ACCEPTED MANUSCRIPT

Clinical Artificial Intelligence and Machine Learning Metrics 101: How to Evaluate Artificial Intelligence Tools for Veterinary Clinical Practices

Brian Hur, DVM, PhD[1,2,*] · Aparna Elangovan, PhD[3] · Laura Hardefeldt, BScBVMS, MPH, PhD, DACVIM[2]

1. Veterinary Information Network, Davis, CA, USA
2. Melbourne Veterinary School, University of Melbourne, Melbourne, VIC, Australia
3. Collate, Menlo Park, CA, USA

[*]Corresponding author: brian.hur@vin.com

This is the authors' accepted manuscript of an article published in final form as: Hur B, Elangovan A, Hardefeldt L. Clinical Artificial Intelligence and Machine Learning Metrics 101: How to Evaluate Artificial Intelligence Tools for Veterinary Clinical Practices. *Veterinary Clinics of North America: Small Animal Practice*. 2026;56(5).

Please cite the published version, available at www.sciencedirect.com/science/article/pii/S0195561626000318

SYNOPSIS

Artificial intelligence tools are rapidly entering veterinary medicine, yet clinicians often lack frameworks to evaluate their performance. This article provides a practical guide to understanding AI evaluation metrics, including sensitivity, specificity, precision, recall, and AUC-ROC, and explains why accuracy alone is insufficient. We address the critical role of inter-annotator agreement in establishing performance ceilings, the importance of external validation, and modality-specific evaluation considerations for AI scribes, digital imaging, and pathology applications. In the absence of regulatory oversight, veterinary professionals must develop evaluation literacy to make informed decisions about AI adoption.

KEY POINTS

- Evaluating AI tools requires understanding core performance metrics (sensitivity, specificity, precision, recall, F1 score, AUC-ROC) and recognizing that reported performance depends heavily on the population tested.
- AI performance is fundamentally limited by inter-annotator agreement among expert labelers, a concept often overlooked in validation studies that establishes the ceiling for model evaluation.
- External validation on independent datasets is essential, as AI tools frequently show degraded performance when deployed outside their original training environment due to domain shift.

- In the absence of regulatory oversight for veterinary AI, clinicians bear responsibility for critically evaluating tools before adoption using standardized evaluation frameworks.

KEYWORDS Artificial intelligence, machine learning, veterinary medicine, performance metrics, model validation, external validation, diagnostic AI, clinical decision support

1. Introduction

The adoption of artificial intelligence (AI) and machine learning (ML) in veterinary medicine has accelerated markedly over the past several years with tools now deployed across diverse clinical applications. [1–3] These range from ambient scribes that automate medical record generation, to computer vision systems that analyze radiographs and pathology slides, to natural language processing pipelines that extract structured data from millions of consultation records. [4–7] The pace of development shows no signs of slowing; if anything, the barrier to creating such tools continues to fall as the same AI technologies powering clinical applications can now assist in their own development.[8]

This lowered barrier to entry for AI development is simultaneously empowering and perilous.[9] In the same way we pay special attention to evaluating other novel therapies in the veterinary profession, we need to also understand how to evaluate AI and ML based tools. There are differences in evaluations that should be understood when applying or building such tools.[10] In the early days, developing tools required the ability to code and a deep understanding of ML, however the same underlying models powering the AI can now help develop the very systems that use AI. This could allow for the proliferation of tools by creators with little or no prior coding experience.[11,12] While these AI and ML tools can be useful, their value can also be overstated, and may in fact be harmful if not evaluated properly. It is imperative that all those engaged in their use understand the mechanisms used for evaluation.[13,14]

The veterinary landscape is unique in many ways which creates challenges when evaluating AI tools.[1,15] Species variation, building AI tools with limited test data compared to human medicine, and diverse practice settings (small animal, large animal, exotic and emergency) all present significant issues.[16] Furthermore, the field is often less regulated than human medicine and in many regions there are no governing boards to control how individual veterinarians use AI tools.[17,18] An AI tool that might be evaluated in one setting, such as a university teaching hospital, could have a drastically decreased performance when tested in another, such as a general practice clinic, due to their many differences in environments and workflows, which is often referred to as domain shift.[19,20] This was notably demonstrated by a widely deployed sepsis prediction algorithm in human medicine which, upon independent external validation, exhibited substantially degraded performance compared to developer-reported metrics while missing the majority of actual sepsis cases.[21]

The veterinary community stands to benefit from the use of AI tools in clinics through improving communication time with patients, simplifying note taking, and reducing the risk of burnout in practitioners. [22–24] Additionally, it can expand research capability through allowing population-based studies rather than relying on manual evaluation of relatively small numbers of clinical records.[25] It also has the potential to evaluate longitudinal records to better understand how treatments can impact patient outcomes.[26]

The purpose of this article is to provide a practical framework for evaluating AI and ML tools in veterinary medicine, applicable both to clinicians assessing commercial products for adoption and to those developing novel applications. The aim is to equip veterinary professionals, regardless of technical background, with the conceptual foundations and practical heuristics necessary for informed decision making. The framework addresses evaluation across the major AI modalities currently deployed in veterinary practice: documentation and scribe tools, diagnostic imaging, pathology and cytology, dental radiography, and cardiology applications. Throughout, we emphasize that evaluation is not a one-time event but an ongoing process extending from initial assessment through real world deployment and monitoring.

2. Background

Veterinary medicine presents distinct context for AI development with three fundamental challenges: biological heterogeneity across species and breeds, data scarcity and quality issues, and an absence of regulatory oversight. We address each in turn.

2.1 Species and Breed Challenges

There are anatomical and physiological diversity among species, breeds, and individual patients.[27,28] Detecting cardiomegaly on canine thoracic radiographs requires accounting for chest conformation, with the barrel-chested Bulldog, the deep narrow Greyhound thorax, and the keel shaped Great Dane chest each representing different normal anatomy. A model claiming to “detect heart disease” must specify its target population - all dogs vs specific breeds vs cats. This challenge compounds across species. AI systems trained on canine data cannot transfer directly to felines, even for the same disease.[29] The radiographic appearance of pneumonia differs fundamentally between dogs, cats, and horses. Each species effectively constitutes a separate domain requiring dedicated validation.

2.2 Limited Data

Human medical AI benefits from decades of incentivized digitization through meaningful use mandates and standardized coding systems.[30] Veterinary medicine lacks equivalent infrastructure, resulting in a “low-resource” data environment of unstructured free text.[31] A diagnosis of pyoderma might exist as a sentence fragment, “skin looks red, susp pyoderma, dispensing ceph” rather than a coded entry. When AI was used on 4.4 million companion

animal records, complete extraction of drug dose, duration, and diagnosis was achievable for only 39% of canine and 37% of feline records.[27] This reflects documentation reality, not algorithmic failure, with direct implications for any AI trained on such data.

However, the veterinary field does benefit from some data advantages. Large-scale clinical databases such as VetCompass (UK and Australia),[32] SAVSNET,[33] and the Dog Aging Project[34] grant researchers access to millions of de-identified clinical records for population-based studies. Unlike human medicine, which is governed by strict regulations in the United States (e.g., Health Insurance Portability and Accountability Act, HIPPA), veterinary records face relatively minimal legal regulation regarding research access, potentially improving data availability for AI development.[35] These resources enable epidemiological research at a scale that would be considerably more difficult to achieve in human medicine due to privacy and regulatory constraints. In addition, growing interest in 'big data' for veterinary research is resulting in increasing annotated data sets.

2.3 Diverse Practice Settings

Veterinary medicine encompasses fundamentally different environments. For example, a high-volume vaccine clinic that uses checkboxes for its exam records, a general practice with abbreviated documentation, a specialty referral with complex cases and detailed records, or emergency care with time pressured notes which may contain incomplete histories. Furthermore, it could be an exotic species or infrequently used medicine. An AI tool validated in referral hospitals may perform poorly in general practice.[19] This is not because of a flawed system, but because population, documentation style, and clinical context differ substantially.

2.4 Regulatory Gaps

The FDA regulates human AI medical devices, requiring safety and efficacy demonstration before market release.[36] No equivalent exists for veterinary AI. Products can reach market without having any oversight beyond what is done by the developer of such products.[17] This means that there are no mandated validation standards, no required transparency about training data, and shifted evaluation burden to individual veterinarians.

2.5 The Domain Transfer Problem

These factors converge in domain shift: models trained in one context fail to generalize to another.[37,38] Standard evaluations use hold-out sets from training data, demonstrating only internal validity, that the model learned patterns in that specific dataset. External validity, measured by performance on genuinely independent data, requires separate assessment. For example, a model trained on university hospital data might learn that "lymphadenopathy predicts"lymphoma," because in referral populations it often does. Deployed in primary care, where lymphadenopathy commonly reflects dental infection or reactive inflammation, the model will systematically over-predict neoplasia. The learned

correlation was domain specific, not medically universal.[39] Beyond population-level shifts, models may also learn non-clinical artifacts in training data. An NLP model might, for instance, base diagnostic predictions partly on documentation style or service-specific phrasing rather than clinical content, learning to associate linguistic patterns with writing styles rather than with underlying pathology.

3. Understanding AI Performance Metrics

Evaluating AI tools requires navigating two metric traditions that use different terminology and emphasize different aspects of performance. The AI/ML community typically reports precision, recall, and F1-score. Alternatively, clinical literature favors sensitivity, specificity, and predictive values. These frameworks overlap but are not equivalent, and understanding both is necessary when assessing tools that may report either. Rigorous evaluation requires understanding the metrics used to characterize AI performance.

3.1 The Confusion Matrix: Foundation of All Classification Metrics

All classification metrics derive from the confusion matrix, the same framework used for evaluating any diagnostic test, which categorizes predictions into four outcomes (Table 1)[40]:

Table 1. Confusion matrix diagram

	Condition Present	**Condition Absent**
AI Positive (Test Positive)	**True Positive** (TP)	**False Positive** (FP)
AI Negative (Test Negative)	**False Negative** (FN)	**True Negative** (TN)

A **true positive** occurs when the AI correctly identifies a condition that is present. A **false positive** (Type I error) occurs when the AI incorrectly flags a condition that is absent. A **false negative** (Type II error) occurs when the AI fails to identify a condition that is present. A **true negative** occurs when the AI correctly identifies absence of a condition.

Veterinarians routinely apply this reasoning when interpreting laboratory results or imaging findings. AI evaluation requires the same logic, applied with equal rigor. The clinical consequences of these errors differ dramatically by context.[40] For example, consider if an AI system is being relied on for diagnosis of a disease such as osteosarcoma. A false negative for osteosarcoma delays diagnosis of a life-threatening condition. A false positive for the same condition may precipitate unnecessary treatment or even euthanasia. Evaluation must consider not just error rates but error consequences.

3.2 The Machine Learning Framework: Precision and Recall

Machine learning research conventionally reports precision and recall, typically combined into an F1-score.[41]

Recall (equivalent to sensitivity) answers: of all positive cases, what proportion did the system identify?

$$\text{Recall} = \frac{TP}{TP + \text{FN}}$$

Precision (equivalent to positive predictive value) answers: of all cases the system flagged as positive, what proportion were correct?

$$\text{Precision} = \frac{TP}{TP + \text{FP}}$$

The F1-score is the harmonic mean (a type of average that penalizes extreme imbalance between two values) of precision and recall:

$$\text{F1} = \frac{2 \times (P \times R)}{P + R}$$

This framework emphasizes positive class performance, appropriate for ML tasks where the negative class is large, heterogeneous, or not well defined.[41] True negatives do not appear in precision or recall calculations. When evaluating an information extraction system that identifies drug mentions in clinical notes, for instance the universe of "non-drug-mentions" is not well defined, making specificity calculations problematic. Precision and recall avoid this issue by focusing exclusively on positive predictions and positive cases.

3.3 The Clinical Framework: Sensitivity and Specificity

Clinical medicine conventionally reports sensitivity and specificity, which treat both classes symmetrically.[42]

Sensitivity (also called recall or true positive rate): of all cases with the condition, what proportion did the system correctly identify as positive?

$$\text{Sensitivity} = \frac{TP}{TP + \text{FN}}$$

High sensitivity means few missed cases, critical for screening applications where failing to detect disease is unacceptable. A pneumonia detection AI with 95% sensitivity misses 5% of pneumonia cases.

Specificity: of all cases without the condition, what proportion did the system correctly identify as negative?

$$\text{Specificity} = \frac{TN}{TN + \text{FP}}$$

High specificity means few false alarms, critical when false positives lead to unnecessary intervention, client anxiety, or irreversible decisions. Consider canine brucellosis: a false positive could lead to euthanasia and significant concern about zoonotic transmission to the family.

Note that specificity and precision are not equivalent, they answer fundamentally different questions.[42] Specificity concerns performance on true negatives (of all healthy patients, how many were correctly identified as healthy?) whereas precision concerns reliability of positive predictions (of all patients flagged as diseased, how many actually were?). A system can exhibit high specificity with low precision, or vice versa, depending on disease prevalence.

Clinicians also use positive **predictive value (PPV)** and **negative predictive value (NPV):**

$$\text{PPV} = \frac{TP}{TP + \text{FP}} = \text{Precision}$$

$$\text{NPV} = \frac{TN}{TN + \text{FN}}$$

PPV is mathematically identical to precision; the clinical and ML communities use different terminology for the same quantity.[43]

3.4 The Sensitivity-Specificity Tradeoff

Sensitivity and specificity are inversely related at any given decision threshold.[44] Adjusting an AI to catch more true positives (high sensitivity) typically increases false positives (lower specificity). Developers chose thresholds based on the relative costs of different errors, but these choices may not align with a specific clinical context, so it is critical to consider this independently.

A screening tool optimized for sensitivity will flag borderline cases, generating false positives that require follow up. A confirmatory tool optimized for specificity will miss borderline cases but provide high confidence when positive. Neither optimization is inherently correct; the appropriate threshold depends on clinical application.[44]

3.5 The Prevalence Problem

Precision (equivalently, PPV) depends heavily on disease prevalence in ways that sensitivity and specificity do not.[42] This is why clinical researchers often favor sensitivity and specificity, they remain stable across populations with different disease rates, enabling comparison across settings.

Consider an AI with 95% sensitivity and 95% specificity applied to two populations (Table 2):

Table 2. Demonstration on effects of prevalence on precision with two populations that have identical specificity and sensitivity.

	Population A (50% prevalence)	Population B (5% prevalence)
True positive	475	47
False negative	25	3
True negative	475	902
False positive	25	48
Precision (PPV)	**475/(475+25) = 95%**	**47/(47+48) = 49%**

The same system, identical sensitivity and specificity, yields 95% precision in one population and 49% in another.[42] If a tool reports precision from a validation set with artificially balanced classes, real-world performance in low-prevalence settings will be substantially worse. Dataset balancing, where researchers intentionally include roughly equal numbers of positive and negative cases, is standard practice in machine learning development because it improves model training. However, it creates validation conditions that do not reflect clinical reality.

This matters for veterinary applications where many conditions are rare. Using the osteosarcoma example, an AI tool may demonstrate excellent precision in a validation study using university hospital data as it is enriched with positive cases; it may produce mostly false positives in general practice where true prevalence is low. When precision or F1 are reported without prevalence information, caution is needed.

3.6 Translating Between Frameworks

In veterinary AI tool evaluation both clinical and computers science terminologies are used. [43] ML originated tools (e.g., Natural Language Processing systems looking at text, general computer vision) typically report precision, recall, and F1. Clinically validated tools (e.g., diagnostic imaging, screening applications) typically report sensitivity, specificity, and PPV/NPV. An ML paper reporting only precision and recall provide no direct information about specificity. A clinical paper reporting sensitivity and specificity enables precision calculation only if prevalence is known. Ideally, the underlying data necessary for a confusion matrix is provided. This allows better comparison when evaluating tools across different validation studies and ensures that the equivalent metrics are being compared on populations with comparable prevalence.

3.7 AUC-ROC: Threshold-Independent Performance

Every AI classification system operates with an internal threshold determining how suspicious it must be before flagging a case as positive. This threshold is not fixed and represents a design choice with direct clinical consequences. Lower that threshold and the

system flags more true disease (higher sensitivity) but generates more false alarms (lower specificity). Raise the threshold and false positives decrease but false negatives increase. Every threshold embodies a tradeoff; there is no setting that optimizes both simultaneously.

The Receiver Operating Characteristic (ROC) curve makes this tradeoff space explicit.[45] Formally, the ROC curve plots sensitivity (true positive rate) against 1 – specificity (false positive rate) across all possible thresholds (Figure 1). Each point on the curve represents a different operating threshold: the upper left corner (0, 1) represents ideal performance, perfect sensitivity with perfect specificity, while the diagonal line represents random chance (AUC = 0.5). Any useful classifier should perform above this diagonal, with curves bowing further toward the upper left corner indicating better discriminative ability.

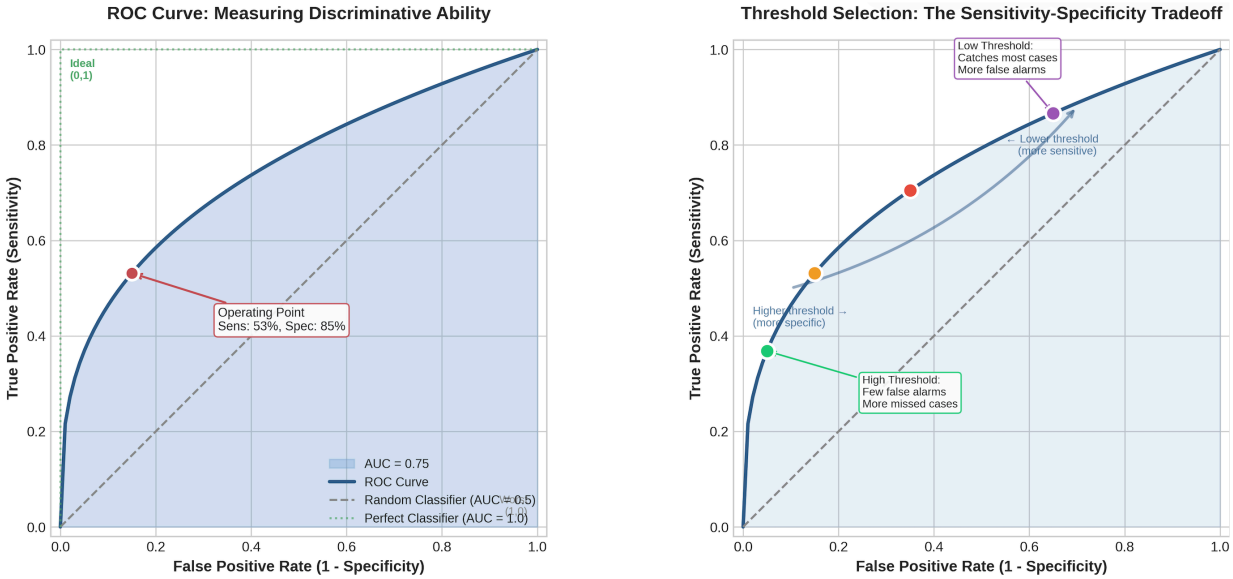

Figure 1. (Left) A ROC curve with AUC = 0.75 showing the area under the curve shaded in blue. The operating point indicates the chosen threshold (sensitivity 53%, specificity 85%). The dashed diagonal represents random chance. (Right) How threshold selection creates the sensitivity-specificity tradeoff: lowering the threshold increases sensitivity but reduces specificity.

The Area Under the ROC Curve (AUC-ROC) summarizes this curve as a single value quantifying discriminative ability independent of threshold selection. AUC answers the question, "Given one case with disease and one without, how often will the system correctly assign higher suspicion to the diseased case?" An AUC of 0.90 means correct ranking occurs 90% of the time.[45] Figure 2 illustrates how classifiers with different AUC values exhibit varying discriminative abilities.

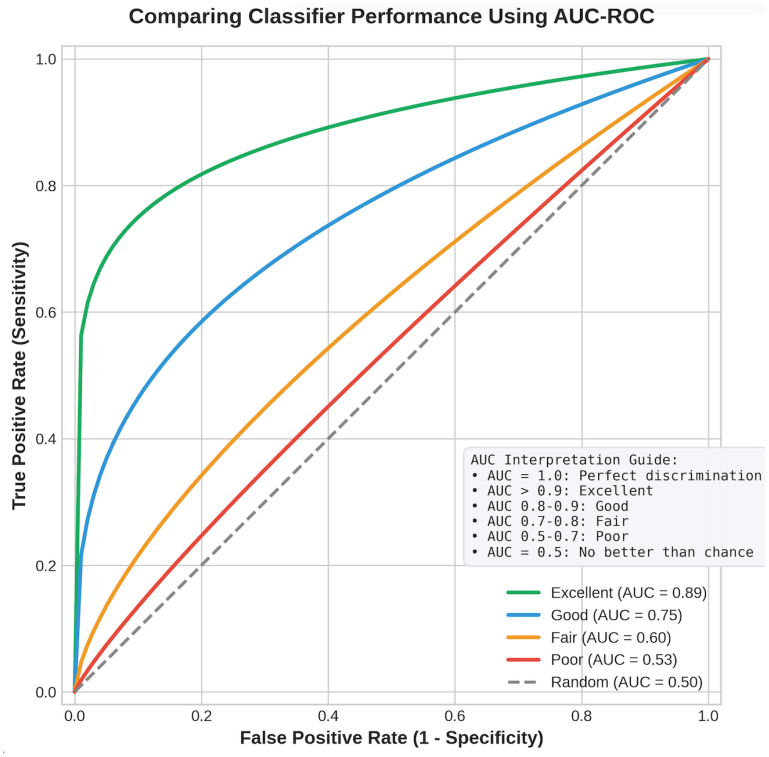

Figure 2. Comparison of classifiers with different discriminative abilities. Higher AUC values correspond to curves that bow further toward the upper left corner, indicating better separation between positive and negative cases. The interpretation guide shows standard benchmarks for AUC assessment.

General AUC interpretation guidelines:

- AUC = 1.0: Perfect discrimination
- AUC > 0.9: Generally considered excellent
- AUC 0.8-0.9: Good
- AUC 0.7-0.8: Fair
- AUC = 0.5: No better than chance

AUC is useful for comparing systems but has significant limitations that must be understood.

First, AUC measures discrimination, whether the system can distinguish diseased from healthy cases, rather than classification accuracy at any specific operating point. An AUC of 0.95 does not mean 95% accuracy; it means that when presented with one diseased and one healthy case, the system assigns higher suspicion to the diseased case 95% of the time. Clinical deployment requires selecting a specific threshold, at which point the sensitivity-specificity tradeoffs reassert themselves. A system with excellent AUC may still perform poorly at the threshold chosen for deployment.

Second, AUC can mislead when disease prevalence is low (a situation ML researchers call “class imbalance”), common in veterinary applications where most patients do not have the condition being screened for.[46] The ROC curve treats sensitivity symmetrically, which can obscure poor positive predictive performance in imbalanced datasets. In such settings, precision-recall curves and the corresponding area under the precision-recall curve (AUC-PR) may provide more informative summaries.[41] A system with high AUC-ROC may produce unacceptably low precision at any practice operating threshold when the positive class is rare.

The practical implication is clear: when AUC is reported, performance metrics at the specific threshold used for deployment are needed for evaluation. AUC alone, without accompanying sensitivity, specificity, and precision at the operational threshold, provides insufficient information for clinical evaluation.

3.8 Calibration: When Confidence Matches Reality

Many AI systems output not just predictions but confidence scores such as"85% probability of pneumonia" rather than simply "pneumonia detected." These probability estimates are useful only if they mean what they claim to mean. A system calibrated well produces confidence scores that match actual frequencies: if the system assigns 80% confidence to 100 predictions, approximately 80 should prove correct.

Calibration matters because confidence scores often inform clinical decisions. A finding flagged with 95% confidence may prompt different action than one flagged at 60%. If these numbers are unreliable, they mislead rather than inform.

Poor calibration is common in deep learning models, which frequently exhibit systematic overconfidence.[47] An AI reporting "95% confident this is osteosarcoma" may be correct far less than 95% of the time since the number reflects the model's internal mathematics rather than true diagnostic probability. This is not merely a minor technical concern; it directly affects how clinicians interpret and act on AI outputs.

Calibration can be assessed through reliability diagrams, which plot predicted probability against observed frequency (Figure 3), and quantified through metrics such as the Brier score or expected calibration error (for both, lower values indicate better calibration).[47] When underlying probability scores are inaccessible, as with many large language models (LLMs), alternative approaches include prompting the model to express confidence verbally or generating multiple responses and treating inconsistency as a signal of uncertainty.[48] These methods remain less validated than traditional calibration assessments.

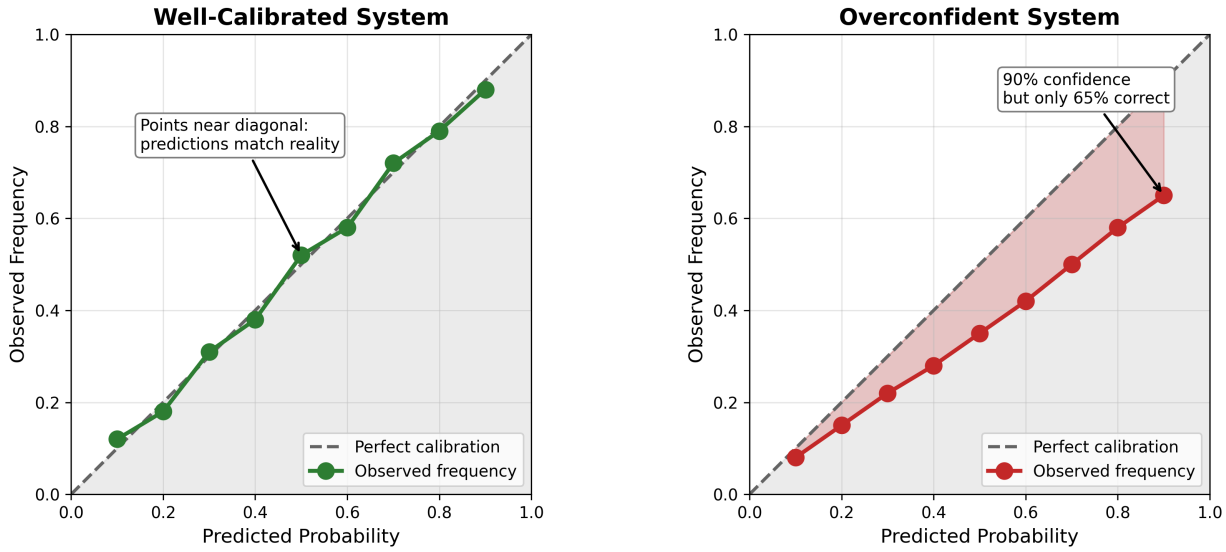

Figure 3. Reliability diagrams comparing a well-calibrated system (left) with an overconfident system (right). Points along the diagonal indicate good calibration; deviation below the diagonal indicates overconfidence.

When a tool provides confidence scores, without an evaluation of the calibration, it should be treated as relative rankings (higher means more suspicious) rather than true probabilities.

3.9 Why Accuracy is Misleading

Accuracy is the proportion of all predictions that are correct and is the most commonly reported metric and frequently the most misleading.[49]

$$\text{Accuracy} = \frac{TP + \text{TN}}{TP + \text{TN} + \text{FP} + \text{FN}}$$

Consider this example. An AI tool is developed to detect a condition that has 5% prevalence in the target population. A system that predicts "negative" for every case, ignoring all input data entirely, achieves 95% accuracy. It has learned nothing, detects nothing, and provides zero clinical value. Yet by the accuracy metric, it appears highly successful.[49] This is not a contrived scenario. Many clinically important conditions are rare, and validation datasets are often constructed with artificial balance (50% positive, 50% negative) that inflates accuracy relative to real-world performance. A system reporting 90% accuracy on a balanced validation set may perform very differently when deployed in a population where true prevalence is 2%.

The practical implication is straightforward: when evaluating AI tools, prior to being considered for use, it should be required to provide sensitivity and specificity (or precision and recall) reported separately for each class. These metrics reveal how the system performs on positive and negative cases independently. Accuracy may be reported for completeness, but it should never be the primary metric, and comparisons of accuracy across datasets with different prevalence rates are not meaningful. Aggregate accuracy

alone, without breaking down the specifics of each class and not having prevalence information, is insufficient for clinical assessment. Its presence as the primary reported metric, without accompanying performance of each class, should prompt skepticism about validation rigor.

4. Inter-Annotator Agreement: The Performance Ceiling

A critical but often overlooked aspect of AI evaluation is the quality of the "ground truth" labels against which performance is measured.

4.1 The Labeling Problem

Supervised AI learns from labeled examples. For example, radiographs annotated as "pneumonia present" or "normal," or histopathology slides classified by tumor grade. These labels are generated by human experts. If experts disagree about correct labels, AI performance is fundamentally constrained by that disagreement. A system cannot be evaluated as better than its training labels when those labels constitute the only available ground truth.

Inter-annotator agreement (IAA) quantifies consistency among human labelers.[50,51] The most common measure is Cohen's Kappa, which calculates agreement between two annotators while correcting for chance:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Where p_o is observed agreement and p_e is chance-expected agreement. Fleiss' Kappa extends this calculation to more than two annotators.

Interpretation guidelines:[50]

- $\kappa > 0.80$: Excellent agreement
- κ 0.60–0.80: Substantial agreement
- κ 0.40–0.60: Moderate agreement
- $\kappa < 0.40$: Poor agreement

These thresholds are conventions rather than absolute standards but provide useful benchmarks for assessing label quality. However, IAA metrics have notable limitations: different statistics can yield substantially different scores for the same data, and inherently complex judgment tasks produce lower agreement regardless of label quality, making context essential when interpreting these values.[52]

4.2 Why IAA Constrains Evaluation

Consider a concrete example: two board-certified radiologists reviewing the same set of thoracic radiographs for pneumonia agree 85% of the time ($\kappa \approx 0.70$). The remaining 15% represents cases where expert opinion genuinely diverges, potentially due to borderline presentations, ambiguous findings, or cases where reasonable specialists reach different conclusions.

An AI system trained and evaluated on labels from one radiologist cannot be meaningfully said to "exceed" 85% performance. If the system achieves 95% agreement with Radiologist A, it may simply be achieving 80% agreement with Radiologist B. Without knowing which radiologist is "correct" in disputed cases, often an unanswerable question, we cannot determine whether the system has learned genuine diagnostic accuracy or merely one expert's labeling tendencies.

This has direct implications for interpreting validation studies. AI performance should be contextualized against IAA, not assumed perfect ground truth.[51] A system achieving 88% accuracy against labels with 85% inter-annotator agreement is performing at approximately the level of expert disagreement. A system achieving 95% accuracy against labels with 70% IAA warrants scrutiny, and it may indicate methodological problems such as test set contamination (training on data inadvertently included in the evaluation set) or label leakage (the model learning to predict labels from artifacts in the data rather than clinical features) rather than genuinely superior performance.

Low IAA signals task ambiguity. When experts cannot agree, the problem may lie not with the AI but with the clinical concept itself. Some diagnostic categories are inherently fuzzy; their boundaries depend on threshold choices that reasonable clinicians make differently. In such cases, the appropriate response may be task redefinition rather than algorithmic improvement.

In rare cases, AI might genuinely detect patterns imperceptible to human experts, as demonstrated by deep learning systems that predict cardiovascular risk factors from retinal images that clinicians cannot visually assess.[53] Validating such claims requires outcome-based validation, correlating predictions with histopathology, treatment response, or survival, rather than relying on expert labels as ground truth.[54]

IAA establishes evaluation ceilings. If IAA calculation is not performed before AI evaluation, there is no way to determine whether reported performance represents genuine capability or reflects artifacts of how labels were generated.

4.3 Implications for Veterinary AI

Veterinary AI faces challenges in establishing reliable ground truth:

Limited specialist availability. Board-certified veterinary radiologists, pathologists, and cardiologists are scarce relative to case volume. Large-scale labeling efforts may rely on general practitioners whose diagnostic accuracy differs from specialist consensus. An expert in canine cardiology may not be an expert in feline cardiology. Labels generated without appropriate species-specific expertise may reflect knowledge gaps rather than ground truth.

Retrospective labeling limitations. Training labels derived from clinical records use outcomes (e.g., histopathology results, response to treatment, clinical progression) as proxies for diagnostic truth. These proxies introduce systematic biases, cases that proceeded to biopsy may be consistently different from those managed conservatively and treatment response conflates diagnosis accuracy with treatment selection.

When evaluating AI tools, the relevant questions must be asked: How were training labels generated? By whom? Was inter-annotator agreement measured and reported? Tools validated against labels from non-specialists, or without IAA assessment, provide insufficient evidence for clinical adoption. The absence of IAA reporting should raise concerns about validation rigor.

5. Validation Hierarchy and External Validity

5.1 Types of Validation

Validation exists on a hierarchy of strength:

Internal validation tests performance on hold out data from the same source as training data. This demonstrates the model learned patterns in that dataset but says nothing about generalizability.

External validation tests on genuinely independent data, from a different institution, time period, or population. This provides evidence of transportability.

Prospective validation tests on new cases collected after model development, under real clinical conditions. This is the strongest evidence but rarely performed.

Multi-site validation across diverse institutions with different equipment, populations, and practices provides the most confidence in generalizability.

5.2 The Generalization Gap

Published evidence demonstrates consistent "generalization gaps" where AI performance degrades on external data.[55] A systematic review of deep learning algorithms for radiologic diagnosis found that the vast majority (81%) of externally validated models showed decreased performance compared to internal validation, with nearly half (49%) showing at least a modest decrease of 0.05 or more in AUC, and nearly a quarter (24%) showing substantial drops of 0.10 or greater.[55]

Several factors drive these performance gaps. Disease prevalence, case demographics, and severity distributions differ between institutions. A model trained where a condition is common may encounter fundamentally different patient populations elsewhere. Image characteristics also vary considerably across sites due to differences in equipment, positioning protocols, and image quality standards. Labeling practices introduce additional variability, as what one institution calls a particular finding may not precisely match another's diagnostic criteria. Perhaps most problematic are spurious correlations, where models learn to rely on features that happen to predict outcomes in training data but have no true diagnostic value. These might include institutional documentation templates, equipment-specific artifacts, or imaging protocols unique to the training environment.

For LLMs, the relationship between memorization and generalization differs by task type: knowledge retrieval tasks depend heavily on whether information appeared in training data, while reasoning tasks show greater generalization, with performance less tied to training data frequency.[56,57]

5.3 Minimum Validation Requirements

For clinical deployment, certain validation standards should be considered non-negotiable. Validation of any AI product being used in a clinic should include:

- External validation on independent data
- Disclosure of validation population characteristics (e.g., disease prevalence, referral vs. primary care settings)
- Performance on relevant subgroups (e.g., species, breeds, disease severity)
- Confidence intervals reflecting estimate precision (which requires adequate sample sizes)

Developer-reported metrics without independent verification warrant skepticism. Peer-reviewed publication provides more confidence than marketing materials, but even published studies may have methodological limitations.

6. Modality-Specific Considerations

6.1 AI Scribes and Documentation

AI scribes generate SOAP (Subjective, Objective, Assessment, Plan, the standard clinical documentation format) notes from consultation audio using speech recognition and LLMs. Unlike diagnostic AI, "ground truth" is partly subjective as note quality involves clinical completeness, appropriate terminology, and accurate representation of the encounter.

Evaluation considerations specific to AI scribes include:

Hallucination risk: Generative AI can produce plausible but fabricated content, and systematic verification against audio recordings is essential.[58] Studies evaluating LLM-generated clinical documentation have observed hallucination rates of 1.47% and omission rates of 3.45%, with errors potentially carrying significant clinical safety and legal implications.[58]

Terminology accuracy: Does the tool handle veterinary-specific terminology, abbreviations, and species-specific language?

Completeness metrics: What proportion of clinically relevant information from encounters is captured?

The evidence base for veterinary AI scribes remains limited, with few peer-reviewed validation studies.

6.2 Diagnostic Imaging AI

Digital imaging AI represents the most mature veterinary application, spanning radiography, echocardiography, and other modalities. However, the peer-reviewed literature remains relatively sparse with few publications utilizing machine learning for imaging-associated tasks across veterinary clinical and biomedical research.[6]

Evaluation considerations include:

Finding coverage: Tools detect finite lists of abnormalities. Conditions outside that list will not be identified.

Species/breed validation: Anatomical conformation varies dramatically (thoracic shape in radiography, cardiac anatomy in echocardiography) and requires validation across relevant populations, including breed-specific conditions such as cardiomyopathies.

Sensitivity vs. specificity tradeoffs: Understanding how the tool is optimized, whether for screening (high sensitivity) or confirmation (high specificity).

Comparison to radiologist standard: Performance should be benchmarked against board-certified interpretation, with inter-reader agreement reported.[59]

6.3 Pathology and Cytology AI

Pathology AI applications include tumor grading, lymphoma classification, and automated cell counting. Unique considerations:

Preparation variation: Staining protocols and slide preparation affect image characteristics.

Scanner calibration: Different whole-slide imaging platforms produce different outputs.

Inter-pathologist agreement: The gold standard itself shows variability; AI performance should be contextualized against expert agreement.

7. Summary

AI tools offer substantial potential to improve veterinary medicine. Realizing this potential requires evaluation literacy, and understanding metrics such as sensitivity, specificity, precision, recall, F1 score, AUC, and AUC-ROC, the role of inter-annotator agreement, as well as the critical importance of external validation.

The core questions are straightforward: What is the sensitivity and specificity? On what population? Compared to what expert baseline? With what external validation? These are not obstacles to adoption but prerequisites for responsible adoption.

In the absence of regulatory oversight, the profession must develop and apply rigorous standards. The framework presented here, grounded in quantitative evaluation metrics and practical checklists, equips veterinary professionals to make informed decisions. As AI proliferates, these evaluation skills become as essential as any other clinical competency.

Care Points

1. Evaluating AI tools requires understanding core performance metrics (sensitivity, specificity, precision, recall, F1 score, AUC, and AUC-ROC) and recognizing that reported performance depends heavily on the population tested.
2. AI performance is fundamentally limited by inter-annotator agreement (IAA) among the expert labelers who created the training data, a concept often overlooked in validation studies.
3. External validation on independent datasets is essential, as AI tools frequently show degraded performance when deployed outside their original training environment, and without regulatory oversight, clinicians bear responsibility for critical evaluation before adoption.

Disclosures

B. Hur is employed by Veterinary Information Network. This employment did not influence the content of this article. The remaining authors have nothing to disclose.

References

1. Sobkowich KE. Demystifying artificial intelligence for veterinary professionals: practical applications and future potential. *American Journal of Veterinary Research*. 2025;86(S1):S6-S15. doi:10.2460/ajvr.24.09.0275
2. Albergante L, O'Flynn C, Meyer GD. Artificial intelligence is beginning to create value for selected small animal veterinary applications while remaining immature for others. *Journal of the American Veterinary Medical Association*. 2025;263(3):388-394. doi:10.2460/javma.24.09.0617

3. Gabor S, Danylenko G, Voegeli B. Familiarity with artificial intelligence drives optimism and adoption among veterinary professionals: 2024 survey. *American Journal of Veterinary Research*. 2025;86(S1):S63-S69. doi:10.2460/ajvr.24.10.0293
4. Hur BA, Hardefeldt LY, Verspoor KM, Baldwin T, Gilkerson JR. Describing the antimicrobial usage patterns of companion animal veterinary practices; free text analysis of more than 4.4 million consultation records. *PLOS ONE*. 2020;15(3):1-15. doi:10.1371/journal.pone.0230049
5. Chu CP. ChatGPT in veterinary medicine: a practical guidance of generative artificial intelligence in clinics, education, and research. *Front Vet Sci*. 2024;11. doi:10.3389/fvets.2024.1395934
6. Hennessey E, DiFazio M, Hennessey R, Cassel N. Artificial intelligence in veterinary diagnostic imaging: A literature review. *Vet Radiol Ultrasound*. 2022;63 Suppl 1:851-870. doi:10.1111/vru.13163
7. McGuire KC. AAVSB Releases Considerations for the Use of AI in Veterinary Medicine. AAVSB. April 21, 2025. Accessed December 5, 2025. https://www.aavsb.org/news/aavsb-releases-guidance-for-the-use-of-ai-in-veterinary-medicine/
8. Noy S, Zhang W. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*. 2023;381(6654):187-192. doi:10.1126/science.adh2586
9. Rubeis G, Dubbala K, Metzler I. "Democratizing" artificial intelligence in medicine and healthcare: Mapping the uses of an elusive term. *Front Genet*. 2022;13. doi:10.3389/fgene.2022.902542
10. Lam TYT, Cheung MFK, Munro YL, Lim KM, Shung D, Sung JJY. Randomized Controlled Trials of Artificial Intelligence in Clinical Practice: Systematic Review. *J Med Internet Res*. 2022;24(8):e37188. doi:10.2196/37188
11. Rodriguez DV, Lawrence K, Gonzalez J, et al. Leveraging Generative AI Tools to Support the Development of Digital Solutions in Health Care Research: Case Study. *JMIR Human Factors*. 2024;11(1):e52885. doi:10.2196/52885
12. Chow M, Ng O. From technology adopters to creators: Leveraging AI-assisted vibe coding to transform clinical teaching and learning. *Med Teach*. 2025;47(12):1927-1929. doi:10.1080/0142159X.2025.2488353
13. European Parliament. Directorate General for Parliamentary Research Services. *Artificial Intelligence in Healthcare: Applications, Risks, and Ethical and Societal Impacts*. Publications Office; 2022. Accessed December 24, 2025. https://data.europa.eu/doi/10.2861/568473
14. Botha NN, Segbedzi CE, Dumahasi VK, et al. Artificial intelligence in healthcare: a scoping review of perceived threats to patient rights and safety. *Arch Public Health*. 2024;82:188. doi:10.1186/s13690-024-01414-1
15. Akinsulie OC, Idris I, Aliyu VA, et al. The potential application of artificial intelligence in veterinary clinical practice and biomedical research. *Front Vet Sci*. 2024;11:1347550. doi:10.3389/fvets.2024.1347550
16. Akbarein H, Taaghi MH, Mohebbi M, Soufizadeh P. Applications and Considerations of Artificial Intelligence in Veterinary Sciences: A Narrative Review. *Vet Med Sci*. 2025;11(3):e70315. doi:10.1002/vms3.70315
17. Cohen EB, Gordon IK. First, do no harm. Ethical and legal issues of artificial intelligence and machine learning in veterinary radiology and radiation oncology. *Veterinary Radiology & Ultrasound*. 2022;63(S1):840-850. doi:10.1111/vru.13171
18. Bellamy JEC. Artificial intelligence in veterinary medicine requires regulation. *Can Vet J*. 64(10):968-970.
19. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*. 2020;21(2):345-352. doi:10.1093/biostatistics/kxz041
20. Goetz L, Seedat N, Vandersluis R, van der Schaar M. Generalization-a key challenge for responsible AI in patient-facing clinical applications. *NPJ Digit Med*. 2024;7(1):126. doi:10.1038/s41746-024-01127-3
21. Wong A, Otles E, Donnelly JP, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Intern Med*. 2021;181(8):1065-1070. doi:10.1001/jamainternmed.2021.2626
22. Steffey MA, Griffon DJ, Risselada M, et al. A narrative review of the physiology and health effects of burnout associated with veterinarian-pertinent occupational stressors. *Front Vet Sci*. 2023;10. doi:10.3389/fvets.2023.1184525

23. Steffey MA, Griffon DJ, Risselada M, et al. Veterinarian burnout demographics and organizational impacts: a narrative review. *Front Vet Sci*. 2023;10. doi:10.3389/fvets.2023.1184526
24. Olson KD, Meeker D, Troup M, et al. Use of Ambient AI Scribes to Reduce Administrative Burden and Professional Burnout. *JAMA Netw Open*. 2025;8(10):e2534976. doi:10.1001/jamanetworkopen.2025.34976
25. Hur B, Hardefeldt LY, Verspoor KM, Baldwin T, Gilkerson JR. Evaluating the dose, indication and agreement with guidelines of antimicrobial use in companion animal practice with natural language processing. *JAC-Antimicrobial Resistance*. 2022;4(1):dlab194. doi:10.1093/jacamr/dlab194
26. Hur B, Verspoor KM, Baldwin T, et al. Using natural language processing and patient journey clustering for temporal phenotyping of antimicrobial therapies for cat bite abscesses. *Preventive Veterinary Medicine*. 2024;223:106112. doi:https://doi.org/10.1016/j.prevetmed.2023.106112
27. Hur B, Hardefeldt L, Verspoor K, Baldwin T, Gilkerson J. Overcoming challenges in extracting prescribing habits from veterinary clinics using big data and deep learning. *Australian Veterinary Journal*. 2022;100(5):220-222. doi:10.1111/avj.13145
28. Appleby RB, Basran PS. Artificial intelligence in veterinary medicine. *Journal of the American Veterinary Medical Association*. 2022;260(8):819-824. doi:10.2460/javma.22.03.0093
29. Xiao S, Dhand NK, Wang Z, et al. Review of applications of deep learning in veterinary diagnostics and animal health. *Front Vet Sci*. 2025;12. doi:10.3389/fvets.2025.1511522
30. Rose C, Chen JH. Learning from the EHR to implement AI in healthcare. *npj Digit Med*. 2024;7(1):330. doi:10.1038/s41746-024-01340-0
31. Hur B, Hardefeldt LY, Verspoor K, Baldwin T, Gilkerson JR. Using natural language processing and VetCompass to understand antimicrobial usage patterns in Australia. *Australian Veterinary Journal*. 2019;97(8):298-300. doi:10.1111/avj.12836
32. McGreevy P, Thomson P, Dhand NK, et al. VetCompass Australia: A National Big Data Collection System for Veterinary Science. Published online 2017:15.
33. Jones PH, Radford AD, Noble PJM, et al. SAVSNET: Collating Veterinary Electronic Health Records for Research and Surveillance. *Online Journal of Public Health Informatics*. 2016;8(1):1. doi:10.5210/ojphi.v8i1.6543
34. Creevy KE, Akey JM, Kaeberlein M, Promislow DEL. An open science study of ageing in companion dogs. *Nature*. 2022;602(7895):51-57. doi:10.1038/s41586-021-04282-9
35. Coghlan S, Quinn T. Ethics of using artificial intelligence (AI) in veterinary medicine. *AI & Soc*. 2024;39(5):2337-2348. doi:10.1007/s00146-023-01686-1
36. Health C for D and R. Artificial Intelligence-Enabled Medical Devices. *FDA*. Published online December 5, 2025. Accessed December 7, 2025. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices
37. Hur B, Baldwin T, Verspoor K, Hardefeldt L, Gilkerson J. Domain Adaptation and Instance Selection for Disease Syndrome Classification over Veterinary Clinical Notes. In: *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. Association for Computational Linguistics; 2020:156-166. Accessed July 8, 2021. https://www.aclweb.org/anthology/2020.bionlp-1.17
38. Elangovan A, He J, Li Y, Verspoor K. Principles from Clinical Research for NLP Model Generalization. *arXiv*. Preprint posted online April 2, 2024:arXiv:2311.03663. doi:10.48550/arXiv.2311.03663
39. Finlayson SG, Subbaswamy A, Singh K, et al. The Clinician and Dataset Shift in Artificial Intelligence. *New England Journal of Medicine*. 2021;385(3):283-286. doi:10.1056/NEJMc2104626
40. Shreffler J, Huecker MR. Diagnostic Testing Accuracy: Sensitivity, Specificity, Predictive Values and Likelihood Ratios. In: *StatPearls*. StatPearls Publishing; 2025. Accessed January 2, 2026. http://www.ncbi.nlm.nih.gov/books/NBK557491/
41. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432
42. Monaghan TF, Rahman SN, Agudelo CW, et al. Foundational Statistical Principles in Medical Research: Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value. *Medicina (Kaunas)*.

2021;57(5):503. doi:10.3390/medicina57050503

43. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv*. Preprint posted online October 11, 2020:arXiv:2010.16061. doi:10.48550/arXiv.2010.16061
44. Šimundić AM. Measures of Diagnostic Accuracy: Basic Definitions. *EJIFCC*. 2009;19(4):203-211.
45. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. *J Thorac Oncol*. 2010;5(9):1315-1316. doi:10.1097/JTO.0b013e3181ec173d
46. Carrington AM, Fieguth PW, Qazi H, et al. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. *BMC Med Inform Decis Mak*. 2020;20(1):4. doi:10.1186/s12911-019-1014-6
47. Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of Modern Neural Networks. *arXiv*. Preprint posted online August 3, 2017:arXiv:1706.04599. doi:10.48550/arXiv.1706.04599
48. Geng J, Cai F, Wang Y, Koeppl H, Nakov P, Gurevych I. A Survey of Confidence Estimation and Calibration in Large Language Models. In: Duh K, Gomez H, Bethard S, eds. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics; 2024:6577-6595. doi:10.18653/v1/2024.naacl-long.366
49. Jeni LA, Cohn JF, De La Torre F. Facing Imbalanced Data Recommendations for the Use of Performance Metrics. *Int Conf Affect Comput Intell Interact Workshops*. 2013;2013:245-251. doi:10.1109/ACII.2013.47
50. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22(3):276-282.
51. Carletta J. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Comput Linguist*. 1996;22(2):249-254.
52. Elangovan A, Liu L, Xu L, Bodapati SB, Roth D. ConSiDERS-The-Human Evaluation Framework: Rethinking Human Evaluation for Generative Large Language Models. In: Ku LW, Martins A, Srikumar V, eds. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics; 2024:1137-1160. doi:10.18653/v1/2024.acl-long.63
53. Poplin R, Varadarajan AV, Katy B, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*. 2018;2(3):158-164. doi:10.1038/s41551-018-0195-0
54. McLeod GA, Stanley EAM, Rosenal T, Forkert ND. Distinct visual biases affect humans and artificial intelligence in medical imaging diagnoses. *npj Digit Med*. 2025;9(1):62. doi:10.1038/s41746-025-02226-5
55. Yu AC, Mohajer B, Eng J. External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review. *Radiology: Artificial Intelligence*. 2022;4(3):e210064. doi:10.1148/ryai.210064
56. Elangovan A, He J, Verspoor K. Memorization vs. Generalization : Quantifying Data Leakage in NLP Performance Evaluation. In: Merlo P, Tiedemann J, Tsarfaty R, eds. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics; 2021:1325-1335. doi:10.18653/v1/2021.eacl-main.113
57. Wang X, Antoniades A, Elazar Y, et al. GENERALIZATION V.S. MEMORIZATION: TRACING LANGUAGE MODELS' CAPABILITIES BACK TO PRETRAINING DATA. Published online 2025.
58. Asgari E, Montaña-Brown N, Dubois M, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digit Med*. 2025;8(1):274. doi:10.1038/s41746-025-01670-7
59. Ndiaye YS, Cramton P, Chernev C, Ockenfels A, Schwarz T. Comparison of radiological interpretation made by veterinary radiologists and state-of-the-art commercial AI software for canine and feline radiographic studies. *Front Vet Sci*. 2025;12. doi:10.3389/fvets.2025.1502790