A Consequentialist Critique of Binary Classification Evaluation: Theory, Practice, and Tools

Gerardo A. Flores
MIT

Abigail Schiff
Brigham & Women's Hospital

Alyssa H. Smith
Northeastern University

Julia A. Fukuyama
Indiana University

Ashia C. Wilson
MIT

Abstract

Machine learning-supported decisions, such as ordering diagnostic tests or determining preventive custody, often rely on binary classification from probabilistic forecasts. A consequentialist perspective, long emphasized in decision theory, favors evaluation methods that reflect the quality of such forecasts under threshold uncertainty and varying prevalence, notably Brier scores and log loss. However, our empirical review of practices at major ML venues (ICML, FAccT, CHIL) reveals a dominant reliance on accuracy and AUC-ROC. To address this disconnect, we introduce a decision-theoretic framework mapping evaluation metrics to their appropriate use cases, along with a practical Python package, `briertools`, designed to make proper scoring rules more usable in real-world settings. Specifically, we implement a bounded-threshold variant of the Brier score and log loss that restricts evaluation to a practitioner-specified range of plausible cost ratios, rather than averaging over the full unit interval. We further contribute a theoretical reconciliation between the Brier score and decision curve analysis, directly addressing a longstanding critique by Assel et al. (2017) regarding the clinical utility of proper scoring rules.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

1 INTRODUCTION

We study a setting in which a binary classifier $\kappa(\cdot\,;\tau)$: $\mathcal{X} \to \{0,1\}$ is developed to map an input $x \in \mathcal{X}$ to a binary decision. Such classifiers are foundational to decision-making tasks across domains, from healthcare to criminal justice, where outcomes depend on accurate binary choices. The decision is typically made by comparing a score $s(x) \in \mathbb{R}$, such as a probability or a logit, to a threshold $\tau \in \mathbb{R}$:

$$\kappa(x;\tau) = \begin{cases} 1 & \text{if } s(x) \geq \tau \\ 0 & \text{if } s(x) < \tau. \end{cases}$$

The threshold τ is a parameter that can be adjusted to control the tradeoff between false positives and false negatives, reflecting the specific priorities or constraints of a given application. For example, consider a scenario in which a classifier is used to make (a) judicial decisions, such as who to sentence, or (b) medical decisions, such as recommending treatments for diagnosed conditions. Which threshold should be chosen and how should the resulting classifiers be evaluated? We advocate for a *consequentialist view* of classifier evaluation: pragmatic classifier evaluation should model the real-world impacts of using that classifier to make concrete decisions.

To this end, we introduce a *value* function, $V(\kappa(x;\tau), y)$, which assigns a value to each prediction given the true label y and the classifier's decision $\kappa(x;\tau)$. The overall performance of a classifier is given by its expected value over a distribution ($\mathcal{D}$: $\mathbb{E}_{(x,y)\sim\mathcal{D}}[V(\kappa(x;\tau), y)]$.)

We identify two key factors that determine appropriate evaluation approaches:

1. **Instance Coupling.** whether decision costs are additive, allowing independent decisions across instances (Indep.), or there is a fixed quota to act on positive predictions (Top-K decisions); and

2. **Threshold Specificity.** whether the decision threshold is known exactly or approximately at model selection time (regardless of what will later be known at deployment time).

Table 1 illustrates how different evaluation metrics align with these settings. We develop a taxonomy for the decision problems for which classifiers are typically used, and we describe which commonly used metrics represent a decision-theoretic evaluation of those problems in potentially uncertain decision contexts. This taxonomy makes it simple to choose metrics that are consistent with deployment scenarios.

		Instance Coupling	
		Independent	**Top-K**
Threshold Specificity	**Mixed τ**	Brier Log Loss	AUC-ROC AUC-PR
	Fixed τ	Net Benefit Accuracy	Precision@K Recall@K

Table 1: **Metric Selection Recommendations.** First, identify whether decisions can be made independently or there is a fixed budget of positive labels (column). Then, identify whether the threshold is uncertain or precisely known (row). The dark blue cell is the most common decision context; the light blue cells' metrics are frequently used despite worse metric-context alignment (note especially Accuracy, an unrealistic special case of Net Benefit that is the most common metric overall). The white cell is included for completeness.

Despite pervasive threshold uncertainty in real-world ML applications, such as healthcare and the criminal legal system, evaluation practices remain misaligned with deployment realities. Evaluations based on independent decisions typically assume fixed thresholds, while evaluations that are based on threshold mixtures typically assume dependent decisions. Our analysis of three major ML conferences (ICML, FAccT, CHIL) reveals a consistent preference for metrics designed for fixed or top-K decisions, which do not match common deployment settings where thresholds are uncertain and decisions are independent.

1.1 Contributions

Core theoretical contributions. Our main new result is a **bounded-threshold extension of proper scoring rules**. For independent decisions with uncertainty over a bounded interval of cost ratios, we derive bounded versions of the Brier score and log loss that average regret only over that interval rather than over the full unit interval. This fills the gap between point-threshold metrics such as net benefit and full-interval proper scoring rules.

Our second contribution is a **reconciliation of proper scoring rules with Decision Curve Analysis** in a common regret-based framework, which clarifies when each applies and shows how bounded-threshold scores address the critique of Assel et al. (2017) that standard Brier evaluation averages over implausible thresholds.

As an appendix-level result, we also derive an interpretation of Concordant Partial AUC in the same framework, showing it implicitly averages over a score-induced threshold distribution.

Practice-facing contributions. We provide a decision-context taxonomy, supported by an LLM-assisted survey of current evaluation practice, for matching binary-classification metrics to deployment settings. To support this use in practice, we release `briertools`, which implements the proposed metrics and associated visualizations, and we illustrate the framework on a breast-cancer treatment case study.

2 PRELIMINARIES: CURRENT PRACTICE AND ESTABLISHED THEORY

This section introduces a consequentialist approach that matches metrics to implicit decision problems. We use this approach to summarize known results about optimal thresholding under calibration, accuracy as a special case of net benefit, regret, and the structure of proper scoring rules. We then survey the usage of these metric families in the literature.

2.1 Related work

Dependent Decisions. ROC/AUC was first adopted in psychophysics and radiology; in ML, it evaluates ranking quality under varying thresholds. See Appendix A for more details. There have been consistent critiques of AUC's lack of calibration information Van Calster et al. (2019); Kwegyir-Aggrey et al. (2023). Hand (2009) and Hernández-Orallo et al. (2012) showed that AUC-ROC can be interpreted as a cost-weighted average regret, particularly under calibrated or quantile-based forecasts. Several attempts have been made to evaluate area under only portions of the ROC curve (McClish, 1989; Carrington et al., 2023), but these have not produced clear decision-theoretic interpretations.

Independent Decisions. Vickers and Elkin (2006); Steyerberg and Vickers (2008) and Assel et al. (2017) introduced decision curve analysis (DCA) as a threshold-restricted net benefit visualization, arguing it offers greater clinical relevance than Brier-based aggregation across all thresholds. Recent work has examined the de-

composition of Brier score and log loss into calibration and discrimination components (Shen, 2005; Siegert, 2017; Dimitriadis et al., 2024), providing implementation and visualization guidance.

Shuford et al. (1966) first characterize proper scoring rules by an integral formulation, and Savage (1971) characterize the variable of integration as the rate of substitution. Schervish (1989) realizes this is an integral over binary decision problems at different cost ratios, and Shen (2005) interprets the weight as describing which thresholds are important to the evaluator. Gneiting and Raftery (2007) then connects this integral representation to convexity properties of proper scoring rules and extends it beyond the binary case.

Hernández-Orallo et al. (2012) then studies the behavior of this integral under different threshold decision rules. Pfohl et al. (2022) conclude overall net benefit is maximized with empirical risk minimization and recalibration, and that picking optimal thresholds is equivalent to this post-hoc recalibration.

Surrogate Loss Consistency The requirement that evaluation metrics align with the downstream decision problem is conceptually related to the notion of *surrogate loss consistency* in statistical learning. In that setting, the question is whether the surrogate loss used for training leads to a model that is optimal for the true target loss. This principle was introduced in inference by Fisher (1922) and developed in a modern machine learning context by Bartlett et al. (2006) and Zhang (2004). The focus there, however, is on the choice of training loss and its consistency with the parameter being estimated. Our setting is different: we are concerned not with training-time surrogate losses, but with deployment-time evaluation metrics. Uncertainty in our case arises from decision contexts and thresholds, rather than from prediction values or parameter estimation.

2.2 Common Metric Choices

We analyze evaluation metrics used in papers from ICML 2024, FAccT 2024, and CHIL 2024 using an LLM-assisted review (see Appendix K for more details on our analysis). Accuracy was the most common metric at ICML and FAccT (> 50%), followed by AUC-ROC; CHIL favored AUC-ROC, with AUC-PR also notable. Proper scoring rules (e.g. Brier score, log loss) were rarely used (< 15% and < 5%, respectively). These findings, summarized in Figure 7, confirm the dominance of accuracy and AUC-ROC in practice. This paper addresses this gap by clarifying when Brier scores and log losses are appropriate; further, it provides theoretical justification and tools to support their adoption.

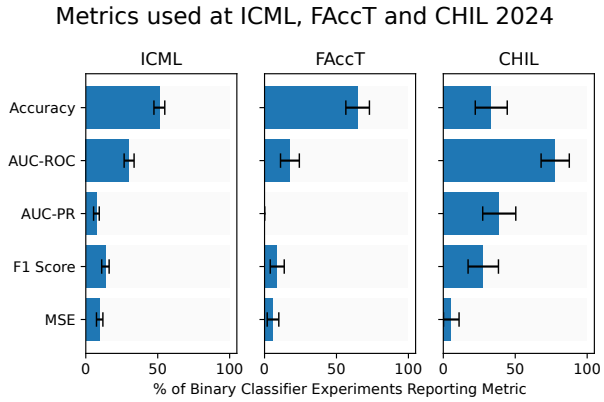

Figure 1: Claude 3.5 Haiku was used to analyze 2,610 papers from three major 2024 conferences. Each plot summarizes the evaluation metrics used for binary classifiers. Accuracy dominates outside healthcare, while AUC-ROC is more prevalent within healthcare domains. Error bars come from binomial confidence intervals.

2.3 Consequentialist Formalism

Our goal is to evaluate binary classifiers in a way that directly reflects their real-world decision consequences. A natural abstraction is **expected regret**, which we define as *the excess cost of using a model compared to the best possible decision rule under a given cost structure.*

Cost model We adopt the cost model introduced by Angstrom (1922), where perfect prediction defines a zero-cost baseline, predicted positives incur an immediate cost C, and false negatives incur a downstream loss L. Without loss of generality, we normalize $L = 1$ and define the relative cost as $c = C/L$. See Appendix B for more details about these choices and their implications.

$V(y, a)$	$a = 0$	$a = 1$
$y = 0$	0 (True Neg)	c (False Pos)
$y = 1$	$1 - c$ (False Neg)	0 (True Pos)

This structure highlights that all evaluation reduces to how well a classifier navigates the cost tradeoff encoded by c.

Regret Let $\pi = P(y = 1)$ denote prevalence, $F_0(\tau)$ the CDF of negative scores, and $F_1(\tau)$ the CDF of positive scores. We define regret as

$$R(\kappa, \pi, c, \tau, \mathcal{D}) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\Big[V(\kappa(x;\tau), y)\Big]$$
$$= c \cdot (1 - \pi) \cdot (1 - F_0(\tau)) + (1 - c) \cdot \pi \cdot F_1(\tau).$$

This quantifies the expected loss of a classifier at threshold τ relative to the idealized zero-cost baseline (Hernández-Orallo et al., 2012).

Theorem 2.1 (Optimal Threshold)**.** *Given a calibrated model, the optimal threshold is the cost:*

$$\arg\min_{\tau} R(\kappa, \pi, c, \tau, \mathcal{D}) = c.$$

See Appendix C.1 for a brief proof; see also Hernández-Orallo et al. (2012).

Assumptions To isolate the role of **cost** in shaping evaluation, we adopt three simplifying assumptions:

- **Fixed prevalence** (π does not shift between training and deployment). This removes confounding from label shift and allows us to focus purely on cost-driven evaluation.
- **Score as probability** The decision-maker treats the model's output $s(x)$ as a probability, that is, acts as though $P(Y = 1|s(X) = p) = p$. This allows us to model the decision-maker's threshold-setting behavior.
- **Optimal thresholding** ($\tau^* = c$ is always chosen). This assumes the decision-maker makes the optimal choice given their beliefs and the cost ratio. This lets us express regret in its minimal form:

$$R^*_\kappa(c) = c \cdot (1 - \pi) \cdot (1 - F_0(c)) + (1 - c) \cdot \pi \cdot F_1(c).$$

These assumptions restrict attention to the core decision problem: how well does a classifier balance false positives and false negatives given a cost ratio? By abstracting away dataset shift and thresholding complications, we can evaluate metrics solely on their ability to reflect cost-sensitive decision quality.

The second assumption is about the behavior of the decision-maker, not the match between the model and the real world (see the concept of small worlds in Savage (1954)). If in fact the model is poorly calibrated, Hernández-Orallo et al. (2012) show theoretically that picking an optimal threshold is tantamount to recalibrating the model and then using the formula from the third assumption. Pfohl et al. (2022) verify this empirically.

In what follows, we show that many standard metrics, like accuracy and AUC-ROC, can be reinterpreted as expectations of regret over particular cost distributions. We clarify which metrics align with which kinds of decision problems.

2.4 Independent Decisions with Fixed Thresholds: a Consequentialist View of Accuracy

Accuracy is the most commonly used metric for evaluating binary classifiers, as it offers a simple measure of correctness that remains the default in many settings Huang and Ling (2005). Formally:

Definition 2.2 (Accuracy)**.** Given data $\{(x_i, y_i)\}_{i=1}^n$ with $y_i \in \{0, 1\}$, and a binary classifier $\kappa(x; \tau)$ thresholded at τ, accuracy is defined as:

$$\text{Accuracy}(\kappa, \mathcal{D}) \triangleq \tfrac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\kappa(x_i; \tau) = y_i).$$

Accuracy corresponds to regret minimization when misclassification costs are equal:

Proposition 2.3. *Let τ denote a (possibly suboptimal) threshold. Then,*

$$\textit{Accuracy}(\kappa, \mathcal{D}) = 1 - 2 \cdot R(\kappa, \pi, c = 1/2, \tau, \mathcal{D}).$$

This equivalence, proved in Appendix C.2 and established by Hernández-Orallo et al. (2012), underscores a key limitation: accuracy implicitly assumes all errors carry equal cost. In many domains, this is neither justified nor appropriate. In criminal sentencing, optimizing for accuracy equates wrongful imprisonment with wrongful release, an assumption at odds with most legal and ethical frameworks. In prostate cancer screening, false negatives may lead to death, while false positives can result in unnecessary treatment that causes erectile dysfunction. It is natural to model the cost of death (normalized to 1) as equal across all patients. However, the implied cost ratio $c = 1/2$ (as if erectile dysfunction were half as bad as death) both oversimplifies outcomes and disregards heterogeneous patient preferences. Accuracy is therefore only meaningful when error costs are balanced, prevalence is stable, and tradeoffs are collectively agreed upon. These conditions are rarely satisfied in practice.

2.5 Dependent Decisions (top K) with Mixtures of Thresholds: A Consequentialist View of AUC-ROC

In resource-constrained settings, such as allocating ICU beds, allocation decisions are coupled by the fixed budget of K predicted positives. When K is known exactly at model selection time, Net Benefit@K is the theoretically appropriate metric; however, we show in Appendix H that it induces the same model ranking as the more popular Precision@K or Recall@K at fixed N, K, and c.

When K is not known at model selection time or varies across deployments, the decision maker faces a mixture over possible thresholds. AUC-ROC reflects this mixture. When scores are calibrated, AUC admits a regret interpretation due to Hand (2009):

$$\text{AUC-ROC}(\kappa) \;=\; 1 - \frac{1}{2\pi(1-\pi)}\, \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[R^*_\kappa\big(s(x)\big)\big]\,.$$

That is, it averages τ^*-regret using the *model's score distribution* as the threshold prior (derivation in Thm. I.6 in Ap. I).

The limitation is that model scores are trained to estimate associations with outcomes rather than the costs of decisions. As Hand (2009) notes, this means the model itself determines the relative importance of false positives and false negatives, effectively assigning how costly it is to miss a cancer diagnosis or how acceptable it is to release a guilty person.

2.6 Independent Decisions with Mixtures of Thresholds: a Consequentialist View of Brier Scores

While accuracy is widely used as an evaluation metric, it is rarely directly optimized. Instead, practitioners generally use surrogate losses such as squared error and log loss (also known as cross-entropy), largely based on their differentiability.

However, these functions can also be used after training to evaluate the performance of a classifier on a given dataset. This represents a use case that is distinctly different from surrogate loss minimization. Decades of research in the forecasting community have demonstrated that these loss functions have a deeper interpretation: they represent distinct notions of average regret, each corresponding to different assumptions about uncertainty and decision-making (see 2.4 and 2.5, which are derived in Shuford et al. (1966), with interpretation clarified in Schervish (1989)). From a consequentialist perspective, these tractable, familiar methods are not being used to their full potential as evaluation measures.

Theorem 2.4 (Brier Score as Uniform Mixture of Regret)**.** *Let $\kappa : \mathcal{X} \to [0,1]$ be a probabilistic classifier with score function $s(x)$, and let $\mathcal{D}$ be a distribution over $(x,y) \in \mathcal{X} \times \{0,1\}$. Then the Brier score of κ is the mean squared error between the predicted probabilities and true labels:*

$$\mathrm{BS}(\kappa, \mathcal{D}) \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[(y - s(x))^2 \right].$$

Moreover, this is equivalent to the expected minimum regret over all cost ratios $c \in [0,1]$, where regret is computed with optimal thresholding:

$$\mathrm{BS}(\kappa, \mathcal{D}) = \mathbb{E}_{c\sim Uniform[0,1]} \left[R^*_\kappa(c) \right].$$

The result is known from Shuford et al. (1966), but a more general proof appears in Appendix D.6, where this unbounded version arises as a special case. The following Theorem 2.5 establishes that unlike the Brier score, which weights regret uniformly across thresholds, log loss emphasizes extreme cost ratios via the weight $\frac{1}{c(1-c)}$.

Theorem 2.5 (Log Loss as a Weighted Average of Regret)**.** *Let $\kappa : \mathcal{X} \to [0,1]$ be a probabilistic classifier with score $s(x)$, and let $\mathcal{D}$ be a distribution over $(x,y) \in \mathcal{X} \times \{0,1\}$. Then:*

$$\begin{aligned}\mathrm{LL}(\kappa, \mathcal{D}) &= \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[-\log \left(s(x)^y (1 - s(x))^{1-y} \right) \right] \\ &= \int_0^1 \frac{R^*_\kappa(c)}{c(1-c)}\, dc = \int_{-\infty}^{\infty} R^*_\kappa \left(\frac{1}{1+e^{-\ell}} \right) d\ell.\end{aligned}$$

Like the Brier score, log loss integrates regret uniformly over log-odds of cost ratios, assigning more weight to rare but high-consequence decisions. This makes log loss more sensitive to tail risks, which may be desirable when one type of error carries disproportionate cost. See Figure 5 in Appendix G for a visualization.

In practice, even though proper scoring rules such as log loss are widely used during training, final model selection often reverts to fixed-threshold metrics. To address this, we introduce bounded-threshold metrics that restrict evaluation to a plausible range of costs without requiring exact specification at the time of model selection.

3 NOVEL RESULTS: BOUNDED THRESHOLD SCORING RULES AND THEIR RELATION TO DCA

Must scoring rules represent regret averaged across all possible thresholds, or can they incorporate expert-level judgment about relevant decision contexts directly? To investigate this question, we analyze Decision Curve Analysis (DCA) (Vickers and Elkin, 2006; Vickers et al., 2019), a decision-theoretic approach from medical informatics that formed the basis of a prominent critique of the Brier score. Assel et al. (2017) argued that averaging over the full threshold range assigns weight to cost ratios no clinician would consider plausible.

We show that DCA can itself be understood as a regret-based evaluation approach, in the same family as the Brier score and log loss. While the specific formulation of DCA makes it ill-suited to averaging across thresholds, the critique highlights the importance of restricting evaluation to clinically relevant cost ratios. Accordingly, we introduce bounded-threshold versions of the Brier score and log loss that *do* support averaging regret over restricted intervals of decision-context uncertainty.

3.1 Decision Curve Analysis

DCA is built around the notion of *net benefit*, the quantity plotted against the threshold in DCA, which we show to be equivalent to regret up to constant terms.

Definition 3.1 (Net Benefit (DCA))**.** As defined by Vickers et al. (2019), the net benefit at decision threshold $\tau \in (0,1)$ is given by:

$$\mathrm{NB}(\tau) = (1 - F_1(\tau)) \cdot \pi - (1 - F_0(\tau)) \cdot (1 - \pi) \cdot \tfrac{\tau}{1-\tau}.$$

Theorem 3.2 (Net Benefit as a function of regret)**.** *Let π denote the prevalence of the positive class, which is the maximum achievable benefit under perfect classification. The net benefit at threshold c is related to the regret as follows:*

$$\mathrm{NB}(c) = \pi - \tfrac{1}{1-c} R^*_\kappa(c).$$

Assel et al. (2017) argue that the Brier score is inadequate for clinical settings where only a narrow range of decision thresholds is relevant (e.g. determining the need for a lymph node biopsy). Comparing the unrestricted Brier score to net benefit at fixed thresholds (e.g., 5%, 10%, 20%), they conclude that net benefit better captures clinical priorities. However, once Brier score is understood as a weighted average of an affine transformation of net benefit across cost ratios, this critique elucidates a useful insight: the appropriate comparison for a fixed-threshold method like DCA is not to average over the full range of costs, as the Brier Score does, but to average only over the relevant interval (e.g., [5%, 20%]). We now turn to this problem.

3.2 Regret over a Bounded Range of Thresholds

We derive a new and computationally efficient expression for expected regret when the cost ratio c is distributed uniformly over a bounded interval $[a,b] \subseteq [0,1]$ by exploiting the duality between pointwise squared error and average regret. This formulation not only improves numerical stability but also simplifies implementation, requiring only two evaluations of the Brier score under projection.

Selecting $[a,b]$ requires the same clinical reasoning already used in decision curve analysis: a practitioner identifies the cost ratios that are plausible for the population of interest by reasoning about what makes a threshold too low (intervention harms outweigh any realistic benefit) or too high (no patient would demand that level of certainty before acting) (Vickers et al., 2016). The uniform prior over this interval is then the least structured choice consistent with those bounds. Prior approaches that require a full distributional specification, such as the Beta priors of Hand (2009) and Zhu et al. (2024), demand dispersion parameters that practitioners cannot easily elicit or justify. See Appendix G for a detailed comparison.

Throughout, we will use notation $\mathrm{clip}_{[a,b]}(z) \triangleq \max(a, \min(b, z))$ to denote the projection of z onto the interval $[a,b]$.

Theorem 3.3 (Bounded Threshold Brier Score)**.** *For a classifier κ, the average minimal regret over cost ratios $c \sim$ Uniform(a,b) is given by:*

$$\mathop{\mathbb{E}}_{c\sim Unif(a,b)} R^*_\kappa(c) = \frac{1}{b-a}\Bigg[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\Big(y - \mathrm{clip}_{[a,b]}(s(x))\Big)^2 - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\Big(y - \mathrm{clip}_{[a,b]}(y)\Big)^2\Bigg].$$

The result follows as a direct extension of the proof of Theorem 2.4. Specifically, the same argument structure applies with the necessary modifications that account for the additional constraints introduced in this setting. For a complete derivation, refer to the proof of Theorem D.5 in the Appendix, where the argument is presented in full detail, including Proposition D.4 which shows that the clipped version is proper but not strictly proper.

Overall, this expression offers two practical advantages. First, it is computationally efficient, requiring only 2 Brier score evaluations, one on predictions and one on labels, after projecting onto $[a,b]$. Second, it is interpretable, recovering the standard Brier score when $a = 0$ and $b = 1$, consistent with the assumption that true labels lie in $\{0,1\}$.

Theorem 3.4 (Bounded Threshold Log Loss)**.** *Let κ be a probabilistic classifier with score function $s(x)$. Let $c = \frac{1}{1+\exp(-\ell)}$ denote the cost ratio corresponding to log-odds ℓ, and suppose ℓ is distributed uniformly over the interval $[\log\frac{a}{1-a}, \log\frac{b}{1-b}]$, where $0 < a < b < 1$. Then the expected regret over this range is given by:*

$$\mathbb{E}_{\ell\sim Uniform\left(\log\frac{a}{1-a},\,\log\frac{b}{1-b}\right)}\left[R^*_\kappa\Big(\frac{1}{1+\exp -\ell}\Big)\right] = \frac{1}{\log\frac{b}{1-b} - \log\frac{a}{1-a}}\Bigg(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\log(1 - |y - \mathrm{clip}_{[a,b]}(s(x))|)] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\log(1 - |y - \mathrm{clip}_{[a,b]}(y)|)]\Bigg).$$

Similarly, this result follows as a direct extension of the proof of Theorem 2.5. The argument structure remains the same, incorporating appropriate modifications to account for the additional constraints in this setting. For a complete derivation, refer to the proof of Theorem D.6 in the Appendix, where the full details are provided. Like the bounded threshold Brier score, our bounded threshold log loss score is practical to implement: it requires only two calls to a standard log loss function with clipping applied to inputs. Moreover, when $a = 0$ and $b = 1$, the second term vanishes, recovering the standard log loss.

Remark 3.5 (Bounded Threshold AUC-ROC)**.** If we assume that a model is calibrated, then a similar idea can be applied to the AUC-ROC. The resulting metric is equivalent to an expected regret at a set of thresholds defined by the score distribution of the model *in the interval* (see Theorem I.5 in the Appendix). However, all of Hand's original critiques remain. The model is still trained to estimate associations with outcomes rather than decision costs. As a result, normative judgments about cost, harm, and acceptability are deferred to the statistical model rather than specified directly.

Remark 3.6 (Skill Scores)**.** Although the bounded threshold metrics are precisely an expected regret, and can be rescaled to match other decision costs, in casual use it can be simpler to use them while rescaling so 100% is a perfect model and 0% is a naive baseline. This is known as a skill score; we expand on this in detail in Appendix J.

3.3 Revisiting the Brier score Critique by Assel et al. (2017)

We reproduce the original results from Assel et al. (2017) (see Appendix F) and show that bounded Brier score rankings closely track those of net benefit at 5%. The scores diverge only when net benefit itself varies substantially across thresholds. This indicates that the main obstacle has been tooling rather than theory. Bounded scoring rules provide a principled and interpretable alternative that respects threshold constraints while aligning more closely with clinical decision-making.

A natural question then arises: why not simply average net benefit over the relevant interval? Area-under-the-curve (AUC) aggregation is not typically practiced in DCA (Steyerberg and Vickers, 2008), in part because it lacks a clear interpretation. By leveraging our equivalence with regret, we show how a mathematically similar procedure yields a well-defined average of net benefit over a bounded range of thresholds.

Theorem 3.7 (Bdd Threshold Net Benefit)**.** *Let* $L(x,y) = \begin{cases} s(x) & \text{if } y=1 \\ (1-s(x)) - \ln(1-s(x)) & \text{if } y=0 \end{cases}$ *be a pointwise loss. For a classifier* κ*, the integral of net benefit (NB) over the interval [a,b] is the loss for the predictions clipped to [a,b] minus the loss for the true labels clipped to [a,b]:*

$$\mathop{\mathbb{E}}_{c\sim Unif(a,b)} \mathrm{NB}(c) = \pi - \frac{1}{b-a}\Bigg[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} L(\mathrm{clip}_{[a,b]}(s(x)), y) - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} L(\mathrm{clip}_{[a,b]}(y), y)\Bigg].$$

While mathematical equivalence resolves formal concerns, it does not address semantic limitations. In prostate cancer screening, for instance, patients may agree on the value of survival as a primary goal but disagree on how to value life with treatment side effects. The direct benefit of treatment is therefore constant, but the net benefit of treatment varies with the threshold. Standard DCA treats the gross benefit minus gross cost of a true positive as fixed across patients, even when their valuations of treatment burdens differ. This is inconsistent in settings with heterogeneous preferences. The Brier score, in contrast, fixes the direct false negative penalty while varying the overtreatment cost with the threshold. This allows the net value of a true positive to adjust with patient preferences, yielding more coherent semantics for population-level averaging under cost heterogeneity. These semantics can also be recovered from decision curves through axis rescaling: quadratic transformations yield the Brier score (Appendix E.3), while logarithmic transformations yield log loss (Appendix E.4). Figure 4 illustrates these relationships.

Although DCA is mathematically close to the Brier score and log loss, its formulation is inherently pointwise. Averaging over thresholds therefore lacks a clear interpretation, limiting DCA to fixed-threshold comparisons.

4 `BRIERTOOLS`: A PYTHON PACKAGE FOR FACILITATING THE ADOPTION OF BRIER SCORES

We introduce a Python package, `briertools`, designed to fill the gap in evaluation support tools by making Brier scores and their bounded-threshold variants easier to apply in practice. The package provides utilities for computing bounded-threshold scoring metrics and for visualizing the associated regret and decision curves. It is installable via `pip` and supports common use cases with minimal overhead.

In addition to providing efficient implementations that exploit the duality between pointwise error and average regret, `briertools` includes plotting functions for regret curves and threshold-based diagnostics. Although plotting regret against thresholds is slower and less precise than our formulas for estimating quadrature, it is valuable for debugging and exploratory analysis. As recommended by Dimitriadis et al. (2024), such visualizations help detect unexpected behaviors across thresholds and offer deeper insight into model performance under varying decision boundaries.

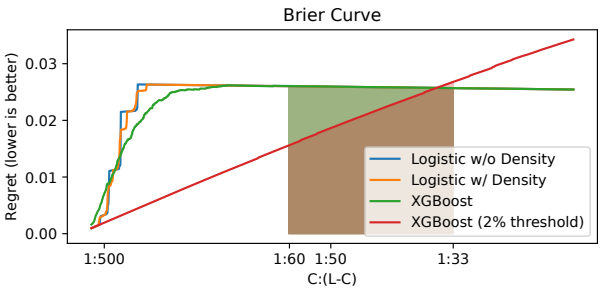

Figure 2: A comparison of breast cancer prediction models' performance over the range of commonly suggested thresholds for treatment.

4.1 Example: Breast Cancer Risk Prediction under Ambiguous Treatment Thresholds

Metric	w/o dens.	w/ dens.	xgb	xgb (2%)
AUC	0.601	**0.606**	0.574	0.591
Brier	0.0263	0.0263	**0.0263**	0.0415
Log Loss	0.156	0.156	**0.155**	0.193
Thr. LogLoss	0.0153	0.0153	0.0153	**0.0123**
Thr. Brier	0.000662	0.000662	0.000662	**0.000545**

Table 2: Performance comparison of classifiers on breast cancer risk prediction. Modifying the internal decision threshold for xgboost during training decreases overall performance but improves performance in the clinically relevant range of thresholds.

Treatment guidelines for breast cancer prevention often recommend Tamoxifen when a patient's estimated cancer risk exceeds a specified threshold. However, there is no consensus as to what this threshold should be. Historical standards have set benefit at 1.66% risk Barlow et al. (2006), while the US Preventive Services Task Force now recommends 3%. A 2% threshold is also frequently cited in the literature. Such disagreement leaves clinicians with the challenge of making treatment decisions under uncertain, contested standards.

Methods We analyze the Breast Cancer Surveillance Consortium Risk Estimation Dataset [1] Barlow et al. (2006), following prior work by Yang et al. (2022). They compared models trained with and without breast tissue density as a predictor, finding little difference in baseline Brier score. Building on this, we additionally train [2] and evaluate two XGBoost classifiers, one of which was modified to use an internal decision threshold of 2%. We assess all models with both traditional metrics (AUC-ROC, Brier score, log loss) and bounded-threshold scoring rules over the clinically plausible range of 1.66%-3%.

[1] Licensing details can be found at `https://www.bcsc-research.org/index.php/datasets/rfdataset`.

[2] We trained using CPU only, on an off-the-shelf laptop.

Findings Across the full range of thresholds, the modified XGBoost model performs worse than both the baseline XGBoost and the logistic models from Yang et al. (2022). However, when evaluation is restricted to the clinically relevant interval, it outperforms all alternatives.

Sensitivity to Bound Selection The Log Loss curve (Figure 2) supports a visual sensitivity analysis. The bounded-threshold Log Loss for any interval $[a, b]$ is the average height of a model's curve over that interval. In the clinically relevant region near [1.66%, 3%], the XGBoost model with a 2% internal threshold (red) has the lowest regret; this ranking is stable under moderate perturbation of the bounds. As the interval widens toward the full $[0, 1]$, the red curve's steep rise at higher cost ratios eventually dominates, which is why the global Log Loss ranks the model worst.

Implications This case study demonstrates that threshold-aware evaluation can change model selection. While global metrics penalize models that deviate from average performance, bounded-threshold scoring rules reveal which models are best suited for specific decision contexts. In domains such as medicine, where professional consensus on treatment thresholds is often absent, evaluation methods must capture performance over plausible ranges rather than relying on fixed cutoffs.

Commensurable Calibration and Discrimination A distinctive feature of `briertools` is that it allows users to evaluate calibration and discrimination on a common, commensurable scale. This is something that standard evaluation practices cannot do.

Top-K metrics, for example, evaluate only the ordering of predicted scores and are entirely insensitive to calibration. They can be used only under the assumption that more than K individuals would benefit from a positive label. In response, it is often recommended that practitioners report both AUC-ROC and Expected Calibration Error (ECE). Yet these numbers cannot be meaningfully combined, leaving practitioners without a principled way to make joint decisions.

In contrast, proper scoring rules such as the Brier score and log loss inherently account for both discrimination and calibration (Shen, 2005; Dimitriadis et al., 2024); they also admit additive decompositions that make this distinction explicit. We obtain the post-hoc calibrated model by applying the Pool Adjacent Violators (PAV) algorithm to the test predictions, following the procedure described in Dimitriadis et al. (2024). The additive decomposition then follows directly. For the Brier Score, this is a standard bias-variance decompo-

sition for squared error, and for log loss, the decomposition separates calibration error from irreducible uncertainty via KL-divergence between the calibrated and uncalibrated models (Shen, 2005).

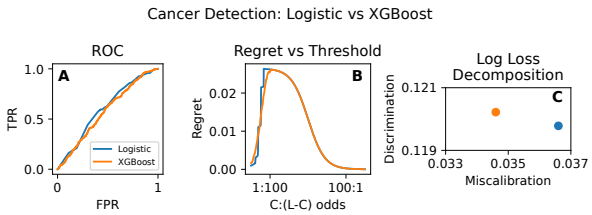

Figure 3: A. ROC plot shows XGBoost has worse discrimination than Logistic Regression models as recommended by Yang et al. (2022). B. Log Loss Curve shows XGBoost has better average regret. C. Decomposition reconciles the two; XGBoost has much better calibration, and only slightly worse discrimination.

This matters for practitioners because miscalibration can strongly affect evaluation outcomes. Subgroup analyses based on top- K metrics, for instance, may suggest misleading fairness conclusions when calibration is poor (Kallus and Zhou, 2019), and AUC-ROC does not reflect error rates at operational thresholds (Kwegyir-Aggrey et al., 2023). Figure 3 shows how log loss decomposition makes these trade-offs transparent: a model with higher AUC but poor calibration (blue) can be revealed as inferior to one with slightly weaker discrimination but much stronger calibration (orange). By surfacing this calibration gap directly, `briertools` enables practitioners to evaluate models in a way that is both rigorous and actionable.

5 DISCUSSION

The regret formalism reveals a unifying structure beneath Table 1: every metric in the table is a weighted average of $R^*_\kappa(c)$, and the metrics differ only in which cost ratios receive weight. This makes the table a prescriptive guide: to choose a metric, identify the deployment setting and use the weighting that matches it. Reading the table against our survey of 2,610 papers from ICML, FAccT, and CHIL 2024 reveals a systematic mismatch between the metrics practitioners report and the settings they occupy.

Dependent decisions. When a fixed budget couples decisions across instances, ranking metrics become appropriate: Net Benefit@K when the budget K is known, AUC-ROC when K varies across deployments. However, when used for independent decisions, AUC-ROC implicitly weights cost ratios by the model's own score distribution (Hand, 2009), letting a model trained to predict outcomes implicitly determine how costly each type of error is.

Fixed thresholds. For independent decisions, accuracy dominates practice (>50% of binary-classification papers at ICML and FAccT), yet it is simply net benefit at the special case $c = 1/2$ (Proposition 2.3), justified only when error costs are equal. DCA allows asymmetric costs but remains pointwise, evaluating R^*_κ at a single cost ratio.

Uncertain thresholds. Eliminating coupled decisions and fixed thresholds places us in the top-left cell of Table 1—the setting of most real deployments, yet the least used in our survey (proper scoring rules appear in fewer than 15% of papers and fewer than 5% at ICML). Standard proper scoring rules average $R^*_\kappa(c)$ over the full unit interval, giving equal weight to irrelevant cost ratios—precisely the over-averaging that Assel et al. (2017) criticized. Our bounded-threshold variants (Theorems 3.3, 3.4) resolve this by averaging only over a practitioner-specified interval $[a, b]$. The DCA reconciliation (Theorem 3.7) confirms that bounded Brier *is* averaged net benefit over $[a, b]$: DCA is the right tool when the threshold is fixed, whereas bounded scores provide a principled average when the threshold is uncertain but confined to a plausible range. The breast cancer case study shows this distinction has practical consequences: bounded evaluation reverses model rankings, selecting the model that dominates in the clinically relevant range.

These bounded scoring rules jointly reward calibration and discrimination, admit additive decompositions (Figure 3), and encode threshold uncertainty as a simple interval rather than a full distributional specification. As with any threshold-based evaluation, these interpretations require scores on a probability scale; this calibration assumption should be stated explicitly whenever bounded metrics are reported.

Limitations. The clean semantics of averaging regret over a cost-ratio interval depend on the assumption that the gross cost of a positive label does not depend on the true class, and that the false-negative cost is invariant across cost-ratio scenarios.

Guidance for Practitioners. To support adoption, we provide `briertools`, an `sklearn`-compatible package for computing and visualizing bounded-threshold scores, regret curves, and calibration decompositions.

Future Work. Extending these results to deployment shift or multi-class settings would broaden their applicability.

Acknowledgments

We thank the participating women, mammography facilities, and radiologists of the BCSC (`http://www.bcsc-research.org/`) for the data.

References

N. Adams and D. Hand. Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*, 32(7):1139–1147, 1999. ISSN 0031-3203. doi: https://doi.org/10.1016/S0031-3203(98)00154-X. URL https://www.sciencedirect.com/science/article/pii/S003132039800154X.

A. Angstrom. On the effectivity of weather warnings. *Nordisk Statistisk Tidskrift*, 1:394–408, 1922.

M. Assel, D. D. Sjoberg, and A. J. Vickers. The brier score does not evaluate the clinical utility of diagnostic tests or prediction models. *Diagnostic and Prognostic Research*, 1(1):19, 2017. doi: 10.1186/s41512-017-0020-3. URL https://doi.org/10.1186/s41512-017-0020-3.

M. Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26(4):641–647, 1955. ISSN 00034851. URL http://www.jstor.org/stable/2236377.

W. E. Barlow et al. Prospective breast cancer risk prediction model for women undergoing screening mammography. *Journal of the National Cancer Institute*, 98(17):1204–1214, Sep 2006.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. doi: 10.1198/016214505000000907. URL https://doi.org/10.1198/016214505000000907.

A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997. ISSN 0031-3203.

G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950. URL https://api.semanticscholar.org/CorpusID:122906757.

A. M. Carrington, D. G. Manuel, P. W. Fieguth, T. Ramsay, V. Osmani, B. Wernly, C. Bennett, S. Hawken, O. Magwood, Y. Sheikh, M. McInnes, and A. Holzinger. Deep roc analysis and auc as balanced average accuracy, for improved classifier selection, audit and explanation. *IEEE Trans Pattern Anal Mach Intell*, 45(1):329–341, Jan 2023. ISSN 1939-3539 (Electronic); 0098-5589 (Linking). doi: 10.1109/TPAMI.2022.3145392.

T. Dimitriadis, T. Gneiting, A. I. Jordan, and P. Vogel. Evaluating probabilistic classifiers: The triptych. *International Journal of Forecasting*, 40(3):1101–1122, 2024. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2023.09.007. URL https://www.sciencedirect.com/science/article/pii/S0169207023000997.

C. Drummond and R. C. Holte. Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, 65(1):95–130, 2006. doi: 10.1007/s10994-006-8199-5. URL https://doi.org/10.1007/s10994-006-8199-5.

T. Fawcett and A. Niculescu-Mizil. Pav and the roc convex hull. *Machine Learning*, 68(1):97–106, 2007. doi: 10.1007/s10994-007-5011-0. URL https://doi.org/10.1007/s10994-007-5011-0.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368, 1922.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

I. J. Good. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 14(1):107–114, 1952. ISSN 00359246. URL http://www.jstor.org/stable/2984087.

D. M. Green and J. A. Swets. *Signal detection theory and psychophysics*. Wiley, New York, 1966.

H. V. Hance. *The optimization and analysis of systems for the detection of pulse signals in random noise*. Sc.d. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1951. URL http://hdl.handle.net/1721.1/12189. Bibliography: leaves 141-143.

D. J. Hand. Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine Learning*, 77(1):103–123, 2009. doi: 10.1007/s10994-009-5119-5. URL https://doi.org/10.1007/s10994-009-5119-5.

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982. ISSN 0033-8419.

J. Hernández-Orallo, P. Flach, and C. Ferri. Brier curves: a new cost-based visualisation of classifier performance. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 585–592, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

J. Hernández-Orallo, P. Flach, and C. Ferri. A unified view of performance metrics: translating threshold choice into expected classification loss. *J. Mach. Learn. Res.*, 13(1):2813–2869, 10 2012.

J. Huang and C. Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17:299–310, 2005. doi: 10.1109/TKDE.2005.50.

Y. Jiang, C. E. Metz, and R. M. Nishikawa. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology*, 201(3):745–750, Dec 1996. ISSN 0033-8419 (Print); 0033-8419 (Linking). doi: 10.1148/radiology.201.3.8939225.

N. Kallus and A. Zhou. *The fairness of risk scores beyond classification: bipartite ranking and the xAUC metric*. Curran Associates Inc., Red Hook, NY, USA, 2019.

K. Kwegyir-Aggrey, M. Gerchick, M. Mohan, A. Horowitz, and S. Venkatasubramanian. The misuse of auc: What high impact risk assessment gets wrong. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 1570–1583, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3593013.3594100. URL https://doi.org/10.1145/3593013.3594100.

J. McCarthy. Measures of the value of information. *Proceedings of the National Academy of Sciences*, 42(9):654–655, 1956. doi: 10.1073/pnas.42.9.654. URL https://www.pnas.org/doi/abs/10.1073/pnas.42.9.654.

D. K. McClish. Analyzing a portion of the roc curve. *Med Decis Making*, 9(3):190–195, 1989. ISSN 0272-989X (Print); 0272-989X (Linking). doi: 10.1177/0272989X8900900307.

C. E. Metz. Basic principles of roc analysis. *Semin Nucl Med*, 8(4):283–98, Oct 1978. doi: 10.1016/s0001-2998(78)80014-2.

D. North. An analysis of the factors which determine signal/noise discrimination in pulsed-carrier systems. *Proceedings of the IEEE*, 51(7):1016–1027, 1963. doi: 10.1109/PROC.1963.2383.

D. O. North. An analysis of the factors which determine signal-noise discrimination in pulse carrier systems. Technical Report PTR-6C, RCA Laboratories Division, Radio Corp. of America, 6 1943.

. Peterson, W. Wesley and T. G. Birdsall. *The theory of signal detectability*. Michigan. University. Department of Electrical Engineering. Electronic Defense Group. Technical report; no. 13. Engineering Research Institute, Ann Arbor, 1953.

W. W. Peterson, T. G. Birdsall, and W. C. Fox. The theory of signal detectability. *Trans. IRE Prof. Group Inf. Theory*, 4:171–212, 1954. URL https://api.semanticscholar.org/CorpusID:206727190.

S. Pfohl, Y. Xu, A. Foryciarz, N. Ignatiadis, J. Genkins, and N. Shah. Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1039–1052, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533166. URL https://doi.org/10.1145/3531146.3533166.

J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. 1999. URL https://api.semanticscholar.org/CorpusID:56563878.

L. Savage. *The Foundations of Statistics*. Wiley Publications in Statistics, 1954.

L. J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2284229.

M. J. Schervish. A general method for comparing probability assessors. *The Annals of Statistics*, 17(4):1856–1879, 1989. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2241668.

Y. Shen. *Loss functions for binary classification and class probability estimation*. PhD thesis, 2005. URL https://www.proquest.com/dissertations-theses/loss-functions-binary-classification-class/docview/305411117/se-2. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-03-03.

E. H. Shuford, A. Albert, and H. Edward Massengill. Admissible probability measurement procedures. *Psychometrika*, 31(2):125–145, 1966. doi: 10.1007/BF02289503. URL https://doi.org/10.1007/BF02289503.

S. Siegert. Simplifying and generalising murphy's brier score decomposition. *Quarterly Journal of the Royal Meteorological Society*, 143(703):1178–1183, 2017. doi: https://doi.org/10.1002/qj.2985. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2985.

K. A. Spackman. Signal detection theory: valuable tools for evaluating inductive learning. In *Proceedings of the Sixth International Workshop on Machine Learning*, pages 160–163, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc. ISBN 1558600361.

E. W. Steyerberg and A. J. Vickers. Decision curve analysis: a discussion. *Med Decis Making*, 28(1):

146–149, 2008. ISSN 0272-989X (Print); 0272-989X (Linking). doi: 10.1177/0272989X07312725.

J. Swets and T. Birdsall. The human use of information–iii: Decision-making in signal detection and recognition situations involving multiple alternatives. *IRE Transactions on Information Theory*, 2(3):138–165, 1956. doi: 10.1109/TIT.1956.1056799.

W. P. Tanner, J. A. Swets, and H. W. Welch. A new theory of visual detection. Technical Report UMR3825, University of Michigan, 1953. URL https://hdl.handle.net/2027.42/7893. Engineering Technical Report.

J. M. C. Thompson and G. W. Brier. The economic utility of weather forecasts. *Monthly Weather Review*, 83:249–253, 1955. URL https://api.semanticscholar.org/CorpusID:122117332.

B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, E. W. Steyerberg, P. Bossuyt, G. S. Collins, P. Macaskill, D. J. McLernon, K. G. M. Moons, E. W. Steyerberg, A. J. Vickers, O. behalf of Topic Group 'Evaluating diagnostic tests, and prediction models'of the STRATOS initiative. Calibration: the achilles heel of predictive analytics. *BMC Medicine*, 17(1):230, 2019. doi: 10.1186/s12916-019-1466-7. URL https://doi.org/10.1186/s12916-019-1466-7.

A. J. Vickers and E. B. Elkin. Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006. doi: 10.1177/0272989X06295361. URL https://doi.org/10.1177/0272989X06295361. PMID: 17099194.

A. J. Vickers, B. Van Calster, and E. W. Steyerberg. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ*, 352:i6, Jan 2016. ISSN 1756-1833 (Electronic); 0959-8138 (Print); 0959-8138 (Linking). doi: 10.1136/bmj.i6.

A. J. Vickers, B. van Calster, and E. W. Steyerberg. A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic and Prognostic Research*, 3(1):18, 2019. doi: 10.1186/s41512-019-0064-7. URL https://doi.org/10.1186/s41512-019-0064-7.

J. von Neumann, O. Morgenstern, and A. Rubinstein. *Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition)*. Princeton University Press, 1944. ISBN 9780691130613. URL http://www.jstor.org/stable/j.ctt1r2gkx.

W. Yang, J. Jiang, E. M. Schnellinger, S. E. Kimmel, and W. Guo. Modified Brier score for evaluating prediction accuracy for binary outcomes. *Statistical Methods in Medical Research*, 31(12):2287–2296, 2022.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004. ISSN 00905364. URL http://www.jstor.org/stable/3448494.

K. Zhu, Y. Zheng, and K. C. G. Chan. Weighted brier score – an overall summary measure for risk prediction models with clinical utility consideration, 2024. URL https://arxiv.org/abs/2408.01626.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]

(b) The license information of the assets, if applicable. [Yes]

(c) New assets either in the supplemental material or as a URL, if applicable. [Yes]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix: Proofs and Further Details

A HISTORICAL DEVELOPMENT OF DEPENDENT VS INDEPENDENT METRICS

Dependent Decisions. The idea of plotting false positive rate (FPR) and true positive rate (TPR) against decision thresholds, but not against each other, originates in World War II-era work on signal detection theory North (1943, 1963); Hance (1951). The ROC plot itself emerged in post-war work on radar signal detection theory Peterson and Birdsall (1953); Peterson et al. (1954) and spread to psychological signal detection theory through the work of Tanner and Swets Tanner et al. (1953); Swets and Birdsall (1956). From there, the ROC plot was adopted in radiology, where detecting blurry tumors on X-rays was recognized as a psychophysical detection problem Metz (1978). The use of the Area under Receiver Operating Characteristics Curve (AUC-ROC) began with psychophysics Green and Swets (1966) and was particularly embraced by the medical community Metz (1978); Hanley and McNeil (1982). From there, as AUC-ROC gained traction in medical settings, Spackman (1989) proposed its introduction to broader machine learning applications. At a time when popular techniques like Naive Bayes and SVMs were wildly miscalibrated, being able to report pure ranking performance was useful to methods developers. This idea was further popularized by Bradley (1997) and extended in studies examining connections between AUC and accuracy Huang and Ling (2005). Although more widespread use of calibration techniques such as Platt scaling (Platt, 1999) and isotonic regression (Ayer et al., 1955) (which is equivalent to computing the convex hull of the ROC curve (Fawcett and Niculescu-Mizil, 2007)) now facilitates reliable probability estimates, AUC-ROC remains prevalent, especially in clinical settings, due to its robustness to score miscalibration. There have been consistent critiques of the lack of calibration information in the ROC curve Van Calster et al. (2019), Kwegyir-Aggrey et al. (2023).

Over the years there have been a few attempts to evaluate area under only a portion of the ROC curve McClish (1989), Carrington et al. (2023), but they haven't produced a clear decision-theoretic interpretation.

Independent Decisions. The idea of maximizing expected value was popularized by von Neumann et al. (1944). The link between incentive-compatible forecast metrics (e.g., Brier score Brier (1950), log loss Good (1952); McCarthy (1956)) and expected regret was formalized by Shuford et al. (1966), clarified by Savage (1971), and later connected to regret curves by Schervish (1989). These ideas were revisited and extended through Brier Curves (Adams and Hand, 1999; Drummond and Holte, 2006; Hernández-Orallo et al., 2011) and Beta-distribution modeling of cost uncertainty (Zhu et al., 2024). Hand (2009) and Hernández-Orallo et al. (2012) showed that AUC-ROC can be interpreted as a cost-weighted average regret, especially under calibrated or quantile-based forecasts. Separately, Vickers and Elkin (2006); Steyerberg and Vickers (2008) and Assel et al. (2017) introduced decision curve analysis (DCA) as a threshold-restricted net benefit visualization, arguing it offers more clinical relevance than Brier-based aggregation. Recent work has further examined the decomposability of Brier and log loss into calibration and discrimination components (Shen, 2005; Siegert, 2017; Dimitriadis et al., 2024), providing guidance on implementation and visualization.

B ANGSTROM COST/LOSS MODEL

Our consequentialist framework evaluates binary decisions using (state, action, value) triplets, connecting the true class (y), predicted label (a), and outcome value ($V(y, a)$). We use a cost structure originating with Angstrom (1922) and used by Thompson and Brier (1955), implicitly based on the idea that negative labels are the majority class, and that the baseline true negative should have a 0 value. The cost/loss parameterization assumes that the failure to detect a positive case (i.e. a false negative) has a downstream cost of L. It then makes the assumption that any positive prediction incurs an immediate cost C (e.g., treatment cost) which does not depend on the true class. This is usually true, and allows us to compare outcomes across different scenarios where different treatment costs are incurred. Where it is false, the formalism is still expressive enough to fit any contingency table we might

want, but comparisons across different treatment costs will no longer be valid.

$V(y,a)$	$a=0$	$a=1$
$y=0$	0 (True Neg)	C (False Pos)
$y=1$	L (False Neg)	C (True Pos)

We assume $C > 0$ because otherwise assigning the label 1 strictly dominates using a classifier. We assume $L > C$ because otherwise assigning the label 0 strictly dominates using a classifier. As such, the ratio C/L is constrained to lie in the interval $(0, 1)$. We want clear semantics for comparing different cost ratios, so we divide thorough by L. This is equivalent to assuming that the loss from false negatives is the same in all scenarios, and only the treatment cost C varies. We then borrow from online learning the idea of measuring the optimality gap rather than the absolute outcome.

$V(y,a)$	$a=0$	$a=1$
$y=0$	0 (True Neg)	c (False Pos)
$y=1$	$1-c$ (False Neg)	0 (True Pos)

Note that in an online setting, regret refers to an optimality gap to the best possible classifier training algorithm. Our setting is offline, so we borrow the terminology to refer to the optimality gap with the best possible classifier.

C REGRET

Theorem C.1 (Optimal Threshold)**.**

$$\arg\min_{\tau} R(\kappa, \pi, \tau, c, \mathcal{D}) = c$$

Proof. We find the stationary points as follows:

$$R(\kappa, \pi, \tau, c, \mathcal{D}) = c \cdot (1-\pi) \cdot (1 - F_0(\tau)) \ + \ (1-c) \cdot \pi \cdot F_1(\tau)$$

$$\begin{aligned} 0 &= \frac{\partial R(\kappa, \pi, \tau, c, \mathcal{D})}{\partial \tau} \\ &= -c(1-\pi) \cdot f_0(\tau) + (1-c)\pi \cdot f_1(\tau) \end{aligned}$$

using the identity $\frac{d}{d\tau} F_0(\tau) = f_0(\tau)$ and $\frac{d}{d\tau} F_1(\tau) = f_1(\tau)$. This gives the condition:

$$c(1-\pi) f_0(\tau) = (1-c)\pi f_1(\tau)$$

Rewriting this in terms of conditional probabilities:

$$\frac{\pi f_1(\tau)}{\pi f_1(\tau) + (1-\pi) f_0(\tau)} = c \quad \Rightarrow \quad c = \frac{P(y=1, s(x)=\tau)}{P(s(x)=\tau)} = P(y=1 \mid s(x)=\tau)$$

□

This will be a minimum if we have convexity, so that

$$\frac{\partial}{\partial \tau} P(y=1 \mid s(x)=\tau) > 0.$$

If the scoring function $s(x)$ is calibrated, then:

$$P(y=1 \mid s(x)=\tau) = \tau,$$

which gives us convexity and therefore,

$$c = \tau.$$

Theorem C.2 (Accuracy as a function of Regret)**.**

$$\textit{Accuracy}(\kappa, \mathcal{D}) = 1 - 2 \cdot R(\kappa, \pi, c = 1/2, \tau, \mathcal{D})$$

Proof.

$$
\begin{aligned}
\text{Accuracy}(\kappa, \mathcal{D}) &\triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{\kappa(x_i;\tau) = y_i\} \\
&= P\big(\kappa(x;\tau) = y\big) \\
&= P\big(\kappa(x;\tau) = 0,\, y = 0\big) \;+\; P\big(\kappa(x;\tau) = 1,\, y = 1\big) \\
&= P(y=0)\, P\big(\kappa(x;\tau) = 0 \mid y = 0\big) \;+\; P(y=1)\, P\big(\kappa(x;\tau) = 1 \mid y = 1\big) \\
&= (1-\pi)\, P(s(x) < \tau \mid y = 0) \;+\; \pi\, P(s(x) \geq \tau \mid y = 1) \\
&= (1-\pi)\, F_0(\tau) \;+\; \pi\Big(1 - F_1(\tau)\Big), \\
&= 1 - \Big((1-\pi)\Big(1 - F_0(\tau)\Big) \;+\; \pi\, F_1(\tau)\Big) \\
&= 1 - 2\Big(\tfrac{1}{2}\Big((1-\pi)\Big(1 - F_0(\tau)\Big) \;+\; \pi\, F_1(\tau)\Big)\Big) \\
&= 1 - 2\; R(\kappa, \pi, c = \frac{1}{2}, \tau, \mathcal{D})
\end{aligned}
$$

□

D APPENDIX: BOUNDED THRESHOLD MIXTURES

The overall plan of this proof is to first use integration by parts to prove an equivalence between pointwise loss functions integrated over the distribution of data, and weighted ℓ^0 loss functions integrated over an interval of costs.

D.1 Lemmas

Lemma D.1 (Positive Class)**.** *Let $0 < a < b < 1$, and let $L(x)$ be a pointwise loss function for the positive class.*

$$
\int_{s=0}^{s=1} \Big(L(\max(a, \min(b, s))) - L(\max(a, \min(b, 1)))\Big)\; dF_1(s) = \int_{c=a}^{c=b} -\frac{dL(c)}{dc}\; F_1(c)\; dc
$$

The proof will simply be integration by parts, with some careful handling of the limits of integration.

Proof.

$$
\begin{aligned}
&\int_{s=0}^{s=1} \Big(L(\max(a, \min(b, s))) - L(\max(a, \min(b, 1)))\Big)\; dF_1(s) \\
&= \int_{s=0}^{s=1} \Big(L(\max(a, \min(b, s))) - L(b)\Big)\; dF_1(s) \\
&= \int_{s=0}^{s=b} \Big(L(\max(a, s)) - L(b)\Big)\; dF_1(s) \\
&= \int_{s=a}^{s=1} \Big(\int_{c=\max(a,s)}^{c=b} -\frac{dL(c)}{dc}\; dc\Big)\; dF_1(s) \\
&= \int_{c=a}^{c=b} \Big(\int_{s=0}^{s=c} dF_1(s)\Big) - \frac{dL(c)}{dc}\; dc \\
&= \int_{c=a}^{c=b} -\frac{dL(c)}{dc}\Big(F_1(c) - F_1(0)\Big)\; dc \\
&= \int_{c=a}^{c=b} -\frac{dL(c)}{dc}\; F_1(c)\; dc
\end{aligned}
$$

□

Lemma D.2 (Negative Class)**.** *Let $0 < a < b < 1$, and let $L(x)$ be a pointwise loss function for the negative class.*

$$\int_{s=0}^{s=1} \Big(L(\max(a, \min(b, s))) - L(\max(a, \min(b, 0))) \Big) \, dF_0(s) = \int_{c=a}^{c=b} \frac{dL(c)}{dc} \, (1 - F_0(c)) \, dc$$

The proof will simply be integration by parts, with some careful handling of the limits of integration.

Proof.

$$\begin{aligned}
&\int_{s=0}^{s=1} \Big(L(\max(a, \min(b, s))) - L(\max(a, \min(b, 0))) \Big) \, dF_0(s) \\
&= \int_{s=0}^{s=1} \Big(L(\max(a, \min(b, s))) - L(a) \Big) \, dF_0(s) \\
&= \int_{s=a}^{s=1} \Big(L(\min(b, s)) - L(a) \Big) \, dF_0(s) \\
&= \int_{s=a}^{s=1} \Big(\int_{c=a}^{c=\min(b,s)} \frac{dL(c)}{dc} \, dc \Big) \, dF_0(s) \\
&= \int_{c=a}^{c=b} \Big(\int_{s=c}^{s=1} dF_0(s) \Big) \frac{dL(c)}{dc} \, dc \\
&= \int_{c=a}^{c=b} \frac{dL(c)}{dc} \Big(F_0(1) - F_0(c) \Big) \, dc \\
&= \int_{c=a}^{c=b} \frac{dL(c)}{dc} \Big(1 - F_0(c) \Big) \, dc
\end{aligned}$$

□

Lemma D.3 (Combining Classes)**.**

$$\begin{aligned}
&\mathbb{E}_{x,y\sim\mathcal{D}} \Big[L(|y - \max(a, \min(b, s(x))|)) - L(|y - \max(a, \min(b, y))|) \Big] \\
&= \int_{c=a}^{c=b} \Big(\frac{dL(c)}{dc} \, (1-\pi)(1 - F_0(c)) - \frac{dL(1-c)}{dc} \, \pi \, F_1(c) \Big) \, dc
\end{aligned}$$

Proof. The proof is a simple application of Lemma D.1 and Lemma D.2.

$$\begin{aligned}
&\mathbb{E}_{x,y\sim\mathcal{D}} \Big[L(|y - \max(a, \min(b, s(x))|)) - L(|y - \max(a, \min(b, y))|) \Big] \\
&= (1-\pi) \int_{s=0}^{s=1} \Big[L(\max(a, \min(b, s(x)))) - L(\max(a, \min(b, 0))) \Big] \, dF_0(s) \\
&\quad + \pi \int_{s=0}^{s=1} \Big[L(1 - \max(a, \min(b, s(x)))) - L(1 - \max(a, \min(b, 1))) \Big] \, dF_1(s) \\
&= (1-\pi) \int_{c=a}^{s=b} \frac{dL(c)}{dc} [1 - F_0(c)] \, dc - \pi \int_{c=a}^{s=b} \frac{dL(1-c)}{dc} F_1(c) \, dc \\
&= \int_{c=a}^{s=b} \Big(\frac{dL(c)}{dc} (1-\pi)[1 - F_0(c)] - \frac{dL(1-c)}{dc} \pi \, F_1(c) \Big) \, dc
\end{aligned}$$

□

Proposition D.4 (Propriety of Bounded-Threshold Scoring Rules)**.** *Let $L(s, y)$ be a strictly proper scoring rule, and let $\mathrm{clip}_{[a,b]}(s) = \max(a, \min(b, s))$. Then the bounded-threshold scoring rule*

$$L_{[a,b]}(s, y) \;=\; L\big(\mathrm{clip}_{[a,b]}(s),\, y\big) \;-\; L\big(\mathrm{clip}_{[a,b]}(y),\, y\big)$$

is proper but not strictly proper.

Proof. Let $S(z,q) = \mathbb{E}_{y\sim q}[L(z,y)]$ be the expected loss under true probability $q = P(y=1)$. Because L is strictly proper, $S(z,q)$ is uniquely minimised at $z = q$.

The expected bounded loss is $S(\text{clip}_{[a,b]}(s), q)$ minus a constant independent of s. Assuming differentiability, the chain rule yields:

$$\frac{\partial}{\partial s} S(\text{clip}_{[a,b]}(s), q) \;=\; \left.\frac{\partial S(z,q)}{\partial z}\right|_{z=\text{clip}_{[a,b]}(s)} \cdot \mathbf{1}_{(a<s<b)}$$

Inside the interval (a,b), the derivative matches the strictly proper rule, vanishing if and only if $s = q$. Outside the interval, the indicator is zero, meaning the expected loss is perfectly flat.

Thus, if $q \in [a,b]$, the expected bounded loss is uniquely minimised at $s = q$. If $q < a$, the unconstrained loss strictly increases for $z > a$, meaning the bounded loss strictly increases for $s > a$ and is flat for $s < a$. Any prediction $s \le a$ (including $s = q$) achieves the identical minimal loss. A symmetric argument applies for $q > b$.

Because setting $s = q$ always achieves the global minimum, the rule is proper. Because the minimiser is not unique when $q \notin [a,b]$, the rule is not strictly proper. □

D.2 Specific Loss Functions

Theorem D.5 (Bounded Threshold Brier Score)**.** *For a classifier κ, the integral of regret over the interval [a,b] is a the Brier Score of the predictions clipped to [a,b] minus the Brier Score of the true labels clipped to [a,b].*

$$\mathop{\mathbb{E}}_{c\sim \mathit{Uniform}(a,b)} R^*_\kappa(c) = \frac{1}{b-a}\left[\mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[(y - \max(a,\min(b,s(x))))^2] \;-\; \mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[(y-\max(a,\min(b,y)))^2]\right]$$

Proof. Let $L(x) = x^2$ be the quadratic pointwise loss. Then $\frac{dL(c)}{dc} = 2c$ and $-\frac{dL(1-c)}{dc} = 2(1-c)$.

$$\begin{aligned}
&\frac{1}{b-a}\left[\mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[(y - \max(a,\min(b,s(x))))^2] \;-\; \mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[(y-\max(a,\min(b,y)))^2]\right]\\
&= \frac{1}{b-a}\mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}\left[(y - \max(a,\min(b,s(x))))^2 \;-\; (y-\max(a,\min(b,y)))^2\right]\\
&\text{Using Lemma D.3, we have}\\
&= \frac{1}{b-a}\int_{c=a}^{c=b}\Big(2c\,(1-\pi)(1-F_0(c)) + 2(1-c)\,\pi\,F_1(c)\Big)\,dc\\
&= \frac{1}{b-a}\int_{c=a}^{c=b} 2R^*_\kappa(c)\,dc\\
&= 2\mathop{\mathbb{E}}_{c\sim\text{Uniform}(a,b)} R^*_\kappa(c)
\end{aligned}$$

□

Theorem D.6 (Bounded Threshold Log Loss)**.** *For a classifier κ, the integral of regret over the interval [a,b] with log-odds uniform weighting is a the Log Loss of the predictions clipped to [a,b] minus the Log Loss of the true labels clipped to [a,b].*

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\ell\sim \mathit{Uniform}\left(\log\frac{a}{1-a},\,\log\frac{b}{1-b}\right)}\left[R^*_\kappa(c = \frac{1}{1+\exp -\ell})\right]\\
&= \frac{1}{\log\frac{b}{1-b} - \log\frac{a}{1-a}}\Bigg[\mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[\log(1 - |y - \max(a,\min(b,s(x)))|)]\\
&\qquad\qquad - \mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[\log(1-|y-\max(a,\min(b,y))|)]\Bigg]
\end{aligned}$$

Proof. Let $L(x) = \log(1-x)$ be the logarithmic pointwise loss. Then $\frac{dL(c)}{dc} = \frac{1}{1-c}$ and $-\frac{dL(1-c)}{dc} = \frac{1}{c}$.

$$\frac{1}{\log\frac{b}{1-b} - \log\frac{a}{1-a}}\Bigg[\mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[\log(1-|y-\max(a,\min(b,s(x)))|)] - \mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[\log(1-|y-\max(a,\min(b,y))|)]\Bigg]$$

$$= \frac{1}{\log\frac{b}{1-b} - \log\frac{a}{1-a}} \mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}\Big[\log(1-|y-\max(a,\min(b,s(x)))|) \;-\; \log(1-|y-\max(a,\min(b,y))|)\Big]$$

Using Lemma D.3, we have

$$= \frac{1}{\log\frac{b}{1-b} - \log\frac{a}{1-a}} \int_{c=a}^{c=b} \Big(\frac{1}{1-c}\,(1-\pi)(1-F_0(c)) + \frac{1}{c}\,\pi\,F_1(c)\Big)\,dc$$

$$= \frac{1}{\log\frac{b}{1-b} - \log\frac{a}{1-a}} \int_{c=a}^{c=b} \Big(c\,(1-\pi)(1-F_0(c)) + (1-c)\,\pi\,F_1(c)\Big)\,\frac{dc}{c(1-c)}$$

$$= \frac{1}{\log\frac{b}{1-b} - \log\frac{a}{1-a}} \int_{c=a}^{c=b} R^*_\kappa(c)\,\frac{dc}{c(1-c)}$$

Now we do a change of variables $\ell = \log\frac{c}{1-c}$, $\frac{d\ell}{dc} = \frac{1}{c(1-c)}$.

$$\frac{1}{\log\frac{b}{1-b} - \log\frac{a}{1-a}} \int_{c=a}^{c=b} R^*_\kappa(c)\,\frac{dc}{c(1-c)}$$

$$= \frac{1}{\log\frac{b}{1-b} - \log\frac{a}{1-a}} \int_{\ell=\log\frac{a}{1-a}}^{\ell=\log\frac{b}{1-b}} R^*_\kappa(c = \frac{1}{1+\exp-\ell})\,d\ell$$

$$= \mathop{\mathbb{E}}_{\ell\sim\text{Uniform}\left(\log\frac{a}{1-a},\,\log\frac{b}{1-b}\right)}\left[R^*_\kappa(c = \frac{1}{1+\exp-\ell})\right]$$

□

D.3 Shifted Brier Score

Definition D.7 (Score Adjustment)**.** Let $s \in (0,1)$ be a predicted probability and let $\mu \in (0,1)$ denote a reference class probability. Define the *score adjustment function* $M : (0,1)\times(0,1) \to (0,1)$ as:

$$M(s,\mu) \triangleq \frac{1}{1+\exp\left(\log\left(\frac{s}{1-s}\right) - \log\left(\frac{\mu}{1-\mu}\right)\right)}$$

That is, $M(s,\mu)$ adjusts the predicted log-odds of s by centering it around the log-odds of μ.

We extend M to the boundary values $s \in \{0,1\}$ by defining:

$$\lim_{s\to a} M(s,\mu) = a \quad \text{for } a \in \{0,1\}$$

Proposition D.8 (Inverse of Score Adjustment)**.** $M(M(s,\mu),-\mu) = s$

Lemma D.9. *Let $G(s,y) : [0,1]\times\{0,1\} \to [0,1]$ be the cumulative distribution function of s for either the*

positive or negative class, and let $G(0, y) = 0$ and $G(1, y) = 1$.

$$\begin{aligned}
&\int_{s=0}^{s=1} (y - M(s, -\mu))^2 \, dG(s, y) \\
&= \int_{s=0}^{s=1} \int_{c=M(y,-\mu)}^{c=M(s,-\mu)} -2(y - c) \, dc \, dG(s, y) \\
&= -2 \int_{c=0}^{c=1} (y - c) \int_{s=M(c,+\mu)}^{s=M(1-y,+\mu)} dG(s, y) \, dc \\
&= -2 \int_{c=0}^{c=1} (y - c) \, [G(M(1 - y, +\mu), y) - G(M(c, +\mu), y)] dc \\
&= -2 \int_{c=0}^{c=1} (y - c) \, [1 - y - G(M(c, +\mu), y)] dc
\end{aligned}$$

Theorem D.10. *If we define a new score such that $s'(x) = M(s(x), \mu)$, then*

$$\mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}} (y - s'(x))^2 = \mathop{\mathbb{E}}_{c\sim Uniform(0,1)} R^*_\kappa(M(c, \mu))$$

Proof. Let $G(s, y) = \begin{cases} F_0(s) & \text{if } y = 0 \\ F_1(s) & \text{if } y = 1 \end{cases}$. Then using Lemma D.9, twice we have:

$$\begin{aligned}
&\int_{s=0}^{s=1} (0 - M(s(x), +\mu))^2 \, dF_0(s) + \int_{s=0}^{s=1} (1 - M(s(x), +\mu))^2 \, dF_1(s) \\
&= 2 \int_{c=0}^{c=1} (0 - c)[1 - 0 - F_0(M(c, -\mu))] + (1 - c)[1 - 1 - F_1(M(c, -\mu))] \, dc \\
&= -2 \int_{c=0}^{c=1} c[1 - F_0(M(c, -\mu))] + (1 - c)[F_1(M(c, -\mu))] \, dc \\
&= -2 \int_{c=0}^{c=1} R^*_\kappa(M(c, -\mu)) \, dc
\end{aligned}$$

□

E AVERAGE NET BENEFIT AS A PROPER SCORING RULE

Theorem E.1 (Restatement of Theorem 3.2)**.**

$$\mathrm{NB}(c) = \pi - \frac{R^*_\kappa(c)}{1 - c}$$

Proof. Once we express the net benefit definition given in Vickers et al. (2019) using the terminology of this paper and arrange the terms, the result follows.

$$\begin{aligned}
NB(c) &= \text{sensitivity} \times \text{prevalence} - (1 - \text{specificity}) \times (1 - \text{prevalence}) \times \frac{\tau}{1 - \tau} \\
&= (1 - F_1(\tau))\pi - (1 - F_0(\tau))(1 - \pi)\frac{\tau}{1 - \tau} \\
&= \frac{1}{1 - \tau}\Big[(1 - \tau)(1 - F_1(\tau))\pi - (1 - F_0(\tau))(1 - \pi)\tau\Big] \\
&= \frac{1}{1 - c}\Big[(1 - c)(1 - F_1(c))\pi - (1 - F_0(c))(1 - \pi)c\Big] \\
&= \pi - \frac{1}{1 - c}\Big[(1 - c)F_1(c)\pi + (1 - F_0(c))(1 - \pi)c\Big] \\
&= \pi - \frac{R^*_\kappa(c)}{1 - c}
\end{aligned}$$

□

Theorem E.2 (Restatement of Theorem 3.7)**.** *Let* $L(x,y) = \begin{cases} s(x) & \text{if } y=1 \\ (1-s(x)) - \ln(1-s(x)) & \text{if } y=0 \end{cases}$ *be a pointwise loss.*

For a classifier κ*, the integral of net benefit over the interval [a,b] is the loss for the predictions clipped to [a,b] minus the loss for the true labels clipped to [a,b].*

$$\mathop{\mathbb{E}}_{c\sim Uniform(a,b)} \mathrm{NB}(c) = \pi - \frac{1}{b-a}\left[\mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}} L(max(a,\min(b,s(x))),y) \;-\; \mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}} L(\max(a,\min(b,y)),y)\right]$$

Proof. Note that $\frac{dL(x,y)}{dc} = \begin{cases} 1 & \text{if } y=1 \\ \frac{c}{1-c} & \text{if } y=0 \end{cases}$. Then.

$$\pi - \frac{1}{b-a}\left[\mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[L(\max(a,\min(b,s(x),y)))] \;-\; \mathop{\mathbb{E}}_{(x,y)\in\mathcal{D}}[L(\max(a,\min(b,y)),y)]\right]$$

Using Lemma D.3, we have

$$= \pi - \frac{1}{b-a}\int_{c=a}^{c=b}\left(\frac{dL(c,0)}{dc}\,(1-\pi)(1-F_0(c)) + \frac{dL(c,1)}{dc}\,\pi\,F_1(c)\right)dc$$

$$= \pi - \frac{1}{b-a}\int_{c=a}^{c=b}\frac{R^*_\kappa(c)}{1-c}\,dc$$

$$= \mathop{\mathbb{E}}_{c\sim\mathrm{Uniform}(a,b)} \pi - \frac{R^*_\kappa(c)}{1-c}$$

$$= \mathop{\mathbb{E}}_{c\sim\mathrm{Uniform}(a,b)} \mathrm{NB}(c)$$

□

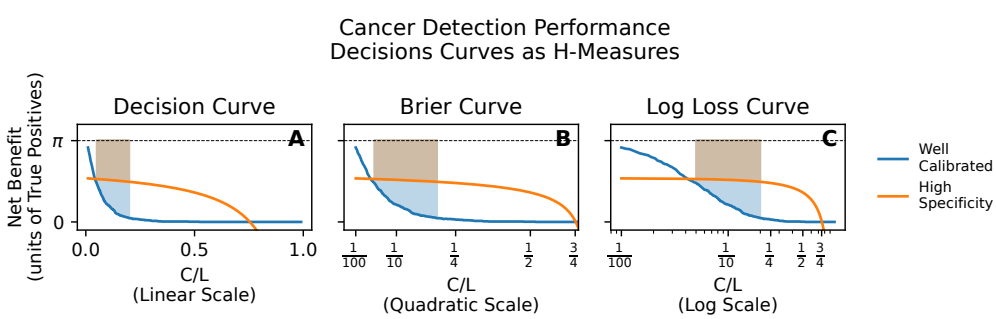

Figure 4: The figure shows the DCA (A), which can be rescaled so that for an interval of cost ratios, the area above the curve and below the prevalence π is equal to the bounded threshold Brier score (B) or bounded threshold log loss (C).

Theorem E.3 (Quadratically Rescaled Decision Curve)**.** *Let* $\phi(c) \triangleq \frac{-(1-c)^2}{2}$ *and therefore* $\frac{d\phi(c)}{dc} = 1-c$*. Note*

that this is invertible on the interval $[0, 1]$.

$$\frac{1}{b-a}\int_{x=\phi(a)}^{x=\phi(b)} \pi - NB(x)\ dx$$

Using Theorem E.2, we have

$$= \frac{1}{b-a}\int_{x=\phi(a)}^{x=\phi(b)} \frac{R^*_\kappa(\phi^{-1}(x))}{1-\phi^{-1}(x)}\ dx$$

$$= \frac{1}{b-a}\int_{c=a}^{c=b} \frac{R^*_\kappa(c)}{1-c}\ (1-c)dc$$

$$= \mathop{\mathbb{E}}_{c\sim Uniform(a,b)} R^*_\kappa(c)$$

Theorem E.4 (Logarithmically Rescaled Decision Curve)**.** *Let* $\phi(c) \triangleq \ln c$ *and therefore* $\frac{d\phi(c)}{dc} = \frac{1}{c}$*. Note that this is invertible on the interval* $(0, 1]$.

$$\frac{1}{\log\frac{b}{1-b} - \log\frac{a}{1-a}}\int_{x=\phi(a)}^{x=\phi(b)} \pi - NB(x)\ dx$$

Using Theorem E.2, we have

$$= \frac{1}{\log\frac{b}{1-b} - \log\frac{a}{1-a}}\int_{x=\phi(a)}^{x=\phi(b)} \frac{R^*_\kappa(\phi^{-1}(x))}{1-\phi^{-1}(x)}\ dx$$

$$= \frac{1}{\log\frac{b}{1-b} - \log\frac{a}{1-a}}\int_{c=a}^{c=b} \frac{R^*_\kappa(c)}{1-c}\ \frac{dc}{c}$$

$$= \mathop{\mathbb{E}}_{\ell\sim Uniform\left(\log\frac{a}{1-a},\ \log\frac{b}{1-b}\right)} \left[R^*_\kappa(c = \frac{1}{1+\exp -\ell})\right]$$

F COMPARISON OF RESULTS FROM Assel et al. (2017)

Assel et al. (2017) provides a table of results showing that the ordering of model quality according to Brier Score fails to match the ordering of model quality according to Net Benefit at a 5% threshold. We reproduce their data generating process and show the same table, with results sorted by the Net Benefit at a 5% threshold for convenience. Note that the ordering by overall Brier Score is indeed quite different. But the ordering by Bounded Threshold Brier Score is almost the same. The single, instructive exception is Assume All Positive, where the Net Benefit at a 20% threshold sharply disagrees with Net Benefit at a 5% threshold. In this case, the Bounded Threshold Brier Score puts some weight on these higher threshold cases.

test	AUC-ROC	Brier	NB 5%	NB 10%	NB 20%	Brier 5%-20%
Highly sensitive	0.73	0.41	0.17	0.15	0.09	0.12
Underestimating risk	0.75	0.15	0.16	0.12	0.06	0.16
Well calibrated	0.75	0.17	0.16	0.12	0.05	0.17
Overestimating risk	0.75	0.20	0.16	0.11	0.03	0.18
Assume all positive	0.50	0.80	0.16	0.11	0.00	0.20
Highly specific	0.73	0.14	0.10	0.10	0.09	0.18
Severely underestimating risk	0.75	0.18	0.09	0.04	0.01	0.29
Assume all negative	0.50	0.20	0.00	0.00	0.00	0.35

Because the data is synthetic, we were able to generate this table based on 1,000,000 samples. The largest standard deviation for any cell in this table is 0.00035, two decimal places further than the precision we have reported for the cells, and as such, standard errors have been omitted from the table cells for clarity.

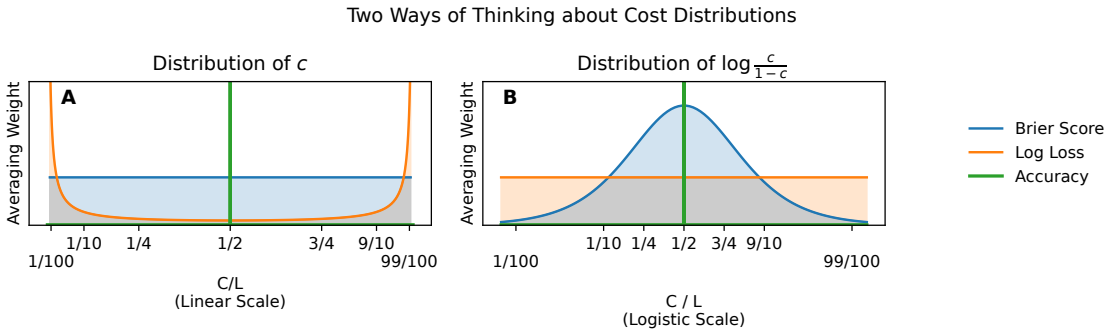

Figure 5: Where the tradeoff between false positives and false negatives is one-to-one, accuracy is a good metric. Where trade-offs are generally moderate, Brier score is a good metric. Where trade-offs are frequently extreme, log loss is better.

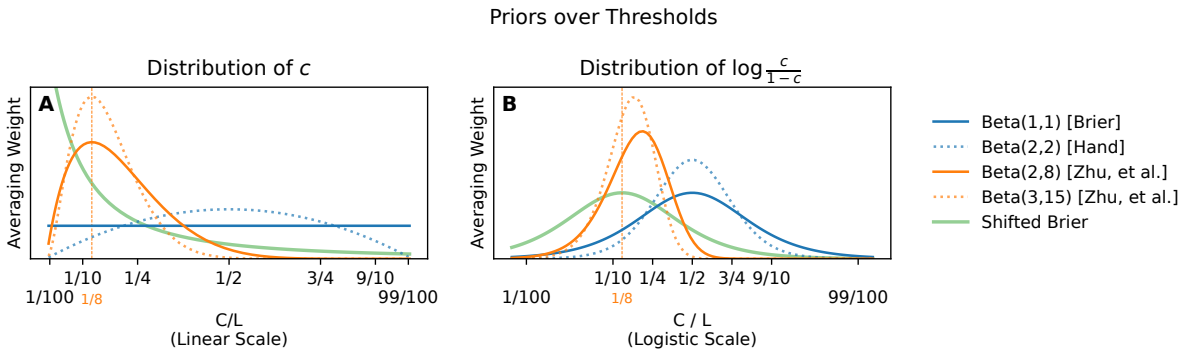

Figure 6: Comparison of cost ratio priors implicit in mixture-of-thresholds metrics. Brier score assumes Beta$(1, 1)$; Hand proposes increasing concentration with Beta$(2, 2)$; Zhu et al. (2024) shifts the mode while inheriting concentration challenges.

G CHOICES OF THRESHOLD DISTRIBUTIONS

Interest in cost-sensitive evaluation during the late 1990s brought renewed attention to the Brier score. Adams and Hand (1999) noted that while domain experts rarely specify exact cost ratios, they can often provide plausible bounds. To improve interpretability, he proposed the LC-Index, which ranks models at each cost ratio and plotting their ranks across the range. Later, Hand (2009) introduced the more general **H-measure**, defined as any weighted average of regret, and recommended a Beta$(2, 2)$ prior to emphasize cost ratios near $c = 0.5$.

Despite its appeal, the H-measure's intuition can be opaque: even the Beta$(1, 1)$ prior used by the Brier score already concentrates mass near parity on the log-odds scale (Figure 6).

Zhu et al. (2024) generalize this idea using asymmetric Beta distributions centered at an expert-specified mode (e.g., Beta$(2, 8)$). However, this raises concerns: the mode is not invariant under log-odds transformation, may be less appropriate than the mean, and requires domain experts to specify dispersion—a difficult task in practice. A simpler alternative is to shift the Brier score to peak at the desired cost ratio via a transformation of the score function $s(x)$, as shown in Appendix D.10.

Rather than infer uncertainty via a prior, Zhu et al. (2024) suggest eliciting threshold bounds directly (e.g., from clinicians). We argue that this approach is better served by constructing explicit threshold intervals rather than encoding beliefs via Beta distributions.

H DEPENDENT DECISIONS WITH FIXED K: EQUIVALENCE OF TOP-K METRICS

When the capacity K is known exactly at model selection time, a metric like Net Benefit@K or True Positives@K (TP@K) is preferable to AUC-ROC. Although Net Benefit@K is conceptually superior because it grounds the

evaluation in actual costs, it induces the exact same model ranking as standard ranking metrics like Precision@K and Recall@K for a fixed dataset, fixed budget K, and fixed cost ratio c.

To see this algebraically, let P be the total number of actual positives in the evaluation dataset. If a model is constrained to predict exactly K positives, let TP denote the number of true positives it selects. The number of false positives is constrained to be $FP = K - TP$, and the number of false negatives left in the unselected pool is $FN = P - TP$.

Under our expected regret formulation, the total cost incurred by the decision rule is:

$$\begin{aligned}\text{Cost} &= c \cdot FP + (1-c) \cdot FN \\ &= c(K - TP) + (1-c)(P - TP) \\ &= cK - cTP + P - TP - cP + cTP \\ &= cK + (1-c)P - TP\end{aligned}$$

Because the cost ratio c, the budget K, and the total positives P are all fixed constants for a given evaluation, the total cost collapses to a constant minus TP.

Consequently, minimizing cost (or maximizing Net Benefit) under a fixed budget is mathematically identical to maximizing TP. Because Precision@$K = TP/K$ and Recall@$K = TP/P$ are merely positive scalar multiples of TP, all of these metrics are affine transformations of one another and will induce the exact same model ranking.

I BOUNDED THRESHOLD AUC-ROC

A natural question to ask is whether the idea of bounding thresholds can be applied to make the AUC-ROC more useful, the same way it makes the Brier score more useful.

To investigate this question, we review the existing attempts to bound the AUC-ROC, and show a new and conceptually clearer way to do it. We then evaluate the strengths and weakness of this approach, and conclude that it is better to use the bounded threshold Brier score instead. For this reason the development of this new bounded threshold AUC-ROC is left to the Appendix rather than the main text.

I.1 Outline

Hand (2009) showed that, under the strong assumption of calibration, the AUC-ROC is equivalent to an integral of expected regret where for our distribution of users' costs, we use the distribution of scores from the model.

We extend this result to show that we can limit the cost distribution over which we take an expectation to those scores from the model that lie in a given interval.

We will first show that we can convert the averages $pAUC_+$ and $pAUC_-$ into expectations over a single distribution. We then combine the terms, and find that they give expected regret plus a couple of simpler terms.

Under calibration, the Bayes-optimal threshold equals the cost ratio c (via $R_{\min}(\tau) = (1-\pi)\tau(1-F_0) + \pi(1-\tau)F_1$ and the first-order condition), so bounding the threshold interval $[a, b]$ is equivalent to bounding plausible cost ratios $c \in [a, b]$ in deployment.

I.2 Background

Our work builds on the ideas of pAUC from McClish (1989) which average TPR based on bounding the FPR, but lacks decision-theoretic interpretation. It is not clear why we would want to limit a range of FPR, and it is not clear why a uniform distribution of FPR produces a meaningful average of TPR.

It also builds on the ideas of PAI from Jiang et al. (1996) which average FPR based on bounding the TPR, but has many of the same problems.

Finally, our work also builds on the idea of Balanced Average Accuracy or Concordant Partial AUC from Carrington et al. (2023) which combines those two averages assuming balanced classes, and adds the idea of bounding by a quantile of the score rather than FPR or TPR.

We add the idea of bounding by the costs encountered in deployment, rather than a quantile of scores, which is a better model of the deployment scenario. Our measure also has a direct interpretation as expected regret, averaged over a single distribution rather than two.

I.3 Notation

Let F_0 and F_1 be the score CDFs for the negative and positive classes with densities $f_0 = \frac{d}{dc}F_0$ and $f_1 = \frac{d}{dc}F_1$. Let

$$F(c) \;=\; (1-\pi)F_0(c) + \pi F_1(c), \qquad f(c) \;=\; \frac{d}{dc}F(c) \;=\; (1-\pi)f_0(c) + \pi f_1(c),$$

be the mixture CDF and density.

Note that the True Positive Rate is the fraction of positive cases where the score exceeds the threshold, so $TPR = 1 - F_1(\tau)$, which can be counterintuitive. Likewise, False Positive Rate is the fraction of negative cases where the score is greater than the threshold, so $FPR = 1 - F_0(\tau)$.

We also define two quantities for the area beneath a subset of the ROC curve as well as the area to the right.

Definition I.1 ($pAUC_+$)**.** This represents the area under a portion of the ROC curve, bounded by the thresholds a and b.

$$pAUC_+(a,b) = \int_{\tau=a}^{b} (1 - F_1(\tau))dF_0(\tau)$$

Note that this is closely related to the partial AUC metric from McClish (1989), although FPR is decreasing in the threshold τ, so the order of the bounds is reversed. Also our function is defined in terms of the threshold value, rather than the FPR value.

$$\text{Partial AUC}(\alpha,\beta) = \int_{FPR=\alpha}^{FPR=\beta} TPR(FPR)\; dFPR = pAUC_+(FPR^{-1}(\beta), FPR^{-1}(\alpha))$$

Definition I.2 ($pAUC_-$)**.** This represents the area to the right of a portion of the ROC curve, bounded by the thresholds a and b.

$$pAUC_-(a,b) = \int_{\tau=a}^{b} F_0(\tau)dF_1(\tau)$$

Note that this is closely related to the partial area index metric from Jiang et al. (1996), although FPR is decreasing in the threshold τ, so the order of the bounds is reversed. Also our function is defined in terms of the threshold value, rather than the TPR value.

$$\text{PAI}(\alpha,\beta) = \int_{TPR=\alpha}^{TPR=\beta} (1 - FPR(TPR))\; dTPR = pAUC_-(TPR^{-1}(\beta), TPR^{-1}(\alpha))$$

I.4 Main Result

We first convert the averages $pAUC_+$ and $pAUC_-$ taken over positive or negative class scores into averages over all scores.

Under calibration, the local shares satisfy $\frac{\pi dF_1(\tau)}{dF(\tau)} = \tau$ and $\frac{(1-\pi)dF_0(\tau)}{dF(\tau)} = 1 - \tau$.

Lemma I.3. *$pAUC_+$ is equivalent to a threshold-weighted average of $1 - F_1(\tau)$ over the scores from the combined distribution of positive and negative examples, up to a constant factor.*

$$\frac{\pi(1-\pi)}{F(b) - F(a)}\; pAUC_+(a,b) = \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}|s(x)\in[a,b]}[\pi\;(1 - s(x))\;(1 - F_1(s(x)))]$$

Proof. We use the assumption of calibration to do a change of variables.

$$\begin{aligned}
pAUC_+(a,b) &= \int_{\tau=a}^{b} (1-F_1(\tau))dF_0(\tau) \\
&= \frac{1}{1-\pi}\int_{\tau=a}^{b} (1-F_1(\tau))\frac{(1-\pi)dF_0(\tau)}{dF(\tau)}dF(\tau) \\
&= \frac{1}{1-\pi}\int_{\tau=a}^{b} (1-F_1(\tau))(1-\tau)dF(\tau) \\
&= \frac{F(b)-F(a)}{1-\pi} \mathop{\mathbb{E}}_{\tau\sim F|\tau\in[a,b]}[(1-F_1(\tau))\ (1-\tau)] \\
&= \frac{F(b)-F(a)}{\pi(1-\pi)} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}|s(x)\in[a,b]}[\pi\ (1-s(x))\ (1-F_1(s(x)))]
\end{aligned}$$

□

An equivalent identity holds for the negative class.

Lemma I.4. *$pAUC_-$ is equivalent to a threshold-weighted average of $F_0(\tau)$ over the scores from the combined distribution of positive and negative examples, up to a constant factor.*

$$\frac{\pi(1-\pi)}{F(b)-F(a)}\ pAUC_-(a,b) = \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}|s(x)\in[a,b]}[(1-\pi)\ s(x)\ F_0(s(x))]$$

Proof. We use the assumption of calibration to do a change of variables.

$$\begin{aligned}
pAUC_-(a,b) &= \int_{\tau=a}^{b} F_0(\tau)dF_1(\tau) \\
&= \frac{1}{\pi}\int_{\tau=a}^{b} F_0(\tau)\frac{\pi dF_1(\tau)}{dF(\tau)}dF(\tau) \\
&= \frac{1}{\pi}\int_{\tau=a}^{b} F_0(\tau)\ \tau\ dF(\tau) \\
&= \frac{F(b)-F(a)}{\pi} \mathop{\mathbb{E}}_{\tau\sim F|\tau\in[a,b]}[F_0(\tau)\ \tau] \\
&= \frac{F(b)-F(a)}{\pi(1-\pi)} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}|s(x)\in[a,b]}[(1-\pi)\ s(x)\ F_0(s(x))]
\end{aligned}$$

□

Since they are now expectations over the same measure, we can combine them.

Theorem I.5. *The weighted sum of $pAUC_+$ and $pAUC_-$ is equivalent to negative expected regret, plus a couple of simpler additive terms.*

$$\frac{\pi(1-\pi)}{F(b)-F(a)}\ [pAUC_+(a,b)+pAUC_-(a,b)] = \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}|s(x)\in[a,b]}[-R^*_\kappa(s(x))] + \pi + (1-2\pi)\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}|s(x)\in[a,b]}[y]$$

Proof. Recall $R^*_\kappa(\tau) = \pi\ (1-s(x))\ F_1(s(x)) + (1-\pi)\ s(x)\ (1-F_0(s(x)))$.

We substitute according to Lemmas I.3 and I.4.

$$
\begin{aligned}
&\frac{\pi(1-\pi)}{F(b)-F(a)}\ [pAUC_+(a,b)+pAUC_-(a,b)] \\
&= \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}|s(x)\in[a,b]}[\overbrace{\pi\ (1-s(x))\ (1-F_1(s(x)))}^{\text{from } pAUC_+}+\overbrace{(1-\pi)\ s(x)\ F_0(s(x))}^{\text{from } pAUC_-}] \\
&= \mathbb{E}[\overbrace{\pi(1-s(x))-\pi\ (1-s(x))\ F_1(s(x))}^{\text{from } pAUC_+}+\overbrace{(1-\pi)s(x)-(1-\pi)\ s(x)\ (1-F_0(s(x)))}^{\text{from } pAUC_-}] \\
&= \mathbb{E}[\overbrace{-\pi\ (1-s(x))\ F_1(s(x))}^{\text{from } pAUC_+}-\overbrace{(1-\pi)\ s(x)\ (1-F_0(s(x)))}^{\text{from } pAUC_-}+\pi(1-s(x))+(1-\pi)s(x)] \\
&= \mathbb{E}[-R^*_\kappa(s(x))+\pi+s(x)-2\pi s(x)]
\end{aligned}
$$

We use the assumption of calibration to simplify the last term. $\mathbb{E}[y|s(x)]=s(x) \implies \mathbb{E}[s(x)]=\mathbb{E}[y]$

$$
\begin{aligned}
&\frac{\pi(1-\pi)}{F(b)-F(a)}\ [pAUC_+(a,b)+pAUC_-(a,b)] \\
&= \mathbb{E}[-R^*_\kappa(s(x))+\pi+s(x)-2\pi s(x)] \\
&= \mathbb{E}[-R^*_\kappa(s(x))+\pi+(1-2\pi)s(x)] \\
&= \mathbb{E}[-R^*_\kappa(s(x))]+\pi+(1-2\pi)\mathbb{E}[y]
\end{aligned}
$$

□

I.5 Generalization of Existing Metrics

Theorem I.6 (Hand's Theorem)**.** *If we don't clip at all (set $(a,b)=(0,1)$), then we get exactly Hand's theorem.*

$$AUC = 1-\frac{1}{2\pi(1-\pi)}\mathbb{E}[R^*_\kappa(c)]$$

Proof.

$$
\begin{aligned}
pAUC_-(a,b)+pAUC_+(a,b) &= \frac{F(b)-F(a)}{\pi(1-\pi)}\left[\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}|s(x)\in[a,b]}[-R^*_\kappa(s(x))]+\pi+(1-2\pi)\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}|s(x)\in[a,b]}[y]\right] \\
&= \frac{1}{\pi(1-\pi)}\left[\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}|s(x)\in[a,b]}[-R^*_\kappa(s(x))]+\pi+(1-2\pi)\pi\right] \\
&= \frac{1}{\pi(1-\pi)}\left[\mathop{\mathbb{E}}_{x,y\sim\mathcal{D}|s(x)\in[a,b]}[-R^*_\kappa(s(x))]+2\pi(1-\pi)\right] \\
AUC+AUC &= -\frac{1}{\pi(1-\pi)}\mathbb{E}[R^*_\kappa(c)]+2 \\
AUC &= 1-\frac{1}{2\pi(1-\pi)}\mathbb{E}[R^*_\kappa(c)]
\end{aligned}
$$

□

Definition I.7. Define $\Delta FPR = FPR(a)-FPR(b) = F_0(b)-F_0(a)$.

Definition I.8. Define $\Delta TPR = TPR(b)-TPR(a) = F_1(b)-F_1(a)$.

Definition I.9. The Balanced Average Accuracy of Concordant Partial AUC is defined in Carrington et al. (2023) as

$$BAA(a,b) = \overbrace{\frac{\Delta FPR}{\Delta TPR+\Delta FPR}}^{\text{Fraction of Negative Cases}}\ \overbrace{\frac{pAUC_+(a,b)}{\Delta FPR}}^{\text{Average Positive Accuracy}} + \overbrace{\frac{\Delta TPR}{\Delta TPR+\Delta FPR}}^{\text{Fraction of Positive Cases}}\ \overbrace{\frac{pAUC_-(a,b)}{\Delta TPR}}^{\text{Average Negative Accuracy}}$$

Lemma I.10. *We can simplify this and express it using our notation*

$$\begin{aligned} BAA(a,b) &= \frac{1}{\Delta TPR + \Delta FPR}\Big[pAUC_+(a,b) + pAUC_-(a,b)\Big] \\ &= \frac{1}{F_1(b) - F_1(a) + F_0(b) - F_0(a)}\Big[pAUC_+(a,b) + pAUC_-(a,b)\Big] \end{aligned}$$

Theorem I.11 (Concordant Partial AUC / Balanced Average Accuracy)**.** *When $\pi = \frac{1}{2}$, then the concordant partial AUC / balanced average accuracy is a linear transformation of expected regret.*

$$BAA(a,b) = 1 - 2 \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}|s(x)\in[a,b]}[R^*_\kappa(s(x))]$$

Proof. We first expand the definition using Theorem I.5.

$$\begin{aligned} BAA &= \frac{1}{F_1(b) - F_1(a) + F_0(b) - F_0(a)}\Big[pAUC_+(a,b) + pAUC_-(a,b)\Big] \\ &= \frac{1}{F_1(b) - F_1(a) + F_0(b) - F_0(a)} \frac{F(b) - F(a)}{\pi(1-\pi)} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}|s(x)\in[a,b]}[-R^*_\kappa(s(x))] + \pi + (1-2\pi) \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}|s(x)\in[a,b]}[y] \\ &= \frac{F(b) - F(a)}{F_1(b) - F_1(a) + F_0(b) - F_0(a)} \frac{1}{\pi(1-\pi)} \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}|s(x)\in[a,b]}[-R^*_\kappa(s(x))] + \pi + (1-2\pi) \mathop{\mathbb{E}}_{x,y\sim\mathcal{D}|s(x)\in[a,b]}[y] \end{aligned}$$

Now consider what happens when $\pi = \frac{1}{2}$.

$$\frac{F(b) - F(a)}{F_1(b) - F_1(a) + F_0(b) - F_0(a)} = \frac{\pi(F_1(b) - F_1(a)) + (1-\pi)(F_0(b) - F_0(a))}{F_1(b) - F_1(a) + F_0(b) - F_0(a)} = \frac{1}{2}$$

$$\frac{1}{\pi(1-\pi)} = 4$$

$$\pi + (1-2\pi)\mathbb{E}[y] = \frac{1}{2} + 0 \cdot \mathbb{E}[y] = \frac{1}{2}$$

Substituting these values back in, we get:

$$\begin{aligned} BAA(a,b) &= \frac{1}{2} \cdot 4 \cdot \left[\mathbb{E}[-R^*_\kappa(s(x))] + \frac{1}{2}\right] \\ &= 1 - 2\mathbb{E}[R^*_\kappa(s(x))] \end{aligned}$$

□

I.6 Discussion

Thus in the case that we know the classifier is calibrated, we can use a combination of the area beneath and to the right of a portion of the AUC curve to get an expected regret. This will still be taking an expectation over the scores of data points whose scores lie in the indicated interval. If there are no such points, then this will not work.

However, this representation raises a conceptual concern: it uses predicted probabilities, intended to estimate outcome likelihoods, as implicit estimates of cost ratios. As Hand (2009) observes, this allows the model to determine the relative importance of false positives and false negatives, *implicitly allowing the model to determine how costly it is to miss a cancer diagnosis, or how acceptable it is to let a guilty person go free.* The model, however, is trained not to encode values or ethical trade-offs, but to estimate outcomes. Using its scores to induce a cost distribution embeds assumptions about harms and preferences that it was never intended to model. While a calibrated model ensures that the mean predicted score equals the class prevalence π, there is no principled reason to treat π as an estimate of the true cost ratio c. Rare outcomes are not necessarily less costly, and often the opposite is true.

This analysis underscores the broader risk of deferring normative judgments, about cost, harm, and acceptability, to statistical models. A more appropriate approach would involve eliciting plausible bounds on cost ratios from domain experts during deployment, rather than allowing the score distribution of a trained model to implicitly dictate them. Finally, this equivalence assumes calibration, which is frequently violated in practice. Metrics that rely on this assumption may be ill-suited for robust evaluation under real-world conditions.

I.7 Recommendation

While bounded AUC-ROC can be shown to represent expected regret under the strong assumption of calibration, its reliance on model scores as implicit cost distributions and the strength of the calibration assumption make it unsuitable for most applications. Practitioners should instead use the Bounded Threshold Brier Score or Log Loss, which allow explicit specification of clinically or operationally relevant cost bounds without these problematic assumptions.

J SKILL SCORES: INTERPRETABLE BOUNDED THRESHOLD METRICS

The newly proposed bounded-threshold metrics have a clear semantic benefit: they explicitly evaluate regret over the relevant range of decision contexts. However, there are two aspects to metric interpretability: (1) clear semantics and (2) easy intuition. Because the calculations are linear, scaling up or recentering our metrics will not change their ordering. This offers us an opportunity.

One of the most helpful guides to interpreting a metric is the existence of anchors with clear meanings. A skill score re-centers a metric so that 0.0 represents the performance of a naive baseline model, 1.0 represents perfect performance, and negative values indicate worse-than-guessing performance. In general, guessing optimally based only on the population positive class prevalence π without any features of the data is a good baseline; this is called the climatological baseline in meteorology.

For example, AUC-ROC of 1.0 means the model predicts everything correctly. The baseline / climatological model would get 0.5, and 0.0 would mean every single label was flipped from the correct value. So a Skill Score for AUC would be $2AUC - 1$.

Brier Score varies between 0.0 and 1.0, with 0.0 being the best possible performance and 1.0 requiring flipping every single label. However, the naive / climatological forecast gets a different score depending on the prevalence - it can be shown to be the Bernoulli variance: $\pi(1-\pi)$. This is the same normalizing factor that occurs in Theorem 4.2 in the connection between expected regret and AUC-ROC.

Thus if BS is the Brier Score, then the Brier Skill Score is $1 - \frac{BS}{\pi(1-\pi)}$. Brier Skill Score is 0.0 for the naive / climatological forecast, and 1.0 for perfect performance, and gives the fraction of the available expected value that is captured by the model.

Consider, for example, a model with $BS = 0.1$, and $\pi = 0.2$. The baseline / climatological model would have a $BS = 0.16$. So the Brier Skill Score suggests that, averaged over the data and over cost ratios between 0 and 1, the model reduces regret by $\frac{0.06}{0.16} = 37.5\%$ when compared to the baseline without a model.

A similar expression exists for the Log Loss where the naive / climatological performance can be shown to be the Bernoulli entropy $H(\pi) = -\pi \log(\pi) - (1-\pi)\log(1-\pi)$. If LL is the Log Loss, then the Skill Score is $1 - \frac{LL}{H(\pi)}$.

J.1 Bounded Brier Score Skill Score

The naive / climatological performance for the Bounded Brier Score is (up to a normalizing factor) $E\left[(y - \text{clip}(\pi))^2 - (y - \text{clip}(y))^2\right]$. Since none of these terms depend on the data, except through the prevalence, this is a valid normalizing factor. We can use it to produce a Bounded Brier Skill Score that shows the fraction of baseline expected regret that is removed by using the classifier.

The same thing is possible for a Bounded Log Loss Skill Score.

In a scenario where the deployment cost of a classifier must be weighed against wholly separate options, it can be useful to rescale our cost matrix such that the loss L matches some quantitative estimate from experts.

In a scenario where we simply want to understand if the classifier is doing something useful, it can be preferable

to rescale using the skill score, so that 1.0 is always the best possible classifier, and 0.0 the result of not using a classifier at all.

K LLM LITERATURE REVIEW

This appendix details our systematic approach to analyzing the use of evaluation metrics across machine learning research. Our primary goal was to determine which metrics researchers prioritize when evaluating binary classifiers across different machine learning domains. The findings provide important context for our main paper's recommendations on metric selection.

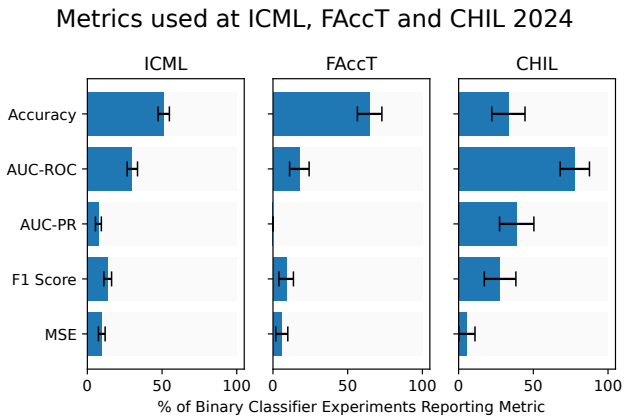

Figure 7: Claude 3.5 Haiku was used to analyze 2,610 papers from three major 2024 conferences. Each plot summarizes the evaluation metrics used for binary classifiers. Accuracy dominates outside healthcare, while AUC-ROC is more prevalent within healthcare domains. Error bars come from binomial confidence intervals.

K.1 Paper Acquisition

Our data collection process focused on gathering papers from three major conferences in 2024: the International Conference on Machine Learning (ICML), the ACM Conference on Fairness, Accountability, and Transparency (FAccT), and the Conference on Health, Inference, and Learning (CHIL). We developed automated scripts to acquire papers from their respective official sources:

- ICML proceedings were accessed through OpenReview's conference platform: `https://openreview.net/group?id=ICML.cc/2024/Conference#tab-accept-oral`
- FAccT papers were obtained from the conference's official website: `https://facctconference.org/2024/acceptedpapers`
- CHIL proceedings were collected from the Proceedings of Machine Learning Research (PMLR): `https://proceedings.mlr.press/v248/`

For text extraction, we employed PyPDF2, a Python-based PDF processing library, to convert all acquired papers from PDF format to plain text.

K.1.1 Classifier Identification

We utilized Anthropic's Claude 3.5 Haiku model ("claude-3-5-haiku-20241022") to search the corpus for papers that mention binary classifiers. The following prompt was sent to Anthropic's API along with the extracted text of each paper.

```
    You are an AI assistant specializing in analyzing research papers in the field of
machine learning and data science. Your task is to examine a given research paper and
analyze its experimental methodology.
```

```
Here is the research paper you need to analyze:

<research_paper>
{{RESEARCH_PAPER}}
</research_paper>

Please follow these steps to analyze the paper:

1. Classifier Detection:
   - Determine if the paper involves a classifier.
   - If yes, identify whether it's binary, multiclass, or multilabel.
   - If no, explain why and continue to the next step.

2. Experiment Detection:
   - Check if the paper includes experimental results.
   - If no, explain why and continue to the next step.

3. Metric Analysis:
   - Identify which of the following metrics are reported:
     a) Classification metrics: Recall, Precision, F1, Accuracy
     b) Probabilistic metrics: Brier Score, Log Loss, Cross Entropy, Perplexity
     c) Error metrics: MSE, RMSE
     d) Cost/benefit metrics: Net Cost, Net Benefit
     e) Curve-based metrics: AUC-ROC, AUC-PR

   Important: Use only these exact metric names in your analysis. For example, use
"AUC-ROC" instead of "AUROC", "AUC", or "AUCROC".

4. Visualization Analysis:
   - Check for the inclusion of these visualizations:
     a) ROC curves
     b) Precision-Recall curves
     c) Brier curves
     d) Decision curves

5. Summary:
   - Provide a JSON object summarizing your findings.

For each step, wrap your thought process in <analysis_breakdown> tags before providing
the final answer. In your analysis breakdown:
    - For classifier detection: List relevant quotes indicating the presence or absence
of a classifier. Classify each quote as supporting binary, multiclass, or multilabel
classification.
    - For experiment detection: List relevant quotes indicating the presence or absence of
experiments. Summarize the type of experiments.
    - For metric and visualization analysis: Create a checklist of all possible metrics and
visualizations mentioned in the instructions. Check them off one by one, citing relevant
quotes for each.

Use the following tags for your responses:
<classifier> : Answer classifier-related questions
<experiments> : Answer experiment-related questions
<metrics> : List reported metrics
<curves> : Indicate included visualizations
```

```
<summary> : Provide the JSON summary

Important guidelines:
- Continue the analysis even if the paper doesn't involve classifiers or experiments.
- Keep explanations concise (maximum 30 words for context in the JSON summary).
- Include relevant quotes from the paper to support your findings.

The JSON summary should follow this structure:
{
    "has_classifier": boolean,
    "classifier_type": "none" | "binary" | "multiclass" | "multilabel",
    "has_experiments": boolean,
    "metrics": {
        "metric_name": {
            "present": boolean,
            "context": "Brief explanation (max 30 words)"
        }
    },
    "visualizations": {
        "visualization_name": boolean
    }
}

metric_name must be one of the following:
"accuracy" | "auc" | "recall" | "precision" | "f1_score" | "mse" | "auprc" |
"cross_entropy" | "rmse" | "mae" | "kl_divergence" | "average_precision"

Please begin your analysis now.
```

We then extracted the JSON summary from the model's response, and found the headline results: that accuracy was dominant at ICML and FAccT, and that AUC-ROC was far more popular at CHIL.

After analyzing 2610 papers across the three conferences, we found significant differences in metric usage patterns:

- At ICML and FAccT, **accuracy** was the dominant evaluation metric for binary classifiers, used in approximately 55.8% and 61.3% of relevant papers respectively.
- At CHIL, **AUC-ROC** was significantly more popular, appearing in 78.8% of papers with binary classifiers, compared to accuracy at 33.6%.
- AUC-PR was reported in 8.7% of ICML papers, only 2.9% of FAccT papers, but 27.7% of CHIL papers, showing domain-specific preferences.
- All other metrics were reported in less than 25% of papers across the board.

These findings suggest substantial domain-specific differences in evaluation practices, particularly between general machine learning and healthcare applications.

K.2 Second Check: More Powerful LLM

We utilized Anthropic's Claude 3.5 Sonnet model ("claude-3-5-sonnet-20241022") to search those papers identified by Haiku as containing binary classifiers. The following prompt was sent to Anthropic's API along with the extracted text of each paper.

```
You are an AI assistant specializing in analyzing research papers in the field of
machine learning and data science. Your task is to examine the research paper given in the
previous message, and analyze its experimental methodology.
```

```
Please follow these steps to analyze the paper:

1. Classifier Detection:
   - Determine if the paper involves a classifier.
   - If yes, identify whether it's binary, multiclass, or multilabel.
   - If no, explain why and continue to the next step.

2. Experiment Detection:
   - Check if the paper includes experimental results.
   - If no, explain why and continue to the next step.

3. Metric Analysis:
   - Identify which of the following metrics are reported:
     a) Classification metrics: Recall, Precision, F1, Accuracy
     b) Probabilistic metrics: Brier Score, Log Loss, Cross Entropy, Perplexity
     c) Error metrics: MSE, RMSE
     d) Cost/benefit metrics: Net Cost, Net Benefit
     e) Curve-based metrics: AUC-ROC, AUC-PR

   Important: Use only these exact metric names in your analysis. For example, use
"AUC-ROC" instead of "AUROC", "AUC", or "AUCROC".

4. Top-K Check:
   - When examining a binary edge classification task evaluated with AUC-ROC, how do we
determine if there are constraints on the number of positive predictions allowed?
   -  For example, is the classifier free to predict any number of positives, or must it
select exactly K edges?
   - What textual indicators or experimental details should we look for in the
methodology to understand these constraints?

5. Summary:
   - Provide a JSON object summarizing your findings.

For each step, wrap your thought process in <analysis_breakdown> tags before providing
the final answer. In your analysis breakdown:
- For classifier detection: List relevant quotes indicating the presence or absence
of a classifier. Classify each quote as supporting binary, multiclass, or multilabel
classification.
- For experiment detection: List relevant quotes indicating the presence or absence of
experiments. Summarize the type of experiments.
- For metric analysis: Create a checklist of all possible metrics mentioned in the
instructions. Check them off one by one, citing relevant quotes for each.

Use the following tags for your responses:
<classifier> : Answer classifier-related questions
<experiments> : Answer experiment-related questions
<metrics> : List reported metrics
<summary> : Provide the JSON summary

Important guidelines:
- Continue the analysis even if the paper doesn't involve classifiers or experiments.
- Keep explanations concise (maximum 30 words for context in the JSON summary).
- Include relevant quotes from the paper to support your findings.

The JSON summary should follow this structure:
```

```
{
    "has_classifier": boolean,
    "has_experiments": boolean,
    "decision_type": "independent" | "top-k" | "other" | "unknown",
    "metrics": {
        "metric_name": {
            "present": boolean,
            "context": "Brief explanation (max 30 words)"
        }
    }
}

Please begin your analysis now.
```

The Sonnet analysis confirmed our initial findings regarding metric preferences and suggested decision scenarios were overwhelmingly independent decisionmaking. However, we left out the decision type from our summary because it required more complex reasoning.

K.3 Human Spot Checks

We spot checked a handful of papers to make sure that the model was accurately reporting the metrics being used. We found that reporting of accuracy, AUC-ROC and AUC-PR was good. Reporting whether Precision and Recall were being used directly as metrics, versus being mentioned in the context of AUC-PR, was sometimes a judgment call. In one case, arguably Mean Squared Error was being used as a loss rather than an evaluation metric, since overall evaluation was not based on model quality.

K.4 Conclusion

This analysis reveals significant differences in how various research communities evaluate binary classifiers. CHIL's preference for AUC-ROC aligns with healthcare's historical connection to ranking metrics, while ICML and FAccT researchers favor accuracy, reflecting their diminished focus on actual costs. These findings inform our main paper's recommendations, showing that consequentialist evaluation remains a niche practice.