



Visualizing Uncertainty in AI-Assisted Canine Acute-Pain Assessment: Dog-Grouped Risk Modeling and Information-Equivalent Interface Prototypes

Chole Zhao¹, Evelyn Liu², Zijian Cao^{*3}

^{1,2}Human-Computer Interaction, Georgia Institute of Technology, GT, USA

³Electrical and Computer Engineering, University of Wisconsin-Milwaukee, WI, USA

Email Address: zijiannc@gmail.com ^(*)

Abstract. Canine acute-pain assessment combines behavioral observation, clinical context, and professional judgment, yet a model probability can be mistaken for a clinical pain score when the two lack a clear visual hierarchy. This study linked retrospective prediction-model findings to three information-equivalent interface prototypes. The structured database contained 594 observations from 184 dogs; 114 complete acute-pain observations from 65 dogs formed the primary cohort, with the short-form Glasgow Composite Measure Pain Scale (CMPS-SF) $\geq 6/24$ as the reference outcome. A hypothetical pre-score feature set excluded direct score components, although the database did not establish recording order. Four calibrated models were evaluated with five repetitions of four-fold dog-grouped nested cross-validation. Logistic regression achieved an area under the receiver operating characteristic curve (AUROC) of .627 (95% CI [.474, .755]), an area under the precision-recall curve (AUPRC) of .476 [.317, .660], and Brier score .210 [.174, .249]. Sensitivity at 0.50 was .189, missing 30 of 37 threshold-positive observations. Full-development bootstrap fits supplied 90% model-stability ranges, distinct from patient-outcome intervals. Exploratory routing reduced Brier score but also reduced AUROC and changed retained event prevalence. An auxiliary vocabulary analysis identified 169 descriptors in 52 grouped rows. A separate transit-display analysis of 12,331 records from 299 participant identifiers did not establish an average payoff or duration advantage over a no-uncertainty display. The numeric dashboard, visual scale, and progressive evidence card retained identical information while varying hierarchy and disclosure. Their structural audit establishes content equivalence, not clinical effectiveness. The design contribution connects model limitations to visible range definitions, separate clinical-reference labels, missing-context cues, and a persistent structured-assessment action.

Keywords. canine acute pain; uncertainty visualization; clinical decision support; probability calibration; information design

INTRODUCTION

Pain assessment in dogs is a problem of inference under uncertainty. Dogs cannot provide verbal self-report, physiological indicators are nonspecific, and behavioral expression varies with temperament, environment, disease, and prior treatment (Hansen, 2003). Clinical guidance therefore emphasizes repeated, multimodal assessment rather than reliance on a single observation or number (Mathews et al., 2014; Monteiro et al., 2023). A decision-support system can organize this evidence, but its visual form determines whether a probability is interpreted as a tentative model estimate, a measured pain score, or an instruction to treat. That distinction is especially important for acute pain. The CMPS-SF is a behavior-based instrument with an established analgesic intervention point, a visual analogue scale (VAS) is an observer rating, and a predictive probability is a model output. Placing the three together can obscure their provenance, while a simplified risk label can create a categorical boundary that neither the scale nor the model validates. The design task is to make the signal, uncertainty, evidence, and required clinical action legible together.

Clinical artificial intelligence creates a second tension: users need enough explanation to judge limitations, but each element competes for attention. Appropriate reliance, rather than

maximal trust, is the goal (Lee & See, 2004). Confidence displays can improve trust calibration, yet explanation and confidence do not automatically improve joint decisions (Zhang et al., 2020). Users therefore need to understand what the uncertainty display means for the assessment at hand.

The study combined computational and information-design evaluation. A structured canine pain database was audited, an acute-pain cohort was constructed, and four calibrated models were tested with dog-grouped nested cross-validation. Contemporaneous CMPS-SF components, VAS, treatment narratives, and target-proximal examination text were excluded from prediction. Robustness analyses examined feature sets, cohort definitions, subgroups, and case-level stability. An auxiliary behavioral file characterized candidate vocabulary, and a separate analysis of transit decisions examined how uncertainty displays related to payoff and response duration (Fernandes et al., 2018).

The model output was organized into a numeric dashboard, continuous risk scale, and progressively disclosed evidence card. A common content schema compared first-view density, disclosure depth, uncertainty encoding, and access to context. The design analysis examined how these arrangements could direct attention or obscure limitations; it did not measure decision time, workload, preference, or clinical effectiveness. Three questions guided the work: how well the retrospectively defined context estimates the CMPS-SF intervention outcome; how performance changes across robustness conditions; and how identical case information can be organized without conflating prediction with the clinical reference. The design contribution is a set of interface requirements tied directly to measured model limitations, with separate evaluation of predictive performance and display structure.

LITERATURE REVIEW

Acute-pain measurement in dogs

Canine pain scores depend on method and observer, so VAS is not an objective ground truth (Holton et al., 1998). Behavior-based development identified posture, mobility, vocalization, attention to a painful area, response to touch, activity, and demeanor (Holton et al., 2001). Formal scaling methods then validated the Glasgow composite approach beyond an unstructured overall impression (Morton et al., 2005). The CMPS-SF contains six behavioral categories. A complete score of at least 6/24 indicates that analgesic intervention should be considered; 5/20 applies only when mobility cannot be assessed (Reid et al., 2007). Because the database did not explain missing mobility values, the primary outcome required all six items. This also respects guidance that acute and chronic instruments serve different purposes (Monteiro et al., 2023).

Pain scores remain part of a broader, repeated examination (Mathews et al., 2014). The proposed interface therefore distinguishes observed context, unavailable information, and a model estimate. Its intended placement before structured scoring is a design assumption rather than an observed sequence in the database. A completed CMPS-SF score is withheld from this view because it is the reference outcome that the model estimates.

Risk and uncertainty as visual information

Clinical uncertainty includes probability, ambiguity from incomplete information, and complexity (Han et al., 2011). A stability interval, a missing postoperative-time field, and a dense narrative represent these different problems. One “confidence” percentage erases the distinction. Risk format must fit the task: graphics are not uniformly superior, and preference is not comprehension (Ancker et al., 2006). Pairing numeric precision with visual magnitude is more defensible than vague labels alone (Lipkus, 2007). The event and denominator should be explicit,

uncertainty should accompany the point estimate, and unsupported bands should be avoided (Spiegelhalter, 2017). Interval encodings also interact with prior beliefs and task goals (Padilla et al., 2021). These principles motivate a continuous probability display without LOW, MEDIUM, or HIGH zones. The CMPS-SF intervention point belongs to the clinical score and does not define a cutoff on a model-probability axis. A visual scale can show the estimate and its resampling range while an adjacent instruction directs the user to complete the structured assessment.

Information hierarchy and progressive disclosure

Shneiderman's (1996) overview-first principle and Nielsen's (2006) progressive disclosure approach motivate an initial view of essential information with further detail available on request. In the progressive card, the event, estimate, model-stability range, assessment instruction, patient and encounter summaries, availability, and concise model limitation remain visible. Association cues, postoperative detail, and reference-status explanation are placed in the expanded layer. This arrangement preserves the limitation message while deferring secondary detail. Reducing visible detail is consistent with the concern for limited processing capacity in cognitive-load theory (Sweller, 1988), but a smaller first view does not demonstrate lower workload. Disclosure adds an interaction cost and can make limitations less likely to be encountered. The audit therefore distinguishes availability of information from its immediate visibility and makes no claim that either layout is easier to use.

Explanation and appropriate reliance

Automation should support calibrated judgment, not maximize agreement (Lee & See, 2004). Confidence can distinguish stronger from weaker advice, but explanations disconnected from the task may add little (Zhang et al., 2020). Cognitive forcing can reduce overreliance at the cost of effort (Buçinca et al., 2021), so a review state must trigger meaningful inspection. Clinicians need explanations connected to context, action, data quality, and limitations (Tonekaboni et al., 2019), not merely faithful feature plots (Amann et al., 2020). Interpretable models are preferable in high-stakes settings when complex alternatives add no clear advantage (Rudin, 2019). The evidence cards therefore use plain-language association directions and avoid causal wording.

Evaluation of small clinical prediction datasets

Small datasets are vulnerable to optimistic estimates when selection and validation share folds. Nested cross-validation reserves outer folds for estimation (Varma & Simon, 2006), and all records from one dog must remain together. Performance should cover discrimination, calibration, and probabilistic error (Steyerberg et al., 2010; Zhong et al., 2023), especially when probabilities will be displayed clinically (Van Calster et al., 2019). Logistic regression provides an auditable linear function; random forests aggregate bootstrapped trees (Breiman, 2001); extremely randomized trees increase split randomization (Geurts et al., 2006); and gradient boosting performs sequential error correction (Friedman, 2001). Brier score measures probabilistic squared error (Brier, 1950), while dog-cluster bootstrap intervals preserve within-dog dependence (Efron & Tibshirani, 1993).

RESEARCH METHODS

Study design and data sources

The study combined retrospective prediction-model evaluation, secondary analyses of observer vocabulary and uncertainty-display decisions, and a structured interface-content audit.

Figure 1 summarizes the clinical-data pathway to interface specification; the transit analysis provides separate evidence about display-related behavior.

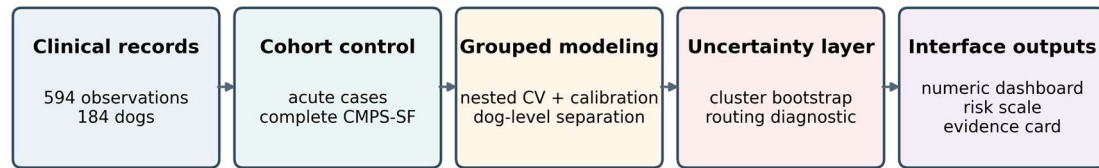


Figure 1. Study pipeline from data verification and cohort construction to dog-grouped modeling, exploratory uncertainty analysis, and interface specification.

The primary file was Dog pain database_V2.xlsx from Dog Pain Database: A Multidimensional Dataset for Investigating Canine Pain, version 2, released on 29 April 2025 (Bremhorst, 2025). The workbook contained one worksheet, 594 nonblank observations, 184 unique dog identifiers, and 76 imported columns: 74 populated data fields and two empty questionnaire-section headers. Supporting files from Zanusso et al. (2024) supplied a 20-dog clinical case table and an observer vocabulary containing 169 distinct Italian descriptors in 52 grouped rows. Transit decision records came from Fernandes et al. (2018). SHA-256 hashes identify the analyzed file versions (Table 1).

Table 1. Verified data inventory and file-level provenance.

File	Role	Contents	SHA-256
Dog pain database_V2.xlsx	Primary structured clinical dataset	594 records; 184 dogs; 74 populated fields; 76 imported columns	9a72e1532f5c8b0c 8200f75135a11bca e9bb60cf7f25e156 5483f0136329e82a
journal.pone.0305925.s003.docx	Auxiliary 20-dog clinical case table	20 records; 20 dogs; 6 fields	fc798afd0c076e27 2d3d7fae20a9b0d2 2718b97d284090fb 4441f3c9e3709926
journal.pone.0305925.s004.docx	Auxiliary observer-generated behavior vocabulary	52 grouped rows; 169 distinct Italian descriptors; 5 fields	2ff64720ad874bec 70832274ee8aeb8f 00704dc7d6b65cc 570e0186a809547f
final_trials.csv	Secondary transit uncertainty-display experiment	19,402 source records; 408 participant identifiers	51c661ea3fb30ec2121bd 86fffd8022bbe 392b8096304cf3825a69b a16a2731c

Cohort construction and reference outcome

Each of the six CMPS-SF item fields was parsed from the final parenthesized integer in the cell. An observation entered the primary cohort when all six values were parsable and the pain-type field contained an acute label but no chronic, subacute, or neuropathic token. This produced 114 observations from 65 dogs. The item values were summed to a 0–24 total, and the binary target equaled 1 when the total was at least 6 (Reid et al., 2007). The mobility-omitted 5/20 rule was not applied because the database did not identify why an item was absent. Figure 2 and Table 2 report every cohort transition.

Feature governance and target exclusion

The workbook did not timestamp the order in which individual fields were entered. The pre-score setting is therefore a hypothetical workflow represented retrospectively, not a verified

temporal sequence. CMPS-SF items, VAS, treatment narratives, orthopedic/neurologic examination narratives, and other target-proximal text were excluded. The eight retained predictors were weight, breed, sex, neuter status, patient type, situation, time since surgery, and anesthesia-recovery state. Some retained context may nevertheless have been recorded after clinical assessment.

Table 2. Construction of the complete canine acute-pain analysis cohorts.

Stage	Obs.	Dogs	CMPS-SF ≥ 6
All nonblank spreadsheet rows	594	184	Not defined
Six parsable CMPS-SF items	236	112	74
Clean acute label and six parsable items (primary)	114	65	37
Clean acute + somatic label (sensitivity)	75	46	24
Any exact acute label, including mixed temporal/mechanistic labels (sensitivity)	165	86	60
<i>Note. Threshold positive denotes CMPS-SF $\geq 6/24$; it is not defined before all six items are available.</i>			

Weight was parsed from the first numeric token, preserving decimal points and commas; cells with multiple numeric tokens received an ambiguous-weight flag in the ablation. Sex abbreviations and full labels were harmonized. Numeric values were median-imputed and standardized within training folds. Categorical values used an explicit unavailable level, and levels occurring fewer than three times in a training partition were grouped. Learned preprocessing parameters were fitted only on training partitions.

Analytic cohort construction

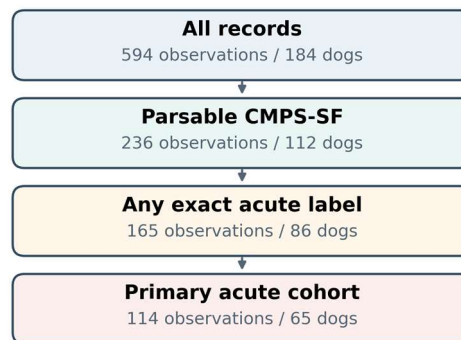


Figure 2. Cohort construction. The primary cohort retained complete CMPS-SF observations with a clean acute-pain label; strict somatic and broad acute definitions were analyzed separately.

Dog identifier was used only for grouping. CMPS-SF items defined the outcome. VAS was retained for descriptive analysis but excluded from prediction because it is a contemporaneous pain judgment. Medication and anesthesia narratives, orthopedic/neurologic examination narratives, and other target-proximal text were excluded. The structured anesthesia-recovery field, including entries that described ongoing sedation, remained a primary encounter-context predictor. Pain type was used only for cohort selection. Pain-origin and diagnosis categories entered an ablation analysis but not the primary model (Table 3).

Models, tuning, and probability calibration

Four models were compared: L2-regularized logistic regression, random forest, extremely randomized trees, and histogram gradient boosting (Breiman, 2001; Friedman, 2001; Geurts et al., 2006). Candidate hyperparameters minimized mean uncalibrated inner-fold Brier score. Logistic C was selected from 0.1, 0.3, and 1, without class weighting. Both tree ensembles used 150 trees, square-root feature subsampling, minimum leaf size 3, and maximum depth 3 or 6.

Histogram boosting used 100 iterations, learning rate .03, seven leaves, minimum leaf size 12, no early stopping, and L2 regularization 1 or 3. Table 4 lists the modal selections.

Table 3. Feature governance for a hypothetical pre-assessment workflow evaluated retrospectively.

Field group	Role	Reason
Dog identifier	Grouping only	Prevents the same dog appearing in training and test folds
CMPS-SF six item fields	Outcome construction only	Direct components of the reference standard
Estimated severity (VAS)	Descriptive/reference annotation only	Direct contemporaneous pain rating; withheld from decision cards
Medication and anaesthesia narratives	Excluded	Contain treatment information that may follow clinical judgment
Orthopedic/neurologic examination narrative	Excluded	May explicitly encode pain response and creates target-proximal leakage
Birth-date field	Excluded	Values were not stable within dog, so age could not be derived reliably
Breed, weight, sex, neuter status	Model input	Retained for the hypothetical pre-score workflow; recording order unverified
Patient type, situation, time since surgery, anaesthesia recovery	Model input	Structured context retained for the hypothetical pre-score workflow; recovery entries may describe sedation
Pain type	Cohort selection only	Acute, chronic, subacute, and neuropathic tokens define analysis eligibility
Pain origin and diagnosis categories	Ablation input only	Fourteen indicators tested without entering the primary model
Pain duration and ambiguous-weight flag	Ablation input only	Two additional fields tested with the clinical-category feature set
Profile/context availability count	Final ablation only	Counts unavailable weight, patient type, situation, surgery time, and recovery state
<i>Note. CMPS-SF = short-form Glasgow Composite Measure Pain Scale; VAS = visual analogue scale.</i>		

Validation used five repetitions of four-fold stratified group cross-validation, with dog identifier as the group. Each outer training partition contained a separate three-fold dog-grouped inner search. After selection, sigmoid calibration was fitted to three-fold grouped out-of-fold predictions within that outer training partition, and the selected base estimator was refitted on the complete outer training partition. No outer test dog entered preprocessing, tuning, or calibration. An observation's final out-of-fold probability was the mean of its five repeat-specific test predictions. This nested design separated primary model selection from performance estimation (Varma & Simon, 2006).

Exploratory feature-set and cohort comparisons fixed the modal logistic hyperparameters selected by the primary analysis, then used five repetitions of four-fold dog-grouped prediction with training-partition sigmoid calibration. This held the specification constant across the sensitivity comparisons. Because the fixed parameters were informed by the full primary analysis, these secondary estimates are data-informed sensitivity comparisons rather than independently nested validation results.

Performance measures and intervals

Discrimination was measured with AUROC and average precision, reported as the AUPRC summary. Probabilistic accuracy used Brier score and log loss; calibration used expected

calibration error (ECE) with ten quantile-defined bins, dropping duplicate boundaries, and a reliability curve. Calibration intercept and slope were jointly estimated by unpenalized logistic regression on the logit of the out-of-fold probabilities. Classification measures at a fixed 0.50 cutoff were balanced accuracy, sensitivity, specificity, and F1. Because 0.50 was not a validated clinical action threshold, a rank-based analysis also measured sensitivity and precision in the highest probability quartile, comprising the ceiling of one-quarter of the observations. These measures distinguish discrimination from calibration (Steyerberg et al., 2010; Van Calster et al., 2019).

Table 4. Candidate models, selected modal parameters, validation, and calibration.

Model	Modal parameters	Validation and calibration
Logistic Regression	$C = 1$; no class weighting	5×4 dog-grouped outer folds; 3 grouped inner folds; Brier selection; sigmoid calibration
Random Forest	150 trees; depth 3; $\sqrt{p}$ features; leaf ≥ 3	5×4 dog-grouped outer folds; 3 grouped inner folds; Brier selection; sigmoid calibration
Extra Trees	150 trees; depth 3; $\sqrt{p}$ features; leaf ≥ 3	5×4 dog-grouped outer folds; 3 grouped inner folds; Brier selection; sigmoid calibration
Histogram Boost	100 iterations; $L_2 = 3$; rate .03; 7 leaves; leaf ≥ 12	5×4 dog-grouped outer folds; 3 grouped inner folds; Brier selection; sigmoid calibration
<i>Note. Preprocessing, hyperparameter selection, and sigmoid calibration in the primary nested evaluation were confined to outer training data. Selection minimized uncalibrated inner-fold Brier score; modal parameters summarize 20 outer folds.</i>		

Ninety-five percent intervals for comparative performance and calibration used 1,000 dog-cluster bootstrap samples of the labels and saved out-of-fold probabilities; all observations of a sampled dog remained together (Efron & Tibshirani, 1993). These intervals describe sampling variability conditional on the fitted predictions and do not repeat nested model selection. Subgroup AUROC intervals used 700 dog-cluster samples. Separately, the modal logistic pipeline was refitted with grouped sigmoid calibration in 200 dog-cluster bootstrap samples of the full primary cohort. A case's 5th and 95th predicted-probability percentiles formed its full-development 90% model-stability range. The range describes variation across refitted models. It is neither a patient-outcome confidence interval nor a confidence interval around the separately estimated out-of-fold probability, and it carries no prospective coverage guarantee.

Robustness analyses

Four nested feature sets tested the contribution of patient profile, encounter context, 14 pain-origin and diagnosis indicators plus pain duration and an ambiguous-weight flag, and an aggregate availability count. The count marked unavailable weight, patient type, situation, time since surgery, and anesthesia-recovery state. Cohort sensitivity analyses used a strict acute-plus-somatic definition and a broader exact-acute definition that retained mixed temporal or mechanistic labels. Subgroup analyses were calculated for sex, postoperative status, emergency/hospitalized status, and profile/context availability. These estimates remained descriptive because subgroup sizes were small.

A post hoc development-sample routing diagnostic ranked observations by 0.65 times full-development stability-range width plus 0.35 times base-2 binary entropy of the out-of-fold probability. Coverage levels from 100% to 50% retained the lowest-scoring observations and assigned the remainder to review. AUROC, Brier score, range width, entropy, and retained event prevalence were recomputed at each level. The weights were exploratory, and the range incorporated full-development fits. This diagnostic therefore cannot validate an operational abstention policy.

Behavioral vocabulary analysis

The auxiliary case table contained 10 dogs with acute pain and 10 healthy dogs (Zanusso et al., 2024). The vocabulary table contained 52 grouped rows, some listing several Italian adjectives. Each row's owner, veterinary-student, and veterinarian frequencies applied to every adjective in that row. Splitting the lists yielded 169 distinct descriptors; Italian terms were kept distinct even when their English translations overlapped. Terms shared across all three groups and pairwise Jaccard similarity were calculated from each group's nonzero term set. Frequencies characterized candidate wording for later comprehension testing. The descriptors were not assigned to the structured cases, used as model-derived cues, or treated as independent diagnostic labels.

Interface construction and content audit

Three prototypes displayed the same modeled event, cross-validated point estimate, full-development 90% model-stability range, patient profile, encounter context, availability state, association cues, and instruction to complete CMPS-SF. The cues came from coefficients of the modal uncalibrated logistic specification fitted to the full primary cohort and describe sample associations rather than causal effects. Interface A used a single-layer numeric dashboard. Interface B placed the probability and range on a continuous horizontal axis with numeric range endpoints. Interface C placed seven content groups in the first view and three in an expanded layer (Nielsen, 2006; Shneiderman, 1996). A concise model limitation remained in its first view, while association cues, postoperative detail, and reference status were deferred. All designs labeled the range as variation across refitted models rather than an interval for patient outcome. No design used clinical risk bands or placed the CMPS-SF intervention point on the probability axis. Completed CMPS-SF and VAS values were withheld from the decision cards and appeared only in the retrospective figure annotation.

The audit applied one ten-group schema: estimate and source, stability range, action and review state, patient summary, encounter summary, postoperative context, availability, association cues, model limitation, and reference status. Fields were mapped to these groups, and a code assertion checked complete coverage across each design's visible and expanded layers. The rendered prototypes were also inspected for separation of model probability from the clinical reference, labeling of interval endpoints, and placement of the limitation message. This audit assessed content and representation, not whether clinicians noticed, understood, or appropriately acted on the information.

Secondary analysis of uncertainty-display decisions

A separate secondary analysis examined behavioral records from the transit decision experiment of Fernandes et al. (2018). The ten display conditions included point predictions without uncertainty, three textual predictive intervals, interval and density plots, a combined density–interval plot, two quantile dotplots, and a cumulative distribution display. Decision quality was measured by the expected payoff of the recorded choice divided by the maximum expected payoff for its scenario. Recorded response duration provided a complementary measure of interaction time. These outcomes were analyzed separately from canine prediction performance and the three interface prototypes.

The source file contained 19,402 records and 408 participant identifiers. Following the original analysis, the binary-choice scenario was excluded, leaving 13,887 records. Removing 1,476 exact duplicate records and 80 records associated with one identifier assigned to two remaining conditions yielded 12,331 records from 299 participant identifiers. Descriptive decision quality gave each participant equal weight; descriptive time was the median of participant-level median durations. Participant bootstrap intervals used 4,000 resamples. Adjusted

comparisons used weighted least squares, with each participant's records summing to unit weight and categorical terms for display format, scenario, presence of a point estimate, and trial order. Duration was log-transformed, and its coefficients were exponentiated as duration ratios. Confidence intervals used participant-cluster robust covariance. Comparisons with the no-uncertainty condition estimated average differences across the recorded trials; they did not estimate the original study's final-trial learning contrasts. Sensitivity analyses retained the original records, collapsed repeated participant–trial entries, excluded identifiers with conflicting stimuli or answers for a trial, and removed durations above the 98th percentile.

Reproducibility

The analyses ran in Python 3.12.14 with pandas 2.2.3, NumPy 2.3.5, SciPy 1.17.0, and scikit-learn 1.8.0. The clinical split and bootstrap streams used seed 20260822. The accompanying analysis package contains source data and checksums, executable code, fold-level selections, out-of-fold predictions, bootstrap arrays, tabular results, and vector and raster figures.

RESULTS AND FINDINGS

Dataset composition and acute-pain cohort

The workbook description and the file contents differed in unit counts: the analyzed worksheet contained 184 unique dog identifiers rather than 200. Among 594 observations, 236 observations from 112 dogs had all six parsable CMPS-SF items. Applying the clean acute definition left 114 observations from 65 dogs; 37 observations (32.5%) reached the 6/24 intervention point. The strict acute-plus-somatic cohort contained 75 observations from 46 dogs, and the broad exact-acute cohort contained 165 observations from 86 dogs.

Table 5. Characteristics of the primary acute-pain cohort.

Characteristic	Value	Detail
Observations	114	count
Unique dogs	65	count
Intervention-threshold observations	37	32.5%
CMPS-SF total	4.0	IQR 2.0-6.0
VAS	32.5	IQR 23.0-43.0; n=114
Weight (kg)	17.3	IQR 11.5-29.2; n=114
Female observations	50	43.9%
Post-surgery observations	38	33.3%
Emergency/hospitalized observations	66	57.9%
Median observations per dog	1.0	range 1-5

The primary cohort had a median CMPS-SF total of 4.0 (IQR 2.0–6.0) and a median VAS of 32.5 (IQR 23.0–43.0). Median weight was 17.3 kg, 50 observations (43.9%) were female, 38 (33.3%) followed surgery, and 66 (57.9%) were emergency or hospitalized encounters. Dogs contributed a median of one observation and a maximum of five (Table 5). The score distribution was concentrated below the intervention point but included observations across the observed 0–15 range (Figure 3).

Comparative predictive performance

The calibrated logistic model had the highest AUROC, .627 (95% CI [.474, .755]), while extra trees had the highest AUPRC, .492 [.339, .703]. Random forest, extra trees, and histogram boosting achieved AUROCs of .606, .595, and .413, respectively (Table 6; Figure 4). All four AUROC intervals included .50. These estimates do not establish reliable discrimination or an advantage for the more complex ensembles; histogram boosting showed particularly weak ranking in this sample.

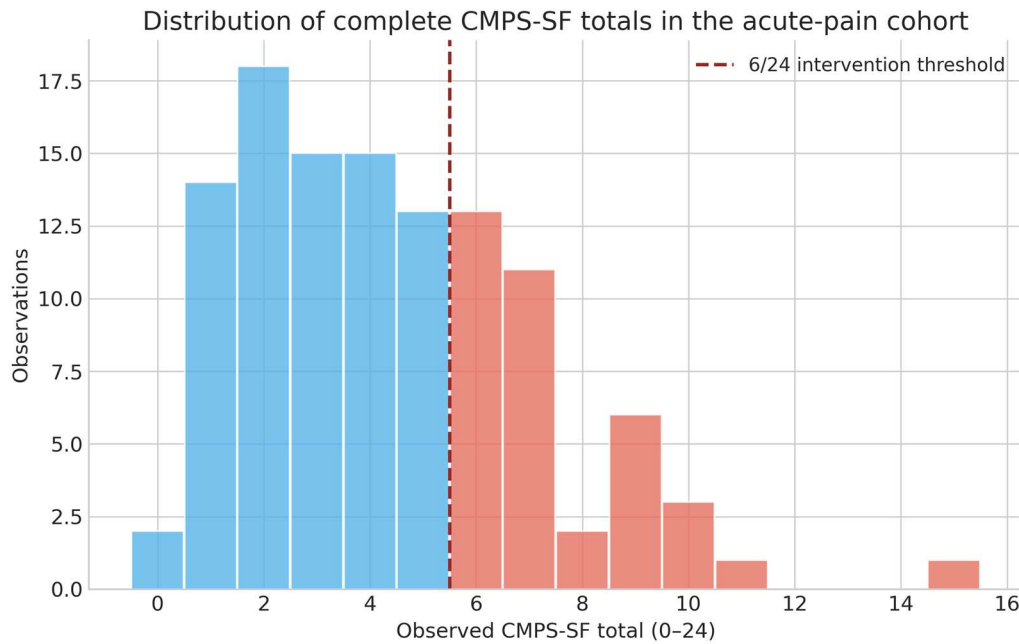


Figure 3. Distribution of complete CMPS-SF totals in the primary cohort. The dashed line marks the 6/24 analgesic intervention point.

At the fixed 0.50 probability cutoff, logistic regression had sensitivity .189 [.028, .390], specificity .961 [.908, 1.000], balanced accuracy .575 [.494, .670], and F1 .298 [.051, .520]. It identified seven of 37 threshold-positive observations and missed 30. Reviewing the highest probability quartile captured .378 [.222, .556] of positive observations, with precision .483 [.250, .741]. These results motivate evaluating review-oriented displays but do not validate clinical triage or an autonomous classification rule. The interface therefore retains the instruction to complete CMPS-SF regardless of the model estimate.

Table 6. Out-of-fold discrimination and classification performance with 95% dog-cluster bootstrap intervals.

Model	AUROC	AUPRC	Bal. acc.	Sens.	Spec.	F1	Top-Q sens.	Top-Q prec.
Logistic Regression	0.627 [0.474, 0.755]	0.476 [0.317, 0.660]	0.575 [0.494, 0.670]	0.189 [0.028, 0.390]	0.961 [0.908, 1.000]	0.298 [0.051, 0.520]	0.378 [0.222, 0.556]	0.483 [0.250, 0.741]
Random Forest	0.606 [0.445, 0.754]	0.462 [0.304, 0.673]	0.541 [0.480, 0.618]	0.108 [0.000, 0.256]	0.974 [0.931, 1.000]	0.186 [0.000, 0.392]	0.324 [0.182, 0.533]	0.414 [0.212, 0.692]
Extra Trees	0.595 [0.429, 0.756]	0.492 [0.339, 0.703]	0.528 [0.479, 0.594]	0.081 [0.000, 0.211]	0.974 [0.931, 1.000]	0.143 [0.000, 0.328]	0.432 [0.242, 0.609]	0.552 [0.290, 0.750]
Histogram Boost	0.413 [0.281, 0.568]	0.325 [0.190, 0.489]	0.500 [0.500, 0.500]	0.000 [0.000, 0.000]	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]	0.162 [0.023, 0.300]	0.207 [0.031, 0.407]

Note. Classification metrics use a fixed 0.50 cutoff only as a diagnostic operating point. Top-Q = highest quartile of predicted probabilities

Logistic regression had the lowest Brier score, .210 [.174, .249], and log loss, .614 [.533, .704]. Its ECE was .101 [.093, .235]. The jointly estimated calibration intercept was $-.180$ [-1.207 , .703] and slope .799 [$-.272$, 1.912] (Table 7). The reliability curves showed irregular bin-level agreement (Figure 5). The display consequently labels the estimate's source and model-

stability range while retaining the structured-assessment instruction; sigmoid calibration alone does not establish dependable clinical probabilities.

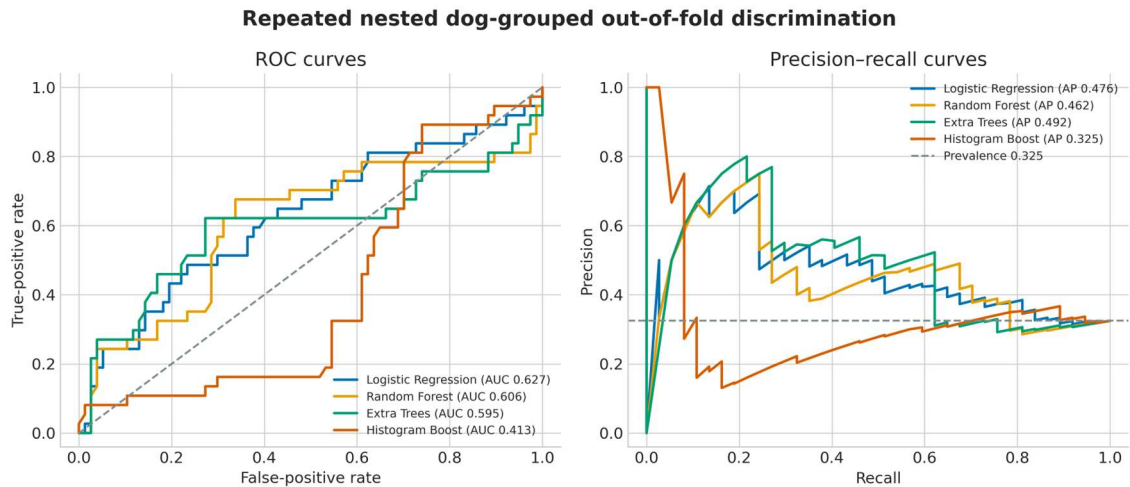


Figure 4. Out-of-fold receiver operating characteristic and precision-recall curves for the four calibrated models. Predictions are means across five repeated outer test assignments.

Feature-set and cohort sensitivity

Patient profile alone yielded AUROC .535, AUPRC .358, and Brier score .220. Adding encounter context produced .635, .497, and .208, respectively. The clinical-category feature set increased AUROC to .652 but yielded Brier .209; adding the availability count produced AUROC .650 and Brier .209 (Table 8). The categories did not improve probabilistic accuracy, and their small ranking difference does not establish incremental clinical value. The eight-variable specification provides a compact basis for comparing displays. These fixed-specification comparisons remain exploratory.

Table 7. Out-of-fold probabilistic accuracy and calibration with 95% dog-cluster bootstrap intervals.

Model	Brier	Log loss	ECE	Cal. intercept	Cal. slope
Logistic Regression	0.210 [0.174, 0.249]	0.614 [0.533, 0.704]	0.101 [0.093, 0.235]	-0.180 [-1.207, 0.703]	0.799 [-0.272, 1.912]
Random Forest	0.215 [0.177, 0.257]	0.624 [0.539, 0.716]	0.236 [0.146, 0.297]	-0.256 [-1.441, 0.871]	0.669 [-0.678, 2.278]
Extra Trees	0.212 [0.170, 0.256]	0.617 [0.526, 0.716]	0.171 [0.141, 0.287]	-0.124 [-1.263, 1.033]	0.836 [-0.455, 2.558]
Histogram Boost	0.224 [0.190, 0.262]	0.642 [0.570, 0.722]	0.229 [0.165, 0.320]	-1.502 [-4.599, 0.482]	-1.090 [-5.445, 1.653]

Note. ECE uses ten quantile-defined bins, dropping duplicate boundaries. Intercept and slope are jointly estimated from out-of-fold logits; ideal values are 0 and 1. Intervals condition on the saved predictions

Table 8. Feature-set ablation for calibrated logistic regression.

Feature set	Raw p	AUROC	AUPRC	Brier	Bal. acc.
Patient profile	4	0.535	0.358	0.220	0.494
Profile + encounter context	8	0.635	0.497	0.208	0.575
Profile + context + clinical categories	24	0.652	0.498	0.209	0.508
Full model + missingness count	25	0.650	0.498	0.209	0.508

With the fixed modal logistic specification, the primary cohort produced AUROC .635, AUPRC .497, and Brier .208. The strict acute-plus-somatic cohort produced AUROC .298, AUPRC .254, and Brier .225, whereas the broad exact-acute cohort produced AUROC .672,

AUPRC .479, and Brier .224 (Table 9). The marked deterioration under the strict definition shows sensitivity to cohort composition. The primary estimate should not be generalized across acute-pain subtypes or used to select a clinical operating threshold.

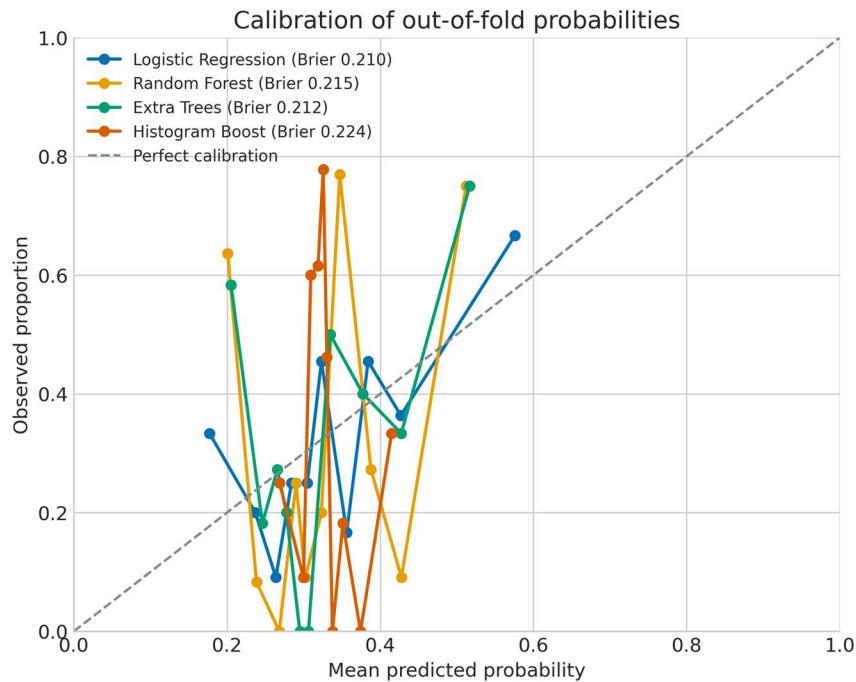


Figure 5. Out-of-fold calibration of the four models. Points summarize quantile bins, and the diagonal represents agreement between predicted and observed event rates.

Table 9. Sensitivity to acute-pain cohort definition.

Cohort definition	Obs.	Dogs	Positive	AUROC	AUPRC	Brier	Bal. acc.
Primary complete 24-point CMPS-SF	114	65	32.5%	0.635	0.497	0.208	0.575
Clean acute plus somatic	75	46	32.0%	0.298	0.254	0.225	0.500
Broad exact acute including mixed labels	165	86	36.4%	0.672	0.479	0.224	0.496

Note. These exploratory comparisons fixed the modal parameters from the primary analysis and repeated dog-grouped prediction with training-partition calibration. Their specification was informed by the primary data; they are not independent nested validations.

Subgroup behavior and post hoc review routing

Subgroup estimates were unstable. AUROC was .592 for female observations and .615 for male observations; the female interval ranged from .200 to .827. AUROC was .275 after surgery and .674 for other observations. Emergency or hospitalized encounters yielded .639, compared with .305 for other encounters. Complete availability of the five specified fields yielded .266, compared with .668 when at least one was unavailable (Table 10; Figure 6). These descriptive contrasts involve small and partly overlapping dog samples and do not establish demographic, missingness, or workflow effects.

At 80% retained coverage, assigning the 23 highest-composite observations to review decreased Brier score from .210 to .181 and mean stability-range width from .377 to .330, while AUROC fell from .627 to .395. At 50% coverage, Brier score was .142 and AUROC .274 (Table 11; Figure 7). Retained event prevalence also decreased from 32.5% at full coverage to 18.7% at 80% coverage and 12.3% at 50% coverage. Lower Brier scores thus accompanied a substantial

change in case mix and do not establish safer automation. The routing rule remains an exploratory development-sample diagnostic.

Table 10. Primary-model performance in prespecified subgroups.

Subgroup	Obs.	Dogs	Positive	AUROC [95% CI]	Brier
Female	50	29	28.0%	0.592 [0.200, 0.827]	0.195
Male	64	36	35.9%	0.615 [0.438, 0.784]	0.221
Post-surgery	38	26	21.1%	0.275 [0.040, 0.531]	0.190
Not post-surgery	76	42	38.2%	0.674 [0.514, 0.814]	0.219
Emergency/hospitalized	66	33	42.4%	0.639 [0.481, 0.792]	0.234
Other encounter	48	32	18.8%	0.305 [0.028, 0.542]	0.177
All five profile/context fields available	39	27	20.5%	0.266 [0.058, 0.497]	0.188
At least one profile/context field unavailable	75	41	38.7%	0.668 [0.516, 0.800]	0.221

Note. Subgroup intervals are descriptive and reflect small numbers of dogs and events.

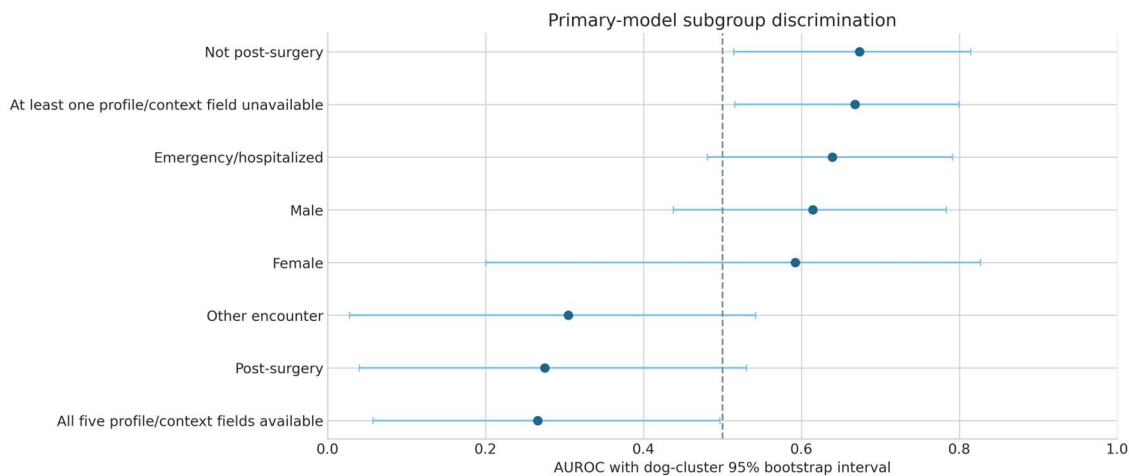


Figure 6. Primary-model AUROC by prespecified subgroup. Horizontal lines are 95% dog-cluster bootstrap intervals and reveal substantial sampling uncertainty.

Table 11. Post hoc review routing based on 65% full-development stability-range width and 35% predictive entropy.

Coverage	Retained	Reviewed	Positive (%)	AUROC	Brier	Mean range width	Mean entropy
100%	114	0	32.5%	0.627	0.210	0.377	0.880
90%	103	11	25.2%	0.513	0.201	0.351	0.870
80%	91	23	18.7%	0.395	0.181	0.330	0.861
70%	80	34	13.8%	0.294	0.159	0.321	0.849
60%	68	46	11.8%	0.296	0.144	0.314	0.834
50%	57	57	12.3%	0.274	0.142	0.312	0.816

Note. Coverage is the proportion retained in the post hoc diagnostic; the remainder is assigned to review. Range width and entropy are on 0–1 scales. Changing event prevalence prevents interpreting lower Brier scores alone as improved safety.

Observer vocabulary and interface evidence

The auxiliary material comprised 20 cases and 169 distinct descriptors in 52 grouped rows. Owners, students, and veterinarians used 123, 79, and 59 descriptors, respectively, with 29 shared by all three groups. Jaccard similarity was .320 for owners versus students, .255 for owners versus veterinarians, and .340 for students versus veterinarians. The most frequent expressions were “sore/in pain” (21 observers), “quiet” (20), “fearful” (17), “restless” (16), and “attentive” (16) (Table 12; Figure 8). These terms remain candidates for comprehension testing rather than

explanations of the model. The downloadable table contains 169 descriptors, although the source article reports 170 in its narrative (Zanusso et al., 2024).

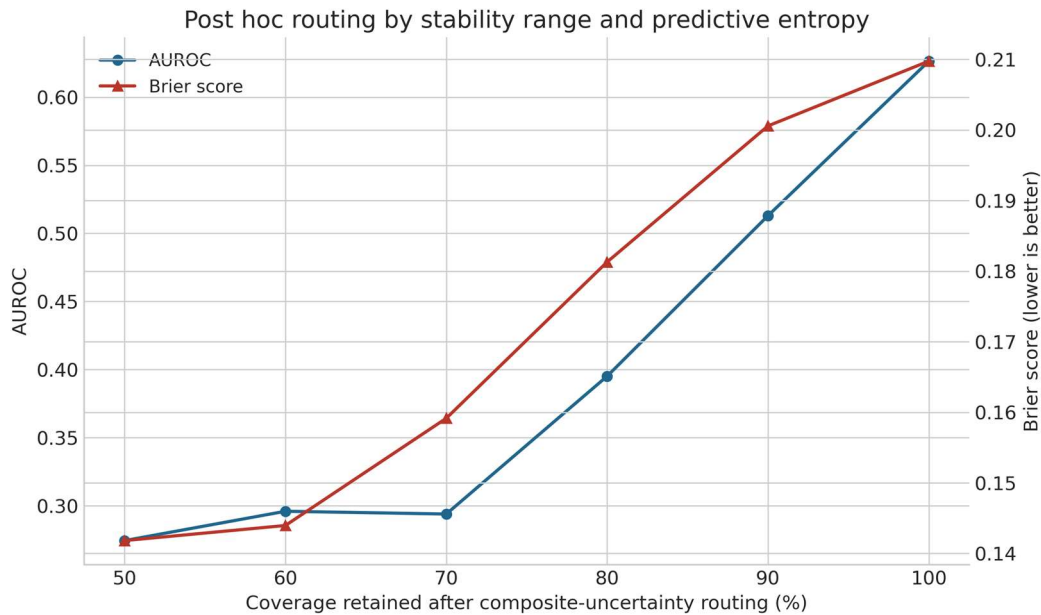


Figure 7. Post hoc routing diagnostic. Cases with the highest composite uncertainty score (65% full-development stability-range width and 35% predictive entropy) were progressively routed to review; lower Brier scores were accompanied by lower AUROC.

Information-equivalent interface prototypes

The prototypes provided access to the same modeled event, probability, stability range, profile, context, availability, association cues, and assessment instruction (Figure 9). The numeric dashboard and visual risk scale placed all ten schema groups in one layer. The progressive card placed seven in the first view and three in an expanded layer (Table 13). These counts describe semantic groups rather than word count, screen area, cognitive load, or usability. Model limitations and range meaning remained visible in every first view.

Table 12. Ten most frequent observer-generated behavioral descriptors.

Descriptor	Observers	Owners	Students	Veterinarians
Dolorante (sore/in pain)	21	6	7	8
Tranquillo (quiet)	20	8	5	7
Impaurito (fearful)	17	6	6	5
Agitato (restless)	16	5	4	7
Attento (attentive)	16	4	7	5
Abbattuto (depressed)	13	2	5	6
Nervoso (nervous)	12	2	3	7
Timoroso (timorous)	11	2	6	3
Curioso (curious)	11	2	5	4
Disorientato (disoriented)	10	3	3	4

From figure 9, it can be seen, that Dog 135, observation 1 has an out-of-fold estimate of 69.1% and a full-development 90% model-stability range of 15.5%–90.9%. The range reflects model refits, not an individual patient's outcome interval. Full-sample coefficient directions are separate sample associations. The retrospective CMPS-SF reference is 9/24 and VAS is 34; completed outcomes are withheld from the decision cards.

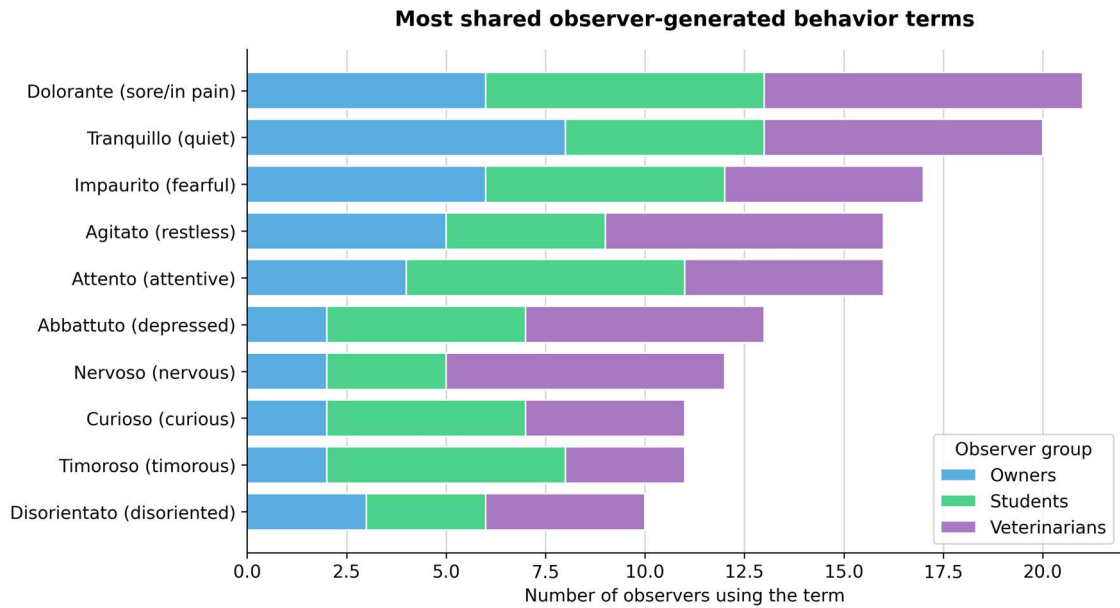


Figure 8. Most frequently used observer-generated behavioral descriptors. Frequencies aggregate owners, veterinary students, and veterinarians in the auxiliary vocabulary file.

Uncertainty displays for one retrospective case

Same information across designs • Probability and clinical score use separate representations

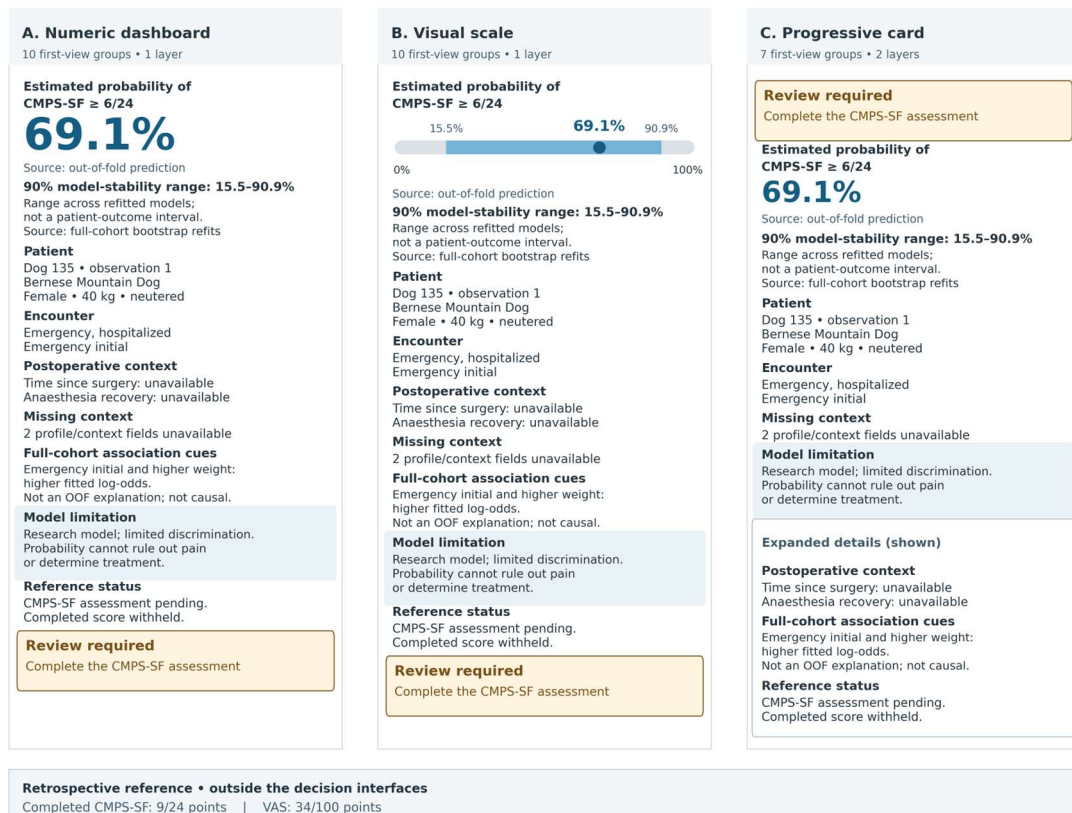


Figure 9. Information-equivalent prototypes for a hypothetical pre-assessment workflow.

The clinical and model quantities were visually separated. A continuous model-probability axis describes the estimated frequency of a CMPS-SF threshold event; the intervention point belongs to the subsequent structured score. Numeric endpoints and range bands describe model-

refit variability, while missing-field labels describe unavailable context. This separation follows Han et al.'s (2011) distinction between forms of uncertainty and makes each quantity's source explicit. It does not establish that users understand those distinctions.

Table 13. Prespecified content-group and disclosure-layer audit of the three interface designs.

Interface	First-view groups	Layers	Uncertainty	Clinical reference and action	Full context
A: Numeric dashboard	10	1	Numeric range with explicit model-refit definition	Complete CMPS-SF; no probability cutoff	Yes
B: Visual scale	10	1	Range band and numeric endpoints with model-refit definition	CMPS-SF criterion separate from probability axis	Yes
C: Progressive evidence card	7	2	Numeric range and definition retained in first view	Complete CMPS-SF; completed score withheld	Yes
<i>Note. The same ten groups are counted in each specification. A and B display all ten together; C retains seven in its first view and exposes postoperative context, association cues, and reference status in a second layer. Completed outcomes are excluded from all decision cards. Counts describe structure and do not measure visual complexity, workload, or clinical performance.</i>					

The numeric dashboard, visual scale, and progressive evidence card differ in hierarchy and disclosure while preserving access to the same information. The audit does not establish that any design improves comprehension, workload, or decision quality. C reduces the number of first-view content groups, but deferred information may receive less attention. All three remain candidates for clinician evaluation. Table 14 connects the empirical findings and design constraints to questions for that evaluation.

Table 14. Connections between computational findings and interface-design requirements.

Empirical limitation	Interface decision	Question for user evaluation
Weak, uncertain discrimination and low sensitivity at 0.50	Keep probability continuous; retain completion of CMPS-SF as the primary action regardless of its value	Does a low model estimate cause inappropriate reassurance or delayed assessment?
Model predictions vary across dog-cluster bootstrap refits	Show numeric range endpoints with a plain-language explanation of model-refit variability	Can users distinguish model stability from an individual patient's outcome uncertainty?
Missing context and heterogeneous encounter information	Label unavailable fields and group observed context separately from model estimates	Can users identify missing evidence without treating it as a confidence score?
No clinically validated probability cutoff	Keep the CMPS-SF intervention point off the model-probability axis	Can users distinguish the probability of the event from the clinical score and intervention point?
Lower Brier scores but lower AUROC in selected subsets	Avoid a validated-safe or automated-clearance label for retained cases	Do review states encourage examination without implying safety of unflagged cases?
Ten content groups with alternative disclosure arrangements	Keep limitation and range meaning visible; defer secondary detail in C	Does disclosure reduce reading burden while preserving awareness of limitations?

Secondary evidence from uncertainty-display decisions

Table 15 summarizes the ten conditions, and Figure 10 shows the adjusted comparisons. The 50-dot display had the highest descriptive mean payoff ratio (0.941), compared with 0.930 for the no-uncertainty condition; participant-level median durations were 9.19 and 9.96 seconds, respectively. The descriptive range across formats was 0.904–0.941 for mean payoff ratio and

7.68–11.50 seconds for median duration. After adjustment, the 50-dot payoff difference was 1.00 percentage point (95% CI [-1.71, 3.70]) and its duration ratio was 1.10 [0.84, 1.42]. The cumulative distribution display had a payoff difference of 0.69 percentage point [-1.93, 3.30] and duration ratio of 1.31 [0.95, 1.79]. All nine display comparisons included the null value for both outcomes, so the analysis did not establish an average advantage over the no-uncertainty condition. The absence of a detected difference does not establish equivalence. The 50-dot payoff estimate remained between 0.94 and 1.42 percentage points across sensitivity analyses, with every corresponding interval including zero.

Table 15. Decision payoff and response duration in the secondary transit analysis.

Display	Partici -pants	Records	Mean payoff / optimum	Median time (s)	Adjusted payoff difference, pp (95% CI)	Adjusted time ratio (95% CI)
No uncertainty	22	911	0.930	9.96	Reference	Reference
Text 60%	32	1295	0.914	7.68	-0.66 [-3.86, 2.54]	0.85 [0.65, 1.12]
Text 85%	25	1016	0.906	7.68	-1.99 [-5.50, 1.53]	0.92 [0.72, 1.17]
Text 99%	37	1601	0.924	9.29	-0.49 [-3.54, 2.56]	1.01 [0.79, 1.28]
Interval plot	20	796	0.904	9.63	-3.10 [-6.29, 0.09]	1.11 [0.84, 1.45]
Density plot	57	2323	0.923	10.98	-0.21 [-2.59, 2.16]	1.23 [0.98, 1.55]
Density + interval	26	1064	0.909	11.38	-1.60 [-5.02, 1.82]	1.23 [0.95, 1.58]
20-dot quantile plot	28	1139	0.925	11.33	-0.02 [-2.78, 2.75]	1.22 [0.93, 1.59]
50-dot quantile plot	23	937	0.941	9.19	1.00 [-1.71, 3.70]	1.10 [0.84, 1.42]
Cumulative distribution	29	1249	0.934	11.50	0.69 [-1.93, 3.30]	1.31 [0.95, 1.79]

Note. Mean payoff is the unweighted mean of participant-specific mean expected/optimal payoff ratios. Median time is the median of participant-specific median durations. Adjusted comparisons use the no-uncertainty condition as the reference and control for scenario, point-estimate variant, and trial order with equal total weight per participant. Confidence intervals are pointwise 95% participant-cluster intervals. Response duration is not a measure of subjective workload. The displayed counts describe records, because nonidentical repeated entries for some nominal trials remain in the primary analysis

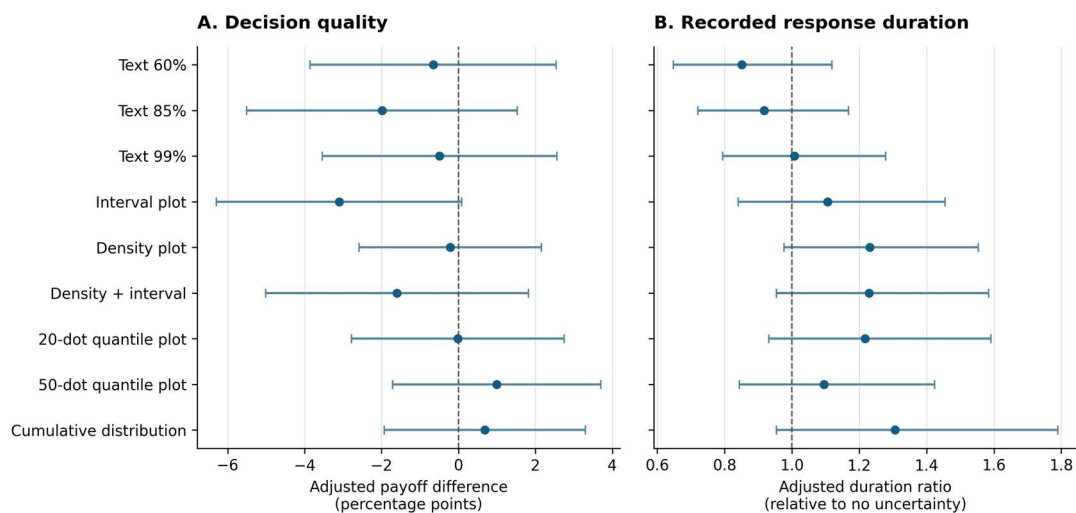


Figure 10. Adjusted differences between uncertainty-display conditions and the no-uncertainty condition in the secondary transit analysis.

Points on Figure 10 show participant-equally-weighted estimates, and bars show 95% participant-cluster confidence intervals, adjusting for scenario, point-estimate variant, and trial

order. Panel A reports percentage-point differences in expected payoff divided by optimal expected payoff; positive values indicate higher payoff. Panel B reports exponentiated differences in log duration; values above one indicate longer recorded response duration. The dashed lines mark the respective null values. The sample comprises 12,331 records from 299 participant identifiers.

DISCUSSION

From model limitations to interface decisions

The graphic-design contribution connects the limitations of the model to the information emphasized by the interface. Weak and uncertain discrimination and poor sensitivity at the diagnostic 0.50 cutoff argue against a reassuring negative state or an automatic treatment instruction. Accordingly, the probability remains continuous, the action is to complete CMPS-SF, and the clinical intervention point is kept separate from the probability axis. The stability range is labeled by its resampling source because it does not describe a patient's outcome or pain-score interval. Missing context receives a separate label because unavailable information and variability across model fits are different problems (Han et al., 2011). These choices translate probabilistic limitations into visible design constraints; they do not demonstrate that the resulting display is safe or effective.

Hierarchy, grouping, and uncertainty salience

Figure 9 illustrates three competing emphases. In A, the large percentage creates a clear entry point, but its typographic prominence may anchor attention before the smaller range and caveat are read. B places the point and range on one continuous axis, making their relative extent visible; numeric endpoints preserve precise reading without implying clinical risk bands. C places the assessment instruction above the probability and groups supporting material in a separate layer. Its first view is less dense in content groups, but those counts do not measure perceptual complexity. Across the designs, patient details and encounter context are grouped apart from model quantities to preserve provenance. The event label, probability units, and separate reference annotation reduce an evident source of semantic confusion: a model percentage is not a CMPS-SF score. Whether users notice and retain that distinction remains an empirical question.

Relation to uncertainty visualization and progressive disclosure

Prior work cautions that graphics are not uniformly better than numbers and that preferred formats need not yield better understanding (Ancker et al., 2006; Lipkus, 2007). The prototypes therefore provide complementary numeric and spatial descriptions rather than treating a graphic range as self-explanatory. Uncertainty visualization depends on the decision task and the quantity being encoded (Padilla et al., 2021; Spiegelhalter, 2017); a model-refit range requires a different explanation from a patient-outcome interval. In a study of clinical practitioners, visualization format affected uncertainty judgments without establishing a direct improvement in decision quality (Frans et al., 2023). Progressive disclosure follows overview-first design principles (Nielsen, 2006; Shneiderman, 1996), but the present audit supports only a structural comparison. Deferring detail can reduce the amount presented at once while also reducing the likelihood that it is inspected. C therefore retains the range explanation and concise limitation in its first view and defers association cues and additional context. This is a testable design rationale, not evidence of lower workload or superiority.

Appropriate reliance and accessible presentation

The review-oriented action follows the goal of appropriate reliance rather than maximal trust (Lee & See, 2004). Confidence and explanation can affect trust without improving joint decisions (Zhang et al., 2020), and requiring additional deliberation may reduce overreliance while increasing effort (Bućinca et al., 2021). Association cues are consequently framed as sample relationships, not treatment explanations. Basic accessibility also affects whether limitations can be perceived: probability and range values should remain available as text, information should not depend on color alone, labels should retain legibility when enlarged, and foreground-background contrast should meet the relevant WCAG 2.2 criteria (World Wide Web Consortium, 2024). An interactive implementation should support keyboard access, visible focus, and programmatically conveyed expansion state. These requirements extend the static audit; conformance and assistive-technology usability were not established here.

Evidence limits and the next evaluation

The predictive analysis, content audit, and the secondary uncertainty-display experiment answer different questions. A secondary experiment in another decision domain can inform display hypotheses but cannot validate these veterinary prototypes. The clinical database lacks field-level timestamps, so excluding direct score components does not verify a real pre-assessment workflow. Small cohorts, wide intervals, and data-informed fixed specifications in the secondary clinical analyses limit performance interpretation, and routing to lower-uncertainty subsets changes their case mix; lower Brier scores with declining AUROC do not establish a safer policy. Clinician evaluation should compare comprehension, decision time, workload, and inappropriate reliance, while recording whether deferred information is opened and remembered. Testing cases with low predicted probability but threshold-positive reference scores is particularly relevant to the model's observed weakness. Until then, the three interfaces remain alternatives supported by a documented design rationale rather than validated decision-support tools.

The transit analysis illustrates why information density, decision quality, and interaction time should be assessed separately. A display's descriptive payoff rank did not establish an adjusted advantage, and uncertainty remained substantial for duration differences. This average-trial analysis addresses a different estimand from the learning and final-trial comparisons reported by Fernandes et al. (2018). Its findings therefore inform hypotheses for future interface evaluation rather than ranking the canine prototypes. Transit payoffs describe choices under a specified utility function; they do not measure clinical appropriateness, comprehension of model stability, workload, or inappropriate reliance. Recorded durations may include reading, response entry, and pauses. The data contain no progressive-disclosure comparison, so they cannot resolve the trade-off between first-view simplicity and awareness of concealed limitations.

CONCLUSION AND RECOMMENDATION

This study linked a retrospective, dog-grouped analysis of canine acute-pain records to three information-equivalent interface designs. Of 594 observations from 184 dogs, 114 observations from 65 dogs met the complete, clean acute-pain definition. Logistic regression achieved AUROC .627, AUPRC .476, and Brier .210, with wide intervals and low sensitivity at 0.50. The hypothetical pre-score setting, unstable cohort results, and absence of clinician testing limit the conclusions to exploratory prediction and interface specification. The findings motivate four design requirements: label the modeled event as CMPS-SF $\geq 6/24$; distinguish the cross-validated estimate from the range across full-development model refits; present missing context separately from model variability; and retain completion of CMPS-SF as the primary action. The

range must not imply a patient-outcome interval, and the model percentage must not resemble a clinical pain score.

Future evaluation should counterbalance interface order, use dog-level-disjoint cases, and prevent exposure to the same case in multiple formats. Outcomes should include correct interpretation of the probability and stability range, distinction from the CMPS-SF reference, decision time, workload, and inappropriate reliance when model output conflicts with clinical evidence. Expansion behavior and recall of limitations should test the specific trade-off introduced by progressive disclosure. The post hoc routing diagnostic remains exploratory because lower retained-subset Brier scores were accompanied by declining AUROC. Neither that analysis nor the interface audit establishes clinical safety or deployment readiness. The design contribution is the explicit translation of model limitations into visual and interaction requirements whose effects can be evaluated directly.

Data Availability Statement

The Dog Pain Database version 2 is available from Zenodo (Bremhorst, 2025; <https://doi.org/10.5281/zenodo.15304081>). The auxiliary case and vocabulary data are the S1 and S2 Tables of Zanusso et al. (2024), distributed as files `journal.pone.0305925.s003.docx` and `journal.pone.0305925.s004.docx` (<https://doi.org/10.1371/journal.pone.0305925>). Transit records and the original analysis are available from the Fernandes et al. (2018) repository at <https://github.com/michael-fernandes/uncertainty-displays-for-transit> and its archived release at <https://doi.org/10.5281/zenodo.1136329>. The accompanying analysis package contains the data versions, code, derived tables, and figure-generation files used here.

AI Usage Statement

OpenAI ChatGPT assisted analysis-code development, statistical analysis, figure preparation, and manuscript drafting and editing.

REFERENCES

- Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20, Article 310. <https://doi.org/10.1186/s12911-020-01332-6>
- Ancker, J. S., Senathirajah, Y., Kukafka, R., & Starren, J. B. (2006). Design features of graphs in health risk communication: A systematic review. *Journal of the American Medical Informatics Association*, 13(6), 608–618. <https://doi.org/10.1197/jamia.M2115>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Bremhorst, A. (2025). Dog pain database: A multidimensional dataset for investigating canine pain (Version 2) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.15304081>
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), Article 188, 1–21. <https://doi.org/10.1145/3449287>
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.

- Fernandes, M., Walls, L., Munson, S., Hullman, J., & Kay, M. (2018). Uncertainty displays using quantile dotplots or CDFs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Article 144, pp. 1–12). Association for Computing Machinery. <https://doi.org/10.1145/3173574.3173718>
- Frans, N., Hummelen, B., Albers, C. J., & Paap, M. C. S. (2023). Visualizing uncertainty to promote clinicians' understanding of measurement error. *Assessment*, 30(8), 2449–2460. <https://doi.org/10.1177/10731911221147042>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3–42. <https://doi.org/10.1007/s10994-006-6226-1>
- Han, P. K. J., Klein, W. M. P., & Arora, N. K. (2011). Varieties of uncertainty in health care: A conceptual taxonomy. *Medical Decision Making*, 31(6), 828–838. <https://doi.org/10.1177/0272989X11393976>
- Hansen, B. D. (2003). Assessment of pain in dogs: Veterinary clinical studies. *ILAR Journal*, 44(3), 197–205. <https://doi.org/10.1093/ilar.44.3.197>
- Holton, L. L., Scott, E. M., Nolan, A. M., Reid, J., Welsh, E., & Flaherty, D. (1998). Comparison of three methods used for assessment of pain in dogs. *Journal of the American Veterinary Medical Association*, 212(1), 61–66. <https://doi.org/10.2460/javma.1998.212.01.61>
- Holton, L., Pawson, P., Nolan, A., Reid, J., & Scott, E. M. (2001). Development of a behaviour-based scale to measure acute pain in dogs. *Veterinary Record*, 148(17), 525–531. <https://doi.org/10.1136/vr.148.17.525>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
- Lipkus, I. M. (2007). Numeric, verbal, and visual formats of conveying health risks: Suggested best practices and future recommendations. *Medical Decision Making*, 27(5), 696–713. <https://doi.org/10.1177/0272989X07307271>
- Mathews, K., Kronen, P. W., Lascelles, D., Nolan, A., Robertson, S., Steagall, P. V. M., Wright, B., & Yamashita, K. (2014). Guidelines for recognition, assessment and treatment of pain. *Journal of Small Animal Practice*, 55(6), E10–E68. <https://doi.org/10.1111/jsap.12200>
- Monteiro, B. P., Lascelles, B. D. X., Murrell, J., Robertson, S., Steagall, P. V. M., & Wright, B. (2023). 2022 WSAVA guidelines for the recognition, assessment and treatment of pain. *Journal of Small Animal Practice*, 64(4), 177–254. <https://doi.org/10.1111/jsap.13566>
- Morton, C. M., Reid, J., Scott, E. M., Holton, L. L., & Nolan, A. M. (2005). Application of a scaling model to establish and validate an interval level pain scale for assessment of acute pain in dogs. *American Journal of Veterinary Research*, 66(12), 2154–2166. <https://doi.org/10.2460/ajvr.2005.66.2154>
- Nielsen, J. (2006, December 3). Progressive disclosure. Nielsen Norman Group. <https://www.nngroup.com/articles/progressive-disclosure/>
- Padilla, L., Kay, M., & Hullman, J. (2021). Uncertainty visualization. In *Wiley StatsRef: Statistics Reference Online* (pp. 1–18). John Wiley & Sons. <https://doi.org/10.1002/9781118445112.stat08296>
- Reid, J., Nolan, A. M., Hughes, J. M. L., Lascelles, D., Pawson, P., & Scott, E. M. (2007). Development of the short-form Glasgow Composite Measure Pain Scale (CMPS-SF) and derivation of an analgesic

- intervention score. *Animal Welfare*, 16(S1), 97–104.
<https://doi.org/10.1017/S096272860003178X>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
<https://doi.org/10.1038/s42256-019-0048-x>
- Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages* (pp. 336–343). IEEE.
<https://doi.org/10.1109/VL.1996.545307>
- Spiegelhalter, D. (2017). Risk and uncertainty communication. *Annual Review of Statistics and Its Application*, 4, 31–60. <https://doi.org/10.1146/annurev-statistics-010814-020148>
- Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, 21(1), 128–138.
<https://doi.org/10.1097/EDE.0b013e3181c30fb2>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
- Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. In *Proceedings of the 4th Machine Learning for Healthcare Conference* (Vol. 106, pp. 359–380). PMLR.
<https://proceedings.mlr.press/v106/tonekaboni19a.html>
- Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, 17, Article 230.
<https://doi.org/10.1186/s12916-019-1466-7>
- Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7, Article 91. <https://doi.org/10.1186/1471-2105-7-91>
- World Wide Web Consortium. (2024, December 12). Web content accessibility guidelines (WCAG) 2.2. <https://www.w3.org/TR/2024/REC-WCAG22-20241212/>
- Zanusso, F., Contiero, B., Normando, S., Gottardo, F., & De Benedictis, G. M. (2024). Qualitative behavioral assessment of dogs with acute pain. *PLOS ONE*, 19(6), Article e0305925.
<https://doi.org/10.1371/journal.pone.0305925>
- Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 295–305). Association for Computing Machinery. <https://doi.org/10.1145/3351095.3372852>
- Zhong, Z. S., Ma, R., & Zhao, H. (2023). Human-uncertainty distillation for calibrated vision models on CIFAR-10H. *Journal of Advanced Computing Systems*, 3(2), 77–89.
<https://doi.org/10.69987/JACS.2023.30206>