphia, pp. 365-384.

Frierson HF Jr, Wolber RA, Berean KW, Franquemont DW, Gaffey MJ *et al*. (1995) Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma. *American Journal of Clinical Pathology*, **103**,195-198.

Goldschmidt M, Peña L, Rasotto R, Zappulli V (2011) Classification and grading of canine mammary tumors. *Veterinary Pathology*, **48**, 117-131.

Harrell F (2014) with contributions from Charles Dupont, many others. Hmisc: Harrell Miscellaneous, Department of Biostatistics, Vanderbilt University School of Medicine, USA, http://biostat.mc.vanderbilt.edu/Hmisc.

Karayannopoulou M, Kaldrymidou E, Constantinidis TC, Dessiris A (2005) Histological grading and prognosis in dogs with mammary carcinomas: application of a human grading method. *Journal of Comparative Pathology*, **133**, 246-252.

Longacre TA, Ennis M, Quenneville LA, Bane AL, Bleiweiss IJ *et al.* (2006) Interobserver agreement and reproducibility in classification of invasive breast carcinoma: an NCI breast cancer family registry study. *Modern Pathology,* **19**, 195-207.

Matos AJ, Baptista CS, Gärtner MF, Rutteman GR (2012) Prognostic studies of canine and feline mammary tumours: the need for standardized procedures. *Veterinary Journal*, **193**, 24-31.

Meyer JS, Alvarez C, Milikowski C, Olson N, Russo I *et al*. (2005) Breast carcinoma malignancy grading by Bloom-Richardson system *vs* proliferation index: reproducibility of grade and advantages of proliferation index. *Modern Pathology,* **18**, 1067-1078.

Mills SW, Musil KM, Davies JL, Hendrick S, Duncan C *et al*. (2015) Prognostic value of histologic grading for feline mammary carcinoma: a retrospective survival analysis. *Veterinary Pathology,* **52**, 238-249.

Misdorp W, Else RW, Hellmén E, Lipscomb TP (1999) Histological classification of mammary tumors of the dog and the cat, 2nd Series. In: *World Health Organization International Histological Classification of Tumours of Domestic Animals*, Vol. VII. Armed Forces Institute of Pathology, Washington DC.

Peña L, De Andrés PJ, Clemente M, Cuesta P, Pérez-Alenza MD (2013) Prognostic value of histological grading in noninflammatory canine mammary carcinomas in a prospective study with two-year follow-up: relationship with clinical and histological characteristics. *Veterinary Pathology,* **50**, 94-105.

Revelle W (2014) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, http://CRAN.R-project.org/package=psych Version = 1.5.1.

Robbins P, Pinder S, de Klerk N, Dawkins H, Harvey J *et al*. (1995) Histological grading of breast carcinomas: a study of interobserver agreement. *Human Pathology*, **26**, 873-879.

Santos AA, Lopes CC, Ribeiro JR, Martins LR, Santos JC *et al*. (2013) Identification of prognostic factors in canine mammary malignant tumours: a multivariable survival study. *BMC Veterinary Research*, **9**, 1.

Santos M, Correia-Gomes C, Santos A, de Matos A, Rocha E *et al.* (2014) Nuclear pleomorphism: role in grading and prognosis of canine mammary carcinomas. *Veterinary Journal*, **200**, 426-433.

Santos M, Marcos R, Faustino AM (2010) Histological study of canine mammary gland during the oestrous cycle. *Reproduction in Domestic Animals*, **45**, e146-e154.

Sim J, Wright CC (2005) The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical Therapy*, **85**, 257-268.

Stenkvist B, Westman-Naeser S, Vegelius J, Holmquist J, Nordin B *et al*. (1979) Analysis of reproducibility of subjective grading systems for breast carcinoma. *Journal of Clinical Pathology*, **32**, 979-985.

Rakha EA, Reis-Filho JS, Baehner F, Dabbs DJ, Decker T *et al.* (2010) Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Research*, **12**, 207.

Tsuda H, Akiyama F, Kurosumi M, Sakamoto G, Yamashiro K *et al*. (2000) Evaluation of the interobserver agreement in the number of mitotic figures of breast carcinoma as simulation of quality monitoring in the Japan National Surgical Adjuvant Study of Breast Cancer (NSAS-BC) protocol. *Japanese Journal of Cancer Research*, **91**, 451-457.

Vieira AJ and Garrett JM (2005) Understanding interobserver agreement: the kappa statistic. *Family Medicine*, **37**, 360-363.

Webster JD, Dennis MM, Dervisis N, Heller J, Bacon NJ *et al*. (2011) Recommended guidelines for the conduct and evaluation of prognostic studies in veterinary oncology. *Veterinary Pathology*, **48**, 7-18.

Figure Legends

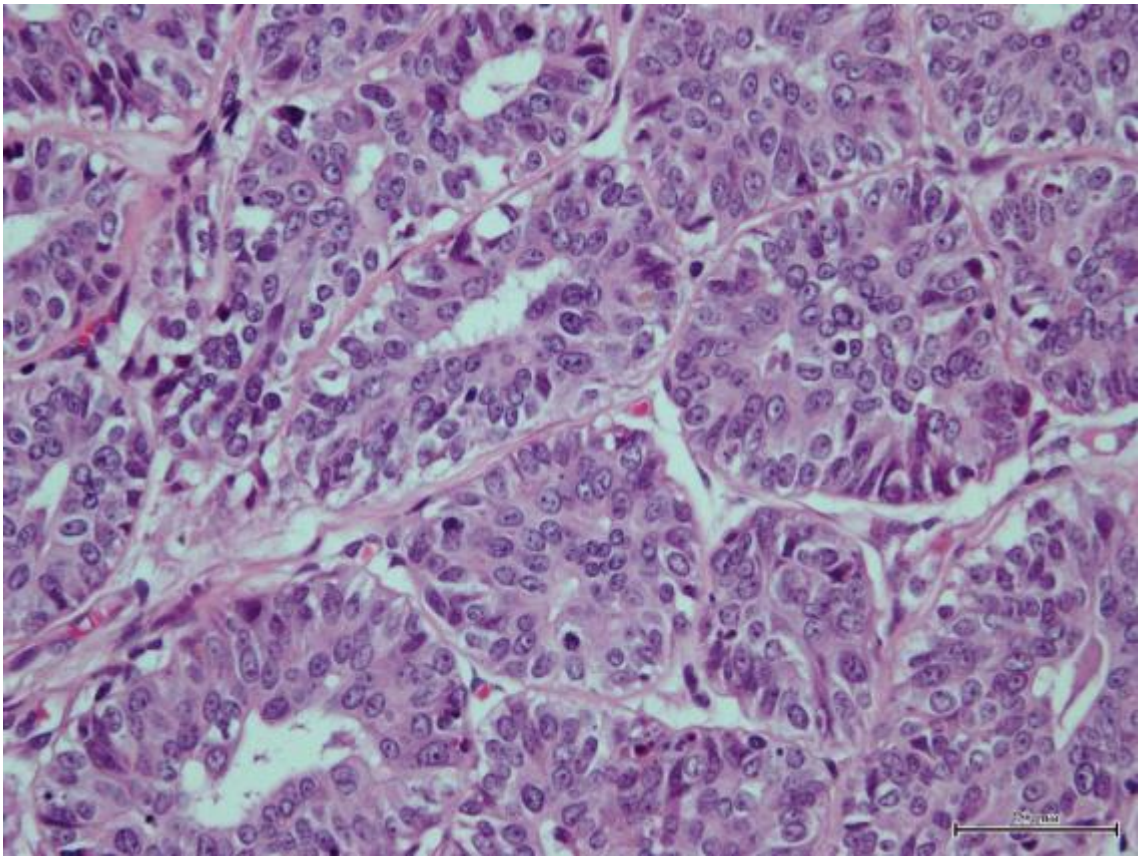

Fig. 1. Grade II canine simple mammary carcinoma. Complete agreement between the three observers in the scores of the grading parameters (score 1 for tubule formation, score 2 for nuclear pleomorphism and score 3 for mitotic count). Haematoxylin and eosin. Bar, 50μm.

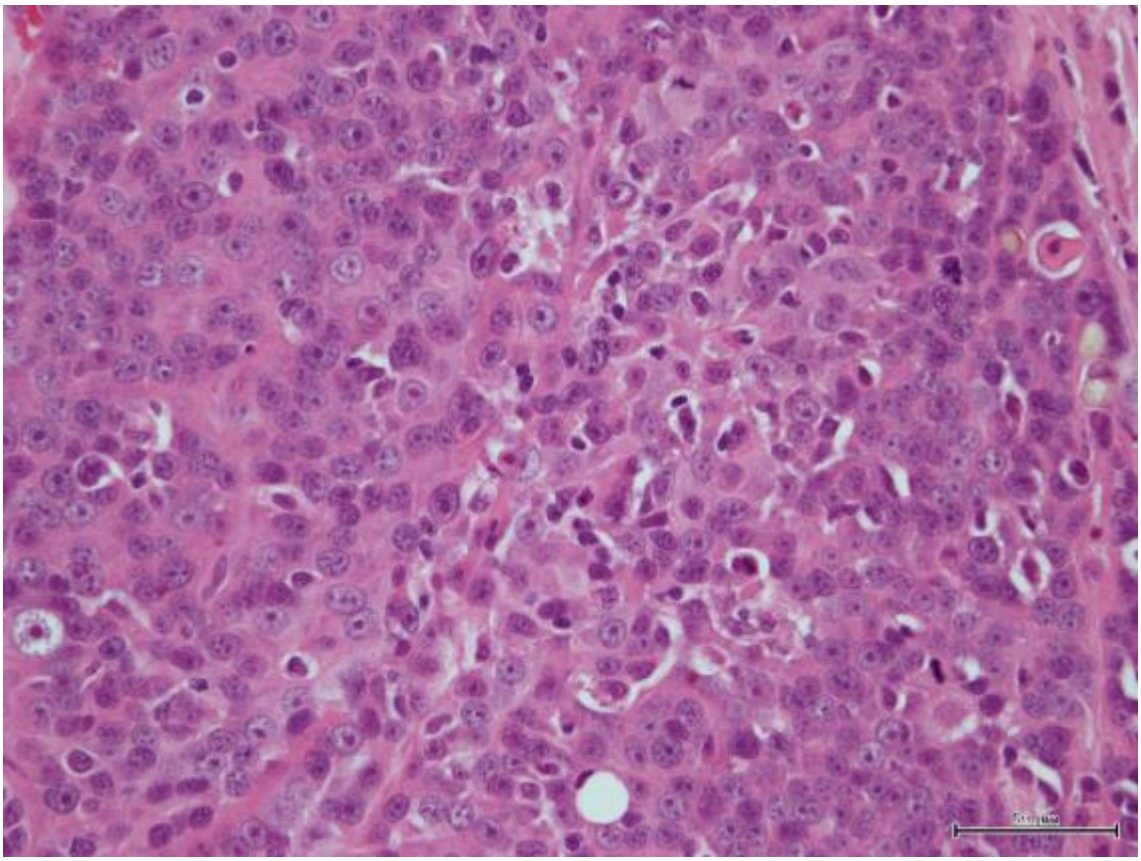

Fig. 2. Grade III canine simple mammary carcinoma. Tubule formation, nuclear pleomorphism and mitotic count were scored as 3 by the three observers. Haematoxylin and eosin. Bar, 50μm.

Table 1

Individual and consensus (observers 1 and 2) grading of 46 canine mammary carcinomas using the Nottingham histological method

	Observer 1	*Observer 2*	*Consensus of observers 1+2*	*Observer 3*
Grade 1	5	9	6	8
Grade 2	20	19	21	17
Grade 3	21	18	19	21

Table 2

Percentage of concordance (C) and kappa agreement values between observers in grading parameter scores and in the Nottingham histological grade

	Tubule formation	*Nuclear pleomorphism*	*Mitotic count*	*Grade*
Observer 1/Observer 2				
C	76%	65%	69%	67%
κ_u	0.61 (0.42–0.81)	0.40 (0.17–0.63)	0.49 (0.29–0.68)	0.47 (0.26–0.69)
κ_w	0.71 (0.52–0.89)	0.57 (0.41–0.74)	0.68 (0.50–0.86)	0.68 (0.53–0.83)
Observer 1/Observer 3				
C	70%	65%	61%	59%
κ_u	0.50 (0.29–0.72)	0.38 (0.15–0.60)	0.32 (0.12–0.51)	0.33 (0.10–0.55)
κ_w	0.69 (0.54–0.85)	0.43 (0.21–0.66)	0.55 (0.34–0.76)	0.59 (0.42–0.76)
Observer 2/Observer 3				
C	72%	65%	74%	78%
κ_u	0.55 (0.34–0.75)	0.39 (0.15–0.62)	0.58 (0.39–0.77)	0.66 (0.46–0.85)
κ_w	0.66 (0.47–0.86)	0.46 (0.21–0.72)	0.77 (0.63–0.92)	0.80 (0.68–0.93)
Consensus/Observer 3				
C	76%	70%	70%	70%
κ_u	0.61 (0.41–0.81)	0.45 (0.22–0.68)	0.50 (0.31–0.69)	0.52 (0.30–0.73)
κ_w	0.76 (0.62–0.90)	0.58 (0.39–0.77)	0.77 (0.67–0.88)	0.71 (0.56–0.86)

κu, kappa unweighted; *κw*, kappa weighted; brackets show 95% confidence intervals