Development, Reporting and Validation of Histologic Tumor Grading Systems
Version: GS 1.0
Guideline Date: Feb. 2023

Communication authors: Christof A. Bertram (christof.bertram@vetmeduni.ac.at), Don Meuten (don_meuten@ncsu.edu), Frances Moore (fmoore1977@centurytel.net), F. Yvonne Schulman (schulmanfy@comcast.net)

Anatomic pathologists: Giancarlo Avallone, Christof A. Bertram, Pompei Bolfa, James Chambers, Michael J. Dark, Melanie Dobromylskyj, Robert A. Foster, Taryn A. Donovan, Andrew Miller, Donald J. Meuten, Frances Moore, Paola Roccabianca, F. Yvonne Schulman, Geoffrey Wood

Clinical pathologists: Melinda Camus, Paula Krimer, Donald J. Meuten, Frances Moore

Medical or surgical oncologists: Nick Dervisis, Milan Milovancev, Laura E. Selmic

Epidemiologists / statisticians: Alexander Bartel, Paula Krimer

All authors are listed alphabetically; contact the communication authors to suggest updates, provide edits and comments.

Recommended Citation: Bertram CA *et al*. Development, Reporting and Validation of Tumor Grading Systems, version 1.0. Veterinary Cancer Guidelines and Protocols. http://vetcancerprotocols.org, accessed on (date).

Contents

Introduction

This guideline provides a framework for developing, reporting, and validating histologic tumor grading systems for veterinary oncology. There are various definitions for tumor grading, but it is essentially *a system that provides probabilities for recurrence, metastases and various survival metrics based on specific microscopic features*.[2] The purpose of a grading system is to subdivide a specific neoplasm into categories that correlate with clinical outcomes. Grading systems should be targeted for one tumor type in one species. They should be simple to apply, reproducible, unambiguous and accurately segregate tumors with different behaviors to ensure the tumor grade provides information that is useful in assigning prognosis and developing treatment plans.

Key criteria of any grading system are:

- Reproducibility within and across observers and laboratories;
- Prognostic relevance;
- Applicability in a routine diagnostic setting.

Grading systems in veterinary oncology rely on defined histologic or cytologic features.[2,7] This guideline addresses histologic grading; cytologic grading guidelines will be addressed in a separate document. Tests currently considered ancillary (histochemical stains other than hematoxylin/eosin, immunohistochemical procedures or molecular tests), are not recommended for histologic tumor grading; however, they may prove critical to future prognostic or predictive algorithms or supplement existing systems.

How useful are histologic features? Given the complexities and nuances of cancer biology, they are surprisingly helpful. However, we know from human oncology that tumors with the same histologic features can have different molecular profiles, and the genetic makeup and immunocompetency of each patient is different.[20] Precision oncology attempts to link these, as well as other components, to provide the most accurate prognoses and direct patient-specific therapies. Given the unique features of the tumor and the host, histologic features will not be 100% prognostic of tumor behavior. Imperfect as it may be, the tumor grade can be a critical factor, when considered in association with other prognostic factors, in directing patient treatment. In human oncology, it is common to develop overall-risk models (nomograms) that include grade, molecular tests, IHC, stage, and patient characteristics, to provide the most accurate prognosis and best therapeutic plan.

The goal should be to develop grading systems that use parameters that can be determined during routine histopathologic evaluation. If ancillary tests are evaluated, the prognostic value should be compared to the results based on standardized histological parameters. Comparisons are needed to determine if newer methods, or combinations of methods, are more predictive of an outcome or treatment selection, and if so at what cost

(spectrum of care).[15] Investigators, clinicians and diagnosticians should determine reasonable goals of a grading system. These goals should be set in relation to how well other, currently available parameters and tests are associated with clinical outcomes, and may be tumor specific.

While a *prognostic* parameter gives information about the patient's potential cancer-induced outcomes regardless of treatment, *predictive* biomarkers provide information about the tumor's potential response to specific therapeutics.[1,18] Some parameters are both prognostic and predictive. Histologic grading systems or individual parameters which provide survival metrics and/or probabilities of metastasis or recurrence have prognostic value. Investigators should also try to determine if a tumor subtype, grade or parameter helps select a treatment option that may help pets live a longer, better quality life.

Grading systems have been published in the veterinary literature for several tumor types in dogs and cats (see supplemental tables). As reviewed recently,[2] most grading systems are based on studies that have drawbacks, which limit the clinical utility, highlighting the need for guidelines. These drawbacks include:

- Retrospective study design;
- Non-representative study populations;
- Small case numbers;
- Poorly defined or absent inclusion criteria;
- Use of subjective rather than objective criteria;
- Inclusion of heterogeneous groups of neoplasms;
- Record-review studies with reliance on pathology reports instead of assessment by independent review of tumor histology with current methods and definitions;
- Failure to include at least one pathologist (ideally two or more) as author(s) to objectively and independently review the histopathology;
- Systems in which grades cannot be accurately assigned due to discordant or overlapping scoring parameters within or between grades;
- Transfer of grading systems from human medicine or other tumor types without validating the value or applicability of the grading system for the tumor and species being studied;
- Lack of standardization of tumor parameter assessments;
- Lack of standardization of outcome assessments;
- Lack of validation studies to determine the reproducibility and prognostic value of each parameter and the grading system in different patient populations;
- Accepting previously published cut-off points without validation;
- Inappropriate selection and/or combination of individual prognostic parameters for developing the grading system;
- Failure to apply uniform selection criteria, including the same standard treatment, and variability across protocols, including surgical techniques and dose (width of surgical margins), and margin assessment;

- Incomplete clinical outcome and treatment data;
- Inadequate recognition of how euthanasia influences outcome data;
- Lack of histologically/cytologically confirmed recurrences and metastases;
- Lack of postmortem examinations to determine the extent of tumor involvement, at least in a representative sample of cases;
- Inappropriate statistical analysis, e.g., “p-value approach” is not adequate for evaluating usefulness of a marker, and statistical significance does not equal prognostic relevance.[6]

The development of new grading systems requires participation of primary care veterinarians, oncologists, radiologists, pathologists, surgeons, and statisticians. Inclusion of statisticians or epidemiologists is essential. All participants, especially statisticians, should contribute to the study design to define an adequate number of cases, sufficient study power, and to help prevent study bias before beginning the study. For grading systems to be clinically useful, it is imperative to apply appropriate statistical assessment to study results, which are based on standardized tumor parameters correlated with standardized outcomes. Clinical outcomes, including histologically confirmed recurrence and metastasis, disease free intervals (DFI) and survival metrics (separated as to euthanasia induced, tumor related, and/or non-tumor related) should be collected on large numbers of cases in which the patients have the same histologic tumor type. Studies should assess as many histologic parameters as possible to identify the criteria that, individually or in combination, predict specific outcomes with the most accuracy. Criteria to determine tumor grades must be sufficiently detailed to ensure others can reproducibly categorize tumors into the grades indicated in the published system. Criteria which are poorly reproducible among pathologists should not be included. Ideally, the criteria for grading should be easy to recognize and assess, unambiguous, and should stratify tumors into distinct grades. Those that come close to this goal will be widely accepted if they predict one or more outcomes with “high probability” and/or help direct treatment options.

The following are recommendations to develop, report, and validate tumor grading systems, prepared under the umbrella of the Veterinary Cancer Guidelines and Protocols initiative (VCGP, https://vcgp.org/). Colleagues are encouraged to send the communication authors of this guideline edits, comments and suggestions that will be used to update the guideline and make it as broadly applicable as possible.

Recommendations for development of grading systems

Development of grading systems is recommended for all tumor types with variable biological behaviors. The key criteria of grading systems (see above) should be achieved by following the recommendations below.

- Identify a specific tumor type in one species.
- Ensure the investigating team includes pathologists, clinicians/oncologists and statisticians with expertise in appropriate study design, inclusion/exclusion criteria, and analyses for oncologic studies.
- Collect a large enough study population with appropriate clinical outcome information for sufficient statistical power.
    - Report the power analysis (calculation used to estimate the smallest sample size needed for the study) that was performed before conducting the study.
    - Detail the demographic characteristics of the patient population
    - Detail the tumor characteristics, such as tumor location, and margins (especially when recurrence is an assessed outcome).
    - Standardize and define the therapeutic intervention(s). See VCGP guideline on outcome assessment (https://www.vcgp.org/) and RECIST [12] for specific recommendations.
    - Define outcomes including “patient-centered” (disease-free survival, progression-free survival, other survival metrics) and “tumor-centered” (recurrence, metastasis).
    - Use categorical outcome measurements (occurrence of events, such as survival rate and DFI, at end of a follow up period) and contiguous outcome measurements (time to an event, such as survival time or DFI).
    - Include as many patients with accompanying autopsy findings as possible, as they verify the extent of tumor progression and cause of death (related or not to the original neoplasm).
- Assess as many histologic tumor parameters as possible (see table 1).
    - Provide methods of how each parameter is evaluated that are detailed enough to allow others to repeat the study with other cases.
    - When available, standardized methods for assessing the parameters (such as provided in the VCGP guidelines and protocols: https://www.vcgp.org/) can be used; however, investigators should consider different methods of parameter assessment and whether different methods are better for different tumor types.
    - Determine reproducibility of each parameter. The methods for determining reproducibility need to be specified.
    - Parameters may be numerical (continuous scale) or categorical (a variable that can take on one of a limited, and usually fixed, number of possible values). For all categorical parameters describe the critical features in detail, with pictorial illustrations for each category.

    - Determination of how each parameter is assessed and evaluation of reproducibility should be done before the study cases are evaluated.
- Reject parameters that are not reproducible. Criteria for determining sufficient/insufficient reproducibility should be defined.
- Determine the discriminant ability of each of the remaining parameters using appropriate statistical methods with consideration of different statistical methods for categorical vs continuous variables, such as the hazard ratio or AUC (Area under the ROC curve).[6] with inclusion of confidence intervals. Non-overlapping confidence intervals indicate a statistically significant difference between the two parameters or grades.
- Reject parameters that are not among the most prognostically relevant based on the statistical results. Criteria for determining the most prognostically relevant parameters of the tumor under investigation must be specifically given.
- Formulate a grading system or algorithm using the combination of tumor parameters that has the strongest statistical correlation with one or more outcomes. Criteria for determining the strongest statistical correlation must be specifically given.
    - The number of grades should be determined by the number of tumor subgroups with different biological behavior that can be identified with reproducibility.
    - The minimum number of parameters needed to retain prognostic value should be included in the final grading system for ease of application.
        - Each tumor parameter included in the grading system should have an “added value” for correlation with outcome (improve the utility of the system; see discussion).
        - For parameters that measure similar features and/or correlate strongly (i.e., parameters with a collinearity/causal relationship), such as karyomegaly and nuclear pleomorphism, both should be reported and compared, and the better representative parameter should be chosen for the final grading system. Researchers are encouraged to evaluate more than one parameter that evaluates similar morphological features, compare their usefulness, tabulate results and justify the choice(s) recommended.
        - The value of each parameter should be demonstrated by using the AUC or other suitable analysis and comparing the reduced model to the full model.[6]
    - The cut-off values (thresholds) of numerical parameters should be based on their statistical discriminability and not arbitrarily selected.
        - A clinically meaningful sensitivity vs. specificity of the cut-off should be selected. Appropriate multivariable statistical analysis, like decision trees or cross-validated parameter tuning, can help to inform this decision.[6]

        - The confidence range associated with each cut-off value should be considered in the weight given to each parameter when calculating the grade; for example, it is unlikely that a tumor with a mitotic count of 2 is really associated with a better prognosis than the same tumor with a mitotic count of 3. A simple way to evaluate this is to compare the confidence ranges and how much they overlap.
    - Categorical parameters can be assigned numerical scores and/or assigned different weights based on confidence ranges.
    - Specific ranges of the parameter score sums should correlate with different grades and outcomes. Grading systems should provide non-overlapping score ranges that cover all possible scores, such as 0-5 = grade 1, >5-10 = grade 2, and >10 = grade 3.
    - Each possible set of grading parameters should be assigned a specific grade. Predetermined classification systems in which a single tumor can have histologic features that belong to 2 different grades are not useful. At the same time, no combination of histologic features should be without an assigned grade.
    - Higher grades should have a worse prognosis.
    - Emphasize outcome as an expected result, such as days of survival, probability of metastasis at a specific time point (e.g., in 6 months, or 1 year, depending on the disease), disease-free interval, etc., and not as good or bad, benign or aggressive.
- Demonstrate the prognostic value of the grading system using appropriate statistical analysis, such as Kaplan-Meier curves, hazard ratios, sensitivity, specificity, number of false and correct classifications, and area under the ROC curve (AUC). All results should have confidence ranges.
- Compare the prognostic value of the grading system (based solely on histological parameters) with ancillary prognostic indicators (such as tumor stage, immunohistochemistry, cytologic features, mutational status, etc.; see table 2), and any previous grading systems.
- Evaluate the reproducibility of the grading system. Investigators developing grading systems should assess interobserver variation. How much is acceptable may depend on the effect of the variation on the tumor grade in their system. If this is not done in the primary study, at least one subsequent validation study is needed.

Table 1. Histologic characteristics of the primary tumor for consideration in grading system development.

Parameter		Comments
Anatomic location	Organ	Specific location within the organ and layers affected*
Tumor size		mm or cm**
Extent of invasion		Define or measure depth and width relative to anatomical landmarks such as epidermal surface for skin tumors, organ capsules, or adjacent tissues
Proliferation	Mitotic count	For VCGP guideline see https://www.vcgp.org/
Lymphovascular invasion		Soft vs hard criteria; for VCGP guideline see https://www.vcgp.org/
Tissue differentiation		Define; consider numerical scale [9]
Nuclear morphology	Karyomegaly	Define; develop specific criteria to define each nuclear feature
	Atypia	
	Bi- or/and multinucleation	
	Pleomorphism	
Cytoplasmic morphology	Basophilia	Define; develop numerical scales for each cytoplasmic feature
	Granularity	
	Anisocytosis	
	Other	
Extracellular matrix	Type and amount	
Inflammation	Type and amount	
Necrosis		Give specifics of assessment; for VCGP guideline see https://www.vcgp.org/
Unique parameter(s)	Pigment	
	Other	

* e.g., skin - location on the body and dermis/subcutis; mucosal organs - mucosa/submucosa/muscularis; bone - axial vs appendicular and epiphysis, metaphysis, or diaphysis; specific area(s) of the brain; etc.

** consider in relation to body weight/surface area of patient

Table 2. Additional information/parameters to be considered with tumor grade in outcome assessment.

Parameter		Comments
Clinical information	Signalment	Species, breed, gender, age, body weight
	Comorbidities	
	Tumor stage	Follow staging protocol for specific tumor type
	Other	
Gross evaluation	Tumor size	mm or cm*
	Necrosis	Develop assessment criteria; express as % of tumor area; for VCGP guideline see https://www.vcgp.org/
	Exact anatomic structure(s) affected; Extent of invasion	Define categories for extent of invasion or measure depth and width
Proliferation	Ki-67 index	Specify how assessed; develop standardized method
	AgNOR	
	PCNA	
	Other	
Margin assessment	Histologic tumor-free distance (HTFD)	For VCGP guideline see https://www.vcgp.org/; consider alternative assessments and trimming methods
	R0-R3	
	Other	
Lymph node metastasis	Confirmed	Specify when confirmed by cytology or histology; specify regional or other lymph node(s)
	Suspected	
Distant metastasis	Confirmed	Specify when confirmed by cytology or histology; specify location; for VCGP guideline see https://www.vcgp.org/
	Suspected	
Lymphovascular invasion		Define method used for evaluation, i.e, type of imaging, etc.; for VCGP guideline see https://www.vcgp.org/
Cytology		
Immunohistochemistry		
Unique parameter(s)	Molecular profile of the tumor	Detailed report of the specific analysis, including the bioinformatics, and deposition of the raw data in public repositories.
	Genetic analysis of the host	

* consider stereology and size relative to body weight/surface

Recommendations for publishing grading systems

It is critical that all proposed grading systems be published in a peer reviewed journal and include details regarding study design, methods, data and outcomes. The published information must be sufficient to permit other investigators to replicate and validate the findings in new investigations. Validation studies are needed. Entire data sets should be made available and can be accommodated in supplemental files. The publication should include comparison between the proposed grading system and existing grading systems based on common clinical outcomes, such as disease-free interval, survival or mortality, metastasis and recurrence. All relevant statistics, such as Kaplan-Meier curves, median survival time, hazard ratios (with 95% confidence intervals), sensitivity, specificity, true positives, true negatives, false positives, false negatives should be reported.

Although there are several sources for oncology study guidelines, such as REMARK,[24] there are insufficient details regarding the characterization of histopathologic parameters necessary for the methods to be replicated by other investigators or diagnostic pathologists. Several of the criteria for determining histologic parameters in many currently used grading systems in veterinary medicine have not been standardized. REMARK [24] outlines what should be included in a study but leaves it to investigators to select methods and study design and provide detailed descriptions. *Veterinary Pathology* provides reporting guidelines [26] for manuscripts on tumor prognosis, advising that details related to assessment of histological features and tumor grading be included in submitted manuscripts. The Veterinary Cancer Guidelines and Protocols (VCGP; www.vcgp.org) website is a resource for standardized methods to evaluate tumor parameters. Table 1 in this document enumerates potential histologic parameters for consideration.

Creating a table (see appendices 1 and 2) or graphic illustration of the grading system, and including recommendations on how to report the grade (including potential synoptic reports) and the grade's interpretation in biopsy reports are recommended.

Eponyms should be avoided for new grading systems. Grading systems can be referred to by year of publication, e.g., melanocytic tumor grading system 2022. If there are 2 in a year, 2022a and 2022b can be used, and so on.

Recommendations for validating grading systems

Studies that develop grading systems are generally observational.[8] Thus study populations may be biased in terms of case selection, including factors such as primary vs secondary vs tertiary care centers, initial therapies, "rescue" therapies, outcome data, retrospective design, influence of euthanasia, etc.. There may also be assessment variation between pathologists. Therefore, the reported prognostic value of the resultant grading system may not be applicable in a different population of pets or for different pathologists, and validation studies are needed before implementation of the system on clinical cases. For many published grading systems, there are few or no validation studies to date.[2] The grading system can be validated by different pathologists in the same

demographic population (replication study) and/or in other populations. Validation studies are also necessary to statistically determine reproducibility of each tumor parameter and of the grading system by independent investigators.

The grading system should be validated for the same tumor type and species for which it was developed using the following recommendations:

- Investigators not involved in the proposed grading system should trial the system by precisely following the materials and methods reported by the initial investigators to determine if similar parameter scores and grades can be assigned and associated with similar patient outcomes.[16,21,25,27] Two different approaches can be used for grading system validation:
    1. Initial investigators provide the complete data sets (all materials, such as blocks, slides, whole slide images [WSI], gross descriptions, photos, clinical data and statistics) to an independent group of investigators. The second group of investigators repeats the study using the same materials and the published methods, in order to determine the reproducibility of the grading system.
    2. A second group of investigators applies the published grading system to a new group of cases of the same tumor type in the same species to determine if a similar association with outcomes is achieved.

  The second method is a more robust validation, as it tests the grading system in a different group of patients assessing its general applicability.

- The same outcome metrics and statistical analysis used in the primary study should be used in the validation study *so long as those are correct for the study*, and the correlation of the grades and outcomes compared. New parameters or ways of assessing parameters can be introduced to modify the original grading system and if done, must be compared to those recommended in the initial publication and the rationale given for the modification. If comparisons are not analyzed and published, it is impossible to know which is more prognostic and/or predictive of response to a treatment. Determination of which methods or grading systems are “better” should be based on statistical comparison of relationships to the same set of clinical outcomes.
- If the independent investigators identify parameters that have significant observer discordance or the grading system does not correlate substantially with clinical outcomes, recommendations to improve these objectives should be considered and either assessed or proposed for future studies.
    - One or more of the following modifications may be proposed based on statistical results:
        - Additions, exchanges, or exclusion of grading parameters;
        - Changes to the cut-off values or stratifications of individual grading parameter and the overall grades;
        - Changes to the methodology of assessing a grading parameter;

        - Changes to weights of the individual parameters within the grading systems;
    - The modified methods must be described in sufficient detail for reproducibility, and the prognostic value of the modified grading system must be compared to the original grading system. The most prognostically useful system should be adopted; multiple grading systems for the same tumor should be avoided.
- When possible, compare the prognostic value of the grading system with other established prognostic tests (such as tumor stage, immunohistochemistry, mutation status, etc.).
- Publish the validation study with the same attention to detail as the primary study (See *Recommendations for publishing grading systems* above).
- If the report of the primary study lacks details for assessing the parameters and determining the grade, the validation study should specify the methods applied.[25]
- The independent investigation group should consider sharing their datasets and results with the initial investigators before the validation study is published (see discussion).

Discussion

Assigning grades to some tumors, such as mast cell tumors, soft tissue tumors and mammary carcinomas, is now standard practice for veterinary pathologists.[2,5,9,11,16,19] The grade is an important piece of data used by clinicians to estimate prognoses and/or develop treatment plans.[2,5,9,11,16,19] Grading systems are tumor and species specific, and they should not be extrapolated to a different tumor type or species. The organ involved should also be considered, as organ specific grading systems may be required for some tumors.

The purpose of a grading system is to stratify tumors of a specific histologic type into those with different biological behaviors based on the microscopic features of the tumor. This stratification can only be accomplished by correlating different tumor parameters (and eventually the proposed grades) with accurate, standardized clinical outcome data on a large number of representative cases using appropriate statistical methods.[10] The applicability of veterinary oncology studies is often limited by assessment of insufficient case numbers. Collaboration among investigators and institutions can be critical to gathering the number of cases necessary to achieve reliable results. Cases collected for oncology studies, that have accurate outcome and/or treatment data, should be saved and combined to increase case numbers and for future studies as new methodologies/techniques are developed. Biobanks at different institutions and a VCGP biobank are potential archives for collecting this type of study material in order to help merge ongoing studies and provide cases for further study.

Multiple parameters should be considered in developing a grading system. To presuppose which parameters may be useful, or not, for a tumor type may bias the data. Trojani et al. [28] looked at 7 parameters and determined the 3 most prognostic ones, based

on outcome data, to use for soft tissue tumor grading in human medicine. In veterinary medicine, the 3 criteria used for assessing STTs in human medicine were adopted without evaluating the larger number of parameters and determining if those three are the most prognostically important in dogs.[9,14] Subsequent studies of canine perivascular wall tumors have shown that these parameters are not as reliable as size of tumor and depth of invasion in predicting recurrence.[2] Table 1 provides a list of parameters that can be evaluated via light microscopy on routinely stained sections of the primary tumor but is not all inclusive, and other parameters may be considered depending on the tumor type. Studies that test potential prognostic criteria but find them to have no significant value or be too subjective provide valuable information and should also be published, either as a primary manuscript or as supplemental materials. This information will complete the list of parameters that were evaluated and help colleagues eliminate these criteria from tumor grading once the results are validated. To only publish positive correlations can produce unintended pressure that may lead to study bias.[3,31]

Tumor grade should not be applied in isolation when attempting to determine outcome(s). Patient factors are critical and should be gathered by researchers when developing new grading systems, e.g., species, breed, size, gender, age, evidence of tumor spread, concurrent disease, treatments, etc., which are not part of the histologic grade but are factors that could impact outcome. Breed and/or specific lineage within a breed may affect outcomes due to genes that suppress or stimulate cancer development, independent of the tumor grade; a low-grade tumor in a golden retriever, Bernese mountain dog or boxer dog may behave aggressively in a larger percentage of cases than in other breeds. The impact of tumor size on prognosis is likely related to the size of the patient, e.g., a 2 cm^3 mass in a 5 kg dog versus 2 cm^3 mass in a 50 kg dog. While metastasis is generally considered a poor prognostic indicator, the issue of whether histologically low-grade tumors that have metastasized to a lymph node or beyond, especially in the case of micrometastases (ie., no lymphadenomegaly or grossly detectable metastatic tumor), have a worse prognosis or not needs to be investigated for the different tumor types. Concurrent disease conditions, different treatments and other clinical factors, such as duration of lesion prior to diagnosis or disease-free survival time,[4,13]can all influence outcome. The histologic information (diagnosis, tumor grade) is a piece of the puzzle. Other information gathered by clinicians, oncologists, and radiologists are critical to determining the most accurate prognosis and providing the best treatment recommendations. Alternatives to tumor grading, that have prognostic relevance, have been published for canine oral melanoma, and this approach may be applicable to other tumors.[4,13]

Technology will help create new methods to evaluate tumors and/or patients, and some owners will pay for new techniques at any cost, while others will decline based on philosophy and considerations, such as cost, age of pet, or emotional value of the pet to their family. How to balance optimal care with concerns of individual owners is not simple but is partially encompassed by the philosophy of *spectrum of care*. Researchers can help address this by comparing grading systems with other clinical information and prognostic tests (Table 2), the latter of which are not used routinely in diagnostic pathology or not widely available, such as immunohistochemistry, molecular testing, and genetic sequencing. Armed with this information, the pathologist/oncologist can provide

insights to clinicians and owners regarding the benefits and costs of pursuing additional testing to further characterize the tumor.

Pathologists must use standardized methods that are reproducible. Our present grading systems need to be repeated with standardized parameters and standardized outcomes, and evaluated by multiple pathologists. It is impossible to replicate studies in which the published methods are not described in sufficient detail or involve non-standardized techniques or units of measure.[15,29] Using non-standardized methods that may cause variations in results is scientifically flawed and can lead to misleading conclusions that should not be applied to clinical cases. It will be difficult to perform prospective studies with large case numbers in reasonable time frames unless we develop multi-institutional collaborations. However, prospective studies are the standard we should strive for to improve the care of pets with cancer.[15,29]

Current methods of assessing specific histologic parameters used in grading systems may not be optimal. Different methods, such as assessing different-sized areas (mm^2) to evaluate the mitotic count (MC), only enumerating metaphase or atypical mitoses in the MC, comparing MC at the invasive front vs in random fields vs in the region of highest mitotic activity, or evaluating karyomegaly only at a specific magnification, may improve prognostic discriminability and/or reproducibility of assessed parameters. The methodology for certain parameters may need to be different for different tumor types, e.g., different-sized areas for MC may optimize prognostic ability. Investigators should be creative and consider alternative means to evaluate tumor characteristics and then compare the new or modified methodologies to current methods to determine if one is more reproducible and/or more prognostic of clinical outcomes. For subjective morphologic features, such as bizarre nuclei, it might not be possible to define the entire range of atypia or to identify each variation. Instead, a normal reference could be defined and deviation from that reference point would be assigned a numerical score which becomes the estimate of atypical nuclear morphology.

Regardless of the list of parameters chosen to investigate, each parameter must be clearly defined and evaluated by reproducible methods for intraobserver and interobserver concordance. While outcome data is critical in developing a grading system, the methodology for assessing the different histopathologic criteria can be developed and tested for reproducibility without outcome data.[17,20,23,30] Parameters with significant interobserver discordance should be discarded from consideration. Studies to determine current and acceptable levels of intra- and interobserver concordance for grading parameters and grades are needed. Modifications to parameter assessments and grading systems should be based on improvements in concordance values and outcome predictions. Because grading systems must be based on well-defined histological criteria with high interobserver concordance, it is critical to have more than one pathologist on the study. Relying solely on data from pathology reports, without review by the study pathologist(s), is insufficient and must be discontinued. All parameters are not evaluated or reported in the same way in all diagnostic pathology reports. Experience of the pathologist providing the initial diagnosis, criteria for diagnosis, and methods of parameter evaluation change over time, and retrospective studies relying on reports that are decades old may not account for this inevitability.

Parameters which are enumerated require a standard unit of measure (e.g., mitotic count in 2.37 mm^2) or a standardized number of cells over which the parameters were quantified (e.g., labeling index in X number of cells). These methods need to be detailed in the materials and methods sections such that others can replicate the methods and should be a requirement for publication. Manuscripts that only report assessed areas in terms of low power fields or high power fields should be rejected.

Trimming protocols for tumor assessment and grade assignment are sorely needed. While one tissue block for every 2 cm of tumor has been suggested for assessing canine tumors,[22] studies providing data on the number of blocks/slides needed for adequate evaluation of each parameter within the primary tumor in veterinary medicine have not been published. The amount of tissue necessary for adequate histologic assessment of different tumor types may vary.

Histologic assessment of different tumor parameters can be affected by the percentage and areas of the tumor examined. As an example, a common bias introduced in measuring the proportion of tumor necrosis histologically is the avoidance of areas of necrosis during tissue trimming. The bias introduced by this commonplace practice of only sampling viable tissue for histological examination undermines the utility of tumor necrosis as a parameter in tumor grading systems by thwarting accurate and repeatable results. Tumor necrosis may be the only histologic parameter that we bias by deliberately avoiding the parameter we are trying to measure! Thus, studies to determine and standardize optimal trimming protocols are encouraged, and the recommendations may end up being tumor specific.

Parameters should be evaluated individually, in various combinations, and with numerical scoring systems to determine which individual or combined parameters are most useful. This will require rigorous statistical analyses that are appropriate for the study design, identify prognostic and/or predictive parameters with statistical and practical relevance, and avoid statistical bias.[6,31] A recent commentary about statistical models that are appropriate and inappropriate for oncology studies should be consulted.[6] This manuscript underscores the necessity that a statistician familiar with oncology be a member of the investigative team from initial design to conclusion. For example, the authors explain why the correct specification of endpoints and the best-suited survival model for the endpoints need to be identified before the study begins to ensure data is collected appropriately. They point out that the commonly used “p-value” is not a sufficient statistical criterion to assess the usefulness of tumor parameters, since statistical significance and statistical discriminability are inherently different concepts. Instead, statistical methods and models, such as the area under a ROC curve and/or hazard ratios with confidence intervals are more appropriate. An interesting section identifies problems that statisticians see when consulted after a study ends.[6]

Each tumor parameter included in the grading system should provide an “added value” when correlated with outcome(s). This helps ensure that the grading system is composed of a relatively small number (for ease of use) of only the most prognostically relevant parameters. When developing the grading system, we suggest starting with the parameter that has the highest probability of predicting an outcome. Then add additional parameters that may increase the likelihood of that outcome and evaluate their utility

statistically. For this purpose, the area under the ROC curve (reduced vs. full mode) is best suited.[6] The prognostic value of the reduced model (baseline) is compared to the full model after adding another parameter. Multivariable hazard ratios do not provide the same statistical information as they divide the weights among the included parameters regardless of their correlation to each other; the weight within the prognostic model is divided among those parameters that correlate (each would have a lower hazard ratio), and it is not apparent if the combination of parameters has an "added value" beyond the use of one of these parameters alone. For example, including karyomegaly, anisokaryosis and nuclear pleomorphism into a grading system might not be useful as they measure similar features of the tumor, and one of these parameters on its own might provide the same prognostic information. Our gradings systems should not create unnecessary work.

When a grading system includes three grades, the middle group should not be used as the default. The parameters that separate grades must have unambiguous features that are easily differentiated histologically. If one parameter can be seen in more than one grade, the parameter should be given a score (weighted) based on well-defined histologic criteria and the sum of the scores used to determine the grade. In general, the intermediate grade should have a prognosis between the ranges of the high and low grades and should not be used as a category for uncertain prognosis due to ambiguous grading criteria.

For a grading system to be of clinical use, the tumor grades must be shown to predict a probability of time (e.g. days) to event, tumor behavior and/or response to therapies in a population.[29] Ideally, each grading system should be validated by other laboratories and pathologists before it is put into clinical practice (see *Recommendations on validating grading systems* above). How well does data determined from referral cases extrapolate to primary care patients? Cases seen at referral institutions may be biased towards more severe disease, animals with other underlying severe health conditions, and/or owners more willing to undertake advanced and prolonged treatments, resulting in skewed data that might not be applicable to primary care practice. Treatment protocols affect outcomes and must be detailed, with comparisons made across different treatment groups that include surgical excision only or no treatment (controls). How well study results extrapolate to another group of clinicians, pathologists and patients is rarely examined in veterinary oncology but, when performed, can add to confidence in the grading system or possibly lead to its rejection.

The necessity for validation of grading systems cannot be overemphasized. Historically, once a grading system is created in veterinary medicine, it is put in use regardless of whether the system has been validated. Consensus statements that recommend which study or grading system be used are not validation. Validation studies should be welcomed by scientific journals. How well a grading system will extrapolate to a different group is unknown until validation studies attempt to replicate the methods on a different set of patients by a different group of pathologists. As new studies that increase the number of cases assessed and provide additional data are published and validated, grading systems should be updated to improve the prognostic discriminability of the grades. Investigators performing validation studies should critically evaluate the criteria of the initial study. Identification of additional parameters or different assessments of the parameters that improve the prognostic or predictive ability of the grading system will

increase acceptance for publication. Citing the original grading system as well as the validation study/studies in the materials and methods of future papers should be the standard.

Further, it is preferable to have one grading system per tumor type. To achieve this goal, all grading system modifications and new grading systems should be compared with the existing systems, outlining the different methods and different outcomes, and determining and validating acceptable inter-pathologist variability, to identify the best system regarding clinical usefulness. Comparisons of the systems require detailed descriptions of the methodologies employed in determining the grades.

Once a validation study is performed, sharing any discrepancies with the primary investigators should be considered to resolve the differences and provide the best final recommendation. These two teams are best equipped to resolve discrepancies and avoid creation of two separate conclusions, which would require the reader, who may be less familiar with the topic, to draw conclusions from the conflicting information.

Validation studies can add confidence to study results; nevertheless, grades or individual parameters are based on statistical probabilities for specific outcomes in a given population of patients with the same tumor type. They do not necessarily predict an outcome for an individual patient (version of theranostics; precision oncology). They predict the probability of how a similar tumor type of a similar grade might behave in a similar population of animals. This principle can be added to pathology reports to help clients realize pathologists are not necessarily predicting how the tumor will behave in an individual patient.

A detailed discussion of outcome data is outside the scope of this guideline, but standardized outcome data collection is critical and is the most important factor determining whether a grading system has clinical value. Outcome data should include clinically relevant outcomes, such as recurrence rate and time to recurrence, metastatic rate and time to metastasis, and tumor-related survival metrics. Euthanasia-induced, tumor-related and/or non-tumor-related causes of death should be separated, when possible, but care must be used to not miscategorize cases that can provide important survival data. Death by euthanasia in a patient with obvious widespread progressive cancer and rapidly deteriorating clinical condition is both euthanasia induced and tumor related. In contrast, a patient with cancer and several comorbidities that is found dead may have died of a tumor-related or a non-tumor-related cause. Investigators should define how the categories were distinguished. When analyzing outcomes, researchers should consider whether the diagnostic test result or grade led to a self-fulling prophecy, e.g,. micrometastases are found in a lymph node and that information contributes to the decision to euthanize. Studies should include specific timelines for follow up, which may be different for different tumor types.

Local recurrence is defined as the presence of the same tumor within the region of the previous surgical site. Surgical dose (extent of the surgery performed) should be considered in determining recurrence rate. Metastasis should be classified as “regional” (defined as locoregional lymph node involvement) or “distant” (organ location). Tumor recurrence and metastatic lesions should be stratified as either “confirmed” (i.e., histologically or cytologically, with histology preferred), or “suspected” (palpation,

imaging; see specific guidelines at vcgp.org). Histology is required to exclude non-neoplastic causes of a mass in the region of the surgical scar (e.g., reactive fibroplasia, suture reaction, gossypiboma), to confirm recurrence or to identify unrelated de novo neoplasms. Cytologic evaluation of fine needle aspirates may not be able to distinguish granulation tissue from neoplastic spindle cells or identify specific types of soft tissue tumors. Ideally, recurrences and metastases should be confirmed histologically, and complete autopsies performed in as many cases as possible in order to fully evaluate the extent of tumor progression. Otherwise, the level of certainty regarding the presence or absence of recurrence and metastasis in the study and, thus, the resultant grading system is compromised. For additional details, please see the guideline for outcome assessment (www.vcgp.org).

Neoplasms that are not metastatic or do not result in increased patient morbidity or mortality do not need to be graded. Similarly if the behavior of a tumor type cannot be subdivided due to uniformly aggressive behavior, a grading system may not be possible or needed. If investigators believe there are different behaviors for a tumor type historically considered indolent or there is an indolent subtype that may be difficult to recognize, then a grading system should be developed. While tumor entities with distinct histomorphologies can be readily assigned into distinct groups, further testing, such as immunohistochemistry, transmission electron microscopy or molecular and genetic tests, may be needed to separate tumor entities with similar histomorphology.

The criteria for tumor selection in a study must be specific enough to confirm the histologic tumor type being investigated. Consistency of tumor diagnoses is essential, especially when tumors within the group have different biological behaviors (some indolent and others aggressive). Tumor nomenclature should not imply a biological behavior at odds with that of tumors in the named category. For instance, the use of soft tissue sarcoma implies an aggressive behavior that does not reflect many of the tumors historically included in this group. It may be preferable to use the term tumor (i.e., soft tissue tumor, melanocytic tumor, hepatocellular tumor, parathyroid tumor, etc.) and provide information, based on grade or subtype if applicable, regarding expected behavior in a comment. Although current grading systems for canine and feline STTs lump different tumor types together, development of grading systems for each tumor within the group is recommended.

The current guideline addresses only histopathological features of tumor grading systems. Prognostic and predictive systems for neoplasms of the hematopoietic system, such as lymphomas and leukemias, rely on cellular and nuclear features, flow cytometry, cluster of differentiation and immunophenotyping. Histopathology is useful in human and veterinary oncology to classify lymphomas, but the integration of multiple methods may lead to the most accurate diagnoses, prognoses, and information for treatment of hematopoietic tumors.

The goal of our standardization effort is to suggest and facilitate uniform and reproducible assessments of tumors; however, investigators should also try to evaluate the judgment of an “experienced” pathologist in assessing histological changes. Every tumor is different, and all parameters cannot necessarily be enumerated or assessed as present or absent. Experience can affect a pathologist’s assessment of anisokaryosis,

atypia, MC, tumor necrosis vs. ischemia, etc.. How much experience impacts a pathologist's assessment of tumor histopathology is unclear. This phenomenon should be tested, comparing subjective tumor assessment among pathologists with different amounts of experience to the application of grading systems.

In addition, there are features that may not be included in current grading systems that can affect prognosis. Grading systems are developed using a small number of criteria for ease of use and may or may not be as good a prognostic assessment as the interpretation of an experienced pathologist, which includes more histologic criteria in conjunction with signalment, clinical history, and other available case information. This is the 'art' of histopathology. However, the pathologist must be familiar with and base their assessment on the current scientific literature.

Tumor grade and tumor markers that provide prognostic information may also prove useful to oncologists offering treatment options as predictive indicators of treatment response. Future grading systems may not just give indications on how long the animal will live (prognostic information) but may also provide information on what treatments to select to help the animal live a longer/better quality of life (predictive information). Oncology studies that correlate biomarkers and grading systems with different treatment outcomes are needed.

Which grading system and other pertinent information to be reported in a pathology report will remain the prerogative of pathologists and the laboratories by which they are employed. Grades should be reported if there is clinical value. There is no governing body that regulates veterinary pathology reports, and litigious situations are much less common in veterinary medicine than human medicine but are a consideration. Ideally, validation of grading systems should be done before they are applied to clinical cases. In the future, full characterization of tumors by histologic morphology, grade, cytology, immunohistochemical or genetic profiles relative to the host immunological status and molecular profile may yield greater utility in formulating prognosis and predicting treatments (precision oncology). The latter may prove to be the most useful application of our grading systems. Collaborating and combining expertise from various fields will enhance our knowledge of tumor behavior and will improve the care of pets with cancer.

References

1. Allison KH. Prognostic and predictive parameters in breast pathology: a pathologist's primer. *Mod Pathol*. 2021;34**:** 94-106. 10.1038/s41379-020-00704-7

2. Avallone G, Rasotto R, Chambers JK, et al. Review of histological grading systems in veterinary medicine. *Vet Pathol*. 2021;58**:** 809-828. doi: 10.1177/0300985821999831

3. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. *Nature*. 2012;483**:** 531-533. doi: 10.1038/483531a

4. Bergin IL, Smedley RC, Esplin DG, Spangler WL, Kiupel M. Prognostic evaluation of Ki67 threshold value in canine oral melanoma. *Vet Pathol*. 2011;48**:** 41-53. 10.1177/0300985810388947

5. Berlato D, Bulman-Fleming J, Clifford CA, et al. Value, limitations, and recommendations for grading of canine cutaneous mast cell tumors: a consensus of the oncology-pathology working group. *Vet Pathol*. 2021;58**:** 858-863. doi: 10.1177/03009858211009785

6. Boracchi P, Roccabianca P, Avallone G, Marano G. Kaplan-Meier curves, cox model, and p-values are not enough for the prognostic evaluation of tumor markers: statistical suggestions for a more comprehensive approach. *Vet Pathol*. 2021;58**:** 795-808. doi: 10.1177/03009858211014174

7. Camus MS, Priest HL, Koehler JW, et al. Cytologic Criteria for Mast Cell Tumor Grading in Dogs With Evaluation of Clinical Outcome. *Vet Pathol*. 2016;53**:** 1117-1123. doi: 10.1177/0300985816638721

8. Caswell JL, Bassel LL, Rothenburger JL, et al. Observational study design in veterinary pathology, part 2: methodology. *Vet Pathol*. 2018;55**:** 774-785. doi: 10.1177/0300985818798121

9. Dennis MM, McSporran KD, Bacon NJ, Schulman FY, Foster RA, Powers BE. Prognostic factors for cutaneous and subcutaneous soft tissue sarcomas in dogs. *Vet Pathol*. 2011;48**:** 73-84. doi: 10.1177/0300985810388820

10. Deyrup AT, Weiss SW. Grading of soft tissue sarcomas: the challenge of providing precise information in an imprecise world. *Histopathol*. 2006;48**:** 42-50. doi: 10.1111/j.1365-2559.2005.02288.x

11. Dobromylskyj MJ, Richards V, Smith KC. Prognostic factors and proposed grading system for cutaneous and subcutaneous soft tissue sarcomas in cats, based on a retrospective study. *J Feline Med Surg*. 2021;23**:** 168-174. doi: 10.1177/1098612x20942393

12. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45**:** 228-247. doi: 10.1016/j.ejca.2008.10.026

13. Esplin DG. Survival of dogs following surgical excision of histologically well-differentiated melanocytic neoplasms of the mucous membranes of the lips and oral cavity. *Vet Pathol*. 2008;45**:** 889-896. doi: 10.1354/vp.45-6-889

14. Kuntz CA, Dernell WS, Powers BE, Devitt C, Straw RC, Withrow SJ. Prognostic factors for surgical treatment of soft-tissue sarcomas in dogs: 75 cases (1986-1996). *J Am Vet Med Assoc*. 1997;211**:** 1147-1151.

15. Meuten DJ, Moore FM, Donovan TA, et al. International guidelines for veterinary tumor pathology: a call to action. *Vet Pathol*. 2021;58**:** 766-794. doi: 10.1177/03009858211013712

16. Mills SW, Musil KM, Davies JL, et al. Prognostic value of histologic grading for feline mammary carcinoma: a retrospective survival analysis. *Vet Pathol*. 2015;52**:** 238-249. doi: 10.1177/0300985814543198

17. Northrup NC, Howerth EW, Harmon BG, et al. Variation among pathologists in the histologic grading of canine cutaneous mast cell tumors with uniform use of a single

grading reference. *J Vet Diagn Invest*. 2005;17**:** 561-564. doi: 10.1177/104063870501700606

18. Oldenhuis CN, Oosting SF, Gietema JA, de Vries EG. Prognostic versus predictive value of biomarkers in oncology. *Eur J Cancer*. 2008;44**:** 946-953. doi: 10.1016/j.ejca.2008.03.006

19. Peña L, De Andrés PJ, Clemente M, Cuesta P, Pérez-Alenza MD. 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. *Vet Pathol*. 2013;50**:** 94-105. doi: 10.1177/0300985812447830

20. Perry JA, Culp WT, Dailey DD, Eickhoff JC, Kamstock DA, Thamm DH. Diagnostic accuracy of pre-treatment biopsy for grading soft tissue sarcomas in dogs. *Vet Comp Oncol*. 2014;12**:** 106-113. doi: 10.1111/j.1476-5829.2012.00333.x

21. Sabattini S, Scarpa F, Berlato D, Bettini G. Histologic grading of canine mast cell tumor: is 2 better than 3? *Vet Pathol*. 2015;52**:** 70-73. doi: 10.1177/0300985814521638

22. Santos IR, Raiter J, Lamego É C, et al. Feline pulmonary carcinoma: Gross, histological, metastatic, and immunohistochemical aspects. *Vet Pathol*. 2023;60**:** 8-20. doi: 10.1177/03009858221122517

23. Santos M, Correia-Gomes C, Santos A, de Matos A, Dias-Pereira P, Lopes C. Interobserver reproducibility of histological grading of canine simple mammary carcinomas. *J Comp Pathol*. 2015;153**:** 22-27. doi: 10.1016/j.jcpa.2015.04.005

24. Sauerbrei W, Taube SE, McShane LM, Cavenagh MM, Altman DG. Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): An Abridged Explanation and Elaboration. *J Natl Cancer Inst*. 2018;110**:** 803-811. doi: 10.1093/jnci/djy088

25. Schott CR, Tatiersky LJ, Foster RA, Wood GA. Histologic grade does not predict outcome in dogs with appendicular osteosarcoma receiving the standard of care. *Vet Pathol*. 2018;55**:** 202-211. doi: 10.1177/0300985817747329

26. Schulman FY, Roccabianca P, Avallone G, et al. Reporting guidelines for manuscripts on tumor prognosis. *Vet Pathol*. 2022;59**:** 397-398. doi: 10.1177/03009858221082207

27. Stefanello D, Buracco P, Sabattini S, et al. Comparison of 2- and 3-category histologic grading systems for predicting the presence of metastasis at the time of initial evaluation in dogs with cutaneous mast cell tumors: 386 cases (2009-2014). *J Am Vet Med Assoc*. 2015;246**:** 765-769. doi: 10.2460/javma.246.7.765

28. Trojani M, Contesso G, Coindre JM, et al. Soft-tissue sarcomas of adults; study of pathological prognostic variables and definition of a histopathological grading system. *Int J Cancer*. 1984;33**:** 37-42. doi: 10.1002/ijc.2910330108

29. Webster JD, Dennis MM, Dervisis N, et al. Recommended guidelines for the conduct and evaluation of prognostic studies in veterinary oncology. *Vet Pathol*. 2011;48**:** 7-18. doi: 10.1177/0300985810377187

30. Yap FW, Rasotto R, Priestnall SL, Parsons KJ, Stewart J. Intra- and inter-observer agreement in histological assessment of canine soft tissue sarcoma. *Vet Comp Oncol*. 2017;15**:** 1553-1557. doi: 10.1111/vco.12300

31. Young S, Kindzierski W, Meyer T. Understanding p-hacking and HARKing. *ResearchersOne*. 2022**:** https://researchers.one/articles/22.01.00007v00001.

Appendix

Appendix 1. Example of a table to summarize a grading approach using a scoring system.

Parameter	Definition	Categories	Score value
1)			
2)			
3)			

Histologic grade	Total score range
Low / I / 1	
Intermediate / II / 2	
High / III / 3	

Appendix 2. Example of a table to summarize the grading approach using a two-tier system.

Parameter	Definition	Threshold value
1)		
2)		
3)		
Histologic grade		
Tumors with a parameter value greater than threshold for more than x number of parameters are high grade, and the others are low grade.		