Performance Measures for Binary Detectors: A Practitioner's Field Guide

Hari Iyer
Statistical Engineering Division

Abstract

Evaluating a binary detector—a system that assigns scores intended to distinguish targets from nontargets—requires more than reporting a single performance measure. Common practices such as using 0.5 as a threshold without justification, selecting a detector by AUC alone, or abandoning ROC analysis because the classes are imbalanced can lead to misleading conclusions.

This guide provides an accessible, mathematically grounded treatment of binary detector evaluation for practitioners. It organizes evaluation around three distinct questions. **Discrimination** concerns how well the detector ranks targets above nontargets. **Calibration and score meaning** concern what the numerical scores represent and how they should be interpreted. **Decision and operation** concern the choice of an operating threshold under specified costs, prevalence, or constraints such as a maximum false-positive rate. These questions are related, but they are not interchangeable, and no single performance measure answers all of them.

The guide distinguishes population performance quantities from their finite-sample estimates and emphasizes the uncertainty associated with estimating ROC curves, threshold-specific performance, operating thresholds, and minimum expected cost. It also distinguishes class imbalance from sample sparsity and explains the dependence of precision on deployment prevalence. For adversarial evaluation, the guide develops a class-conditional framework for evasion and spoofing attacks, including attack-success measures, paired score changes, operating-point effects, and the role of the attack budget.

The aim is to provide practitioners with a consistent framework for choosing, interpreting, and reporting detector-performance measures, together with the assumptions and uncertainty needed to support the resulting conclusions.

Disclosure: This document was developed in support of the NIST GenAI evaluation series (https://ai-challenges.nist.gov/genai). Super intelligence (SI) tools were used to assist with research on relevant publications, organization of the guide, editing, and grammar checking.

Disclaimers: A contribution from U.S. National Institute of Standards and Technology, not subject to copyright. All views expressed in this paper are those of the authors and of others whom attribution is given and are not necessarily those of NIST nor of any of the institutions cited therein. Certain commercial equipment, instruments, or materials are identified in this report to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Contents

1 The Core Mental Model and Basic Vocabulary **4**
- 1.1 Scores and Decisions Are Different Objects . . . 4
- 1.2 Non-response and Inconclusive Detector Outcomes . . . 5
- 1.3 Three Distinct Evaluation Questions . . . 5
- 1.4 Population Quantities vs. Sample Estimates . . . 6
- 1.5 The Empirical Tally: Confusion Counts and Rates . . . 7

2 Discrimination: Ranking Ability **8**
- 2.1 The Receiver Operating Characteristic (ROC) . . . 8
- 2.2 The Detection Error Tradeoff (DET) . . . 10
- 2.3 Common Summaries and Operating-Point Measures . . . 14
- 2.4 Comparing Detectors: ROC Dominance and Crossing Curves . . . 15
- 2.5 Comparing the Same Detector Across Datasets or Conditions . . . 16

3 Calibration and Score Meaning **16**
- 3.1 When Does Probability Calibration Make Sense? . . . 17
- 3.2 Reliability Diagrams and Binned Expected Calibration Error . . . 17
- 3.3 Brier Score and Class-Conditional Brier Components . . . 19
- 3.4 The Murphy Decomposition: Reliability, Resolution, and Uncertainty . . . 20
- 3.5 Likelihood-Ratio Scores: LR, LLR, C_{llr}, and Empirical Cross-Entropy . . . 21

4 Decision and Operations: Choosing a Threshold **23**
- 4.1 Expected Cost and the Population Operating Target . . . 24
- 4.2 Finite Evaluation Data: A Range of Defensible Operating Thresholds . . . 26
- 4.3 A Fixed False-Positive-Rate Requirement . . . 27
- 4.4 Common Thresholding Shortcuts and Their Assumptions . . . 28
- 4.5 Precision, Recall, and Prevalence . . . 28
- 4.6 F_β: A Composite Summary, Not an Imbalance Cure . . . 29
- 4.7 ROC or Precision–Recall? Ask Which Question You Need Answered . . . 31

5 Uncertainty Quantification: What the Evaluation Data Can Support **31**
- 5.1 Imbalance Is Not the Same as Sparsity . . . 31
- 5.2 The Sampling Unit Matters: Clustered and Repeated Data . . . 32
- 5.3 Sampling Uncertainty Does Not Account for Population Shift . . . 33
- 5.4 Pointwise Intervals and Simultaneous Confidence Bands Answer Different Questions 33
- 5.5 A Distribution-Free ROC Band: DKW–Massart . . . 34
- 5.6 A Practical Alternative: Recompute the Entire ROC in a Stratified Bootstrap . . . 35
- 5.7 Uncertainty for Operating Thresholds and Expected Cost . . . 37
  - 5.7.1 A fixed population false-positive-rate requirement . . . 37
  - 5.7.2 The population minimum expected cost . . . 37
- 5.8 Uncertainty for Other Common Metrics . . . 39

6 Evaluating Adversarial Attack Efficacy **40**
- 6.1 Two Common Class-Conditional Attack Objectives . . . 40
- 6.2 Attack Success at a User-Facing Threshold . . . 41
- 6.3 How Far Did the Attacked Scores Move? . . . 42
- 6.4 Operational Consequence at a Specified System Threshold . . . 44

6.5 ROC and AUC as Secondary Detector-Level Summaries 45
6.6 Attack Efficacy Versus Attack Budget 45

7 Practitioner's Reference Matrix and Red Flags **46**
7.1 Matching the Question to the Quantity 46
7.2 Six Evaluation Red Flags 48

A Formal Mathematical Definitions (Population Quantities) **49**
A.1 Class-Conditional Score Distributions and Threshold Rates 50
A.2 AUC and Equal Error Rate 50
A.3 Precision, Recall, and Prevalence 51
A.4 Probability Scores, Calibration, and Brier Quantities 51
A.5 Likelihood Ratios, Log-Likelihood Ratios, and Logarithmic Loss 52
A.6 Decision-Theoretic Population Quantities 54
A.7 Population Quantities for Adversarial Evaluation 54

1 The Core Mental Model and Basic Vocabulary

Every binary detection task distinguishes between two classes:

- **Class 1 (Target):** The class designated as the event of interest (e.g., a fraudulent transaction, a malignant tumor, a malware payload, or whichever class the application defines as $Y = 1$).
- **Class 0 (Nontarget):** The other class (e.g., a legitimate payment, benign tissue, clean software, or whichever class the application defines as $Y = 0$).

Throughout this guide, *target* simply means the class labeled $Y = 1$; its substantive meaning depends on the application.

We focus on **score-producing binary detectors**. Given an input x (an image, transaction, audio clip, feature vector, etc.), such a detector is intended to produce a numerical score $s(x)$. We adopt the convention that larger scores indicate stronger evidence for the target class.

The numerical scale of $s(x)$ does *not* by itself tell us what the score means. A detector score might be

- an arbitrary ranking score,
- an estimated posterior probability in $[0, 1]$,
- a likelihood ratio in $[0, \infty)$,
- a log-likelihood ratio in $(-\infty, \infty)$, or
- some other application-specific quantity.

Thus, seeing a score such as $s(x) = 0.80$ does **not** automatically justify interpreting it as an 80 % probability. That interpretation is appropriate only when the score is intended to represent a probability and is adequately calibrated.

1.1 Scores and Decisions Are Different Objects

A deployed system may report a score, a decision, or both. In this guide we distinguish the detector's numerical score from the decision rule applied to that score. A common rule declares the target class whenever the score exceeds a threshold θ:

$$\text{Decision} = \begin{cases} \text{Declare Target (Alarm)} & \text{if } s(x) \geq \theta, \\ \text{Declare Nontarget (Pass)} & \text{if } s(x) < \theta. \end{cases}$$

Changing θ does not change the underlying scoring function; it changes the operating rule applied to the scores. A lower threshold generally catches more targets but also produces more false alarms. A higher threshold generally reduces false alarms but misses more targets.

Score $\neq$ decision. The detector supplies evidence on a numerical scale; the operating rule determines what action is taken from that evidence.

1.2 Non-response and Inconclusive Detector Outcomes

A detector may sometimes fail to produce a usable score or decision, or it may explicitly return a non-response or inconclusive outcome. Such an outcome is different from an ordinary score that happens to fall on one side of a threshold. At the *detector level*, it should therefore be retained as a separate response category. The treatment of inconclusive outcomes as a distinct performance issue has also been emphasized in forensic-method evaluation (Swofford et al., 2024).

When labeled evaluation data are available, the detector's non-response behavior should be reported separately for the two true classes. If R denotes the event that the detector produces no usable response, the primary quantities are

$$P(R \mid Y = 1) \qquad \text{and} \qquad P(R \mid Y = 0).$$

These are class-conditional detector-performance quantities and do not depend on deployment prevalence. If the deployment prevalence is $\pi = P(Y = 1)$, the corresponding overall non-response probability is

$$P(R) = \pi P(R \mid Y = 1) + (1 - \pi)P(R \mid Y = 0).$$

At the *decision-system level*, an operational policy may specify how a non-response is resolved. For example, the item may be processed again, passed to a second detector, referred to a human reviewer, or assigned a prespecified fallback decision. A direct mapping of non-response to class 0 or class 1 is one such policy. Once the complete policy produces a final binary decision for every trial, the ordinary binary decision measures in this guide apply to the assessment of system-level performance. The non-response policy should be specified in advance and reported as part of the evaluated system, while the detector-level class-conditional non-response rates should still be reported separately.

1.3 Three Distinct Evaluation Questions

For the score-based analyses below, suppose each trial under analysis has a numerical detector score; detector-level non-responses are tracked separately as described in Section 1.2. We then evaluate the detector on a labeled sample of N examples,

$$\{(s_i, y_i)\}_{i=1}^{N},$$

where $y_i \in \{0, 1\}$ is the ground-truth class label and s_i is the corresponding detector score. The same labeled score sample can be used to ask three conceptually different questions (Figure 1):

1. **Discrimination (ranking):** Does the scoring system tend to place targets above nontargets? This question concerns relative ordering and does not require the score values themselves to have a literal probabilistic meaning.

2. **Calibration / score meaning:** Do the numerical score values have the meaning claimed for them? For example, if a score of 0.80 is intended to mean an 80 % target probability, is that interpretation empirically justified?

3. **Decision / operation:** What action rule should be used in deployment? The answer may depend on explicit false-alarm and miss costs, target prevalence, or an operational constraint such as "keep the false-positive rate below 1 %."

These are different questions, and an answer to one does not automatically answer the others. A detector can rank examples extremely well while producing poorly calibrated probabilities; conversely, choosing a deployment threshold requires information about the intended operating environment that a ranking curve alone cannot supply.

Some metrics primarily address one of these questions, while others combine information from more than one. The purpose of the three-part framework is therefore *not* to force every metric into a mutually exclusive box. Its purpose is to keep the scientific questions separate before choosing a metric.

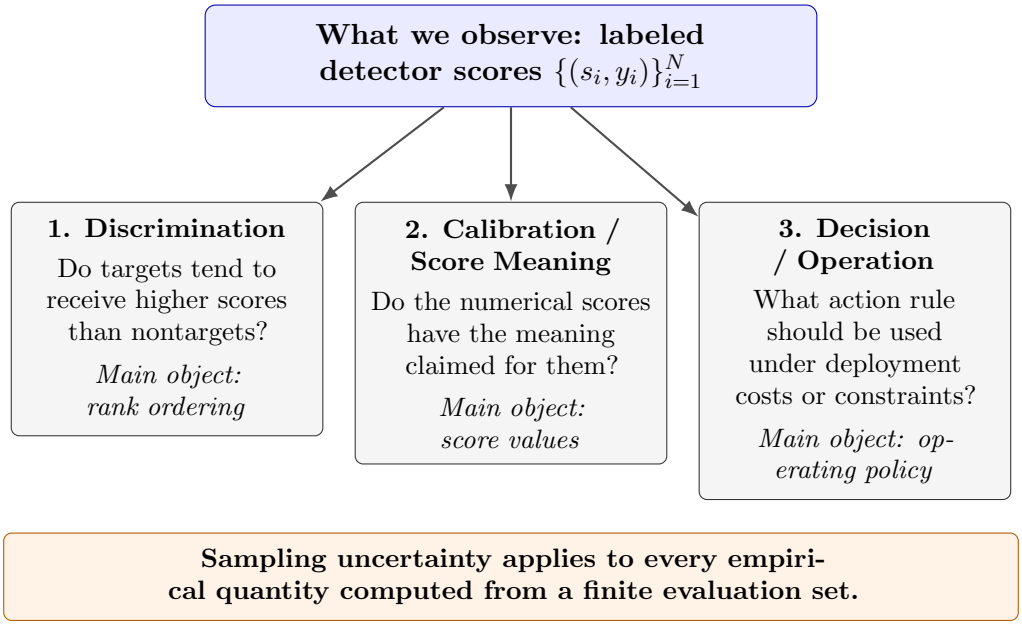


Figure 1: Three distinct questions in binary-detector evaluation. The questions should be kept conceptually separate, even though some metrics combine information from more than one. All empirical answers are subject to sampling uncertainty.

1.4 Population Quantities vs. Sample Estimates

A second distinction is equally important: the difference between a performance quantity defined for an underlying population and the estimate obtained from a finite evaluation sample.

Let

$$N_1 = \#\{i : y_i = 1\}, \qquad N_0 = \#\{i : y_i = 0\}, \qquad N = N_1 + N_0.$$

Then:

- **Population quantities** such as $\mathrm{TPR}(\theta)$, $\mathrm{FPR}(\theta)$, and $\mathrm{PPV}(\theta)$ describe performance under a specified population or operating distribution. They are generally unknown.
- **Sample estimates** such as $\widehat{\mathrm{TPR}}(\theta)$ and $\widehat{\mathrm{FPR}}(\theta)$ are computed from the finite evaluation data and are used to estimate the corresponding population quantities.

Because an evaluation set is finite, every reported metric is subject to sampling uncertainty. If an evaluation contains only 20 fraud cases, for example, its measured detection rate is necessarily noisy even if the same dataset contains ten thousand clean transactions. The absolute numbers of target and nontarget trials therefore matter, not merely their ratio. We return to this issue in Section 5.

1.5 The Empirical Tally: Confusion Counts and Rates

The tally below applies when every trial has a final binary decision, including any trial whose initial non-response has been resolved by a prespecified operational policy as described in Section 1.2. At a fixed threshold θ, every such labeled trial falls into one of four cells:

	Declare Target	**Declare Nontarget**
True Target	$TP(\theta)$ (True Positive)	$FN(\theta)$ (False Negative)
True Nontarget	$FP(\theta)$ (False Positive)	$TN(\theta)$ (True Negative)

Thus,

$$N_1 = TP(\theta) + FN(\theta), \qquad N_0 = FP(\theta) + TN(\theta).$$

More formally,

$$\begin{aligned} TP(\theta) &= \sum_{i:y_i=1} \mathbf{1}\{s_i \geq \theta\} \quad \text{(targets correctly declared target)}, \\ FP(\theta) &= \sum_{i:y_i=0} \mathbf{1}\{s_i \geq \theta\} \quad \text{(nontargets incorrectly declared target)}, \\ TN(\theta) &= \sum_{i:y_i=0} \mathbf{1}\{s_i < \theta\} \quad \text{(nontargets correctly declared nontarget)}, \\ FN(\theta) &= \sum_{i:y_i=1} \mathbf{1}\{s_i < \theta\} \quad \text{(targets incorrectly declared nontarget)}. \end{aligned} \tag{1}$$

Here $\mathbf{1}\{\cdot\}$ denotes the indicator function:

$$\mathbf{1}\{A\} = \begin{cases} 1 & \text{if } A \text{ is true}, \\ 0 & \text{if } A \text{ is false}. \end{cases}$$

The corresponding sample rates are

$$\begin{aligned} \widehat{\mathrm{TPR}}(\theta) &= \frac{TP(\theta)}{N_1} && \text{(True Positive Rate / Sensitivity / Recall)}, \\ \widehat{\mathrm{FPR}}(\theta) &= \frac{FP(\theta)}{N_0} && \text{(False Positive Rate / False Alarm Rate)}, \\ \widehat{\mathrm{FNR}}(\theta) &= \frac{FN(\theta)}{N_1} = 1 - \widehat{\mathrm{TPR}}(\theta) && \text{(False Negative Rate / Miss Rate)}, \\ \widehat{\mathrm{TNR}}(\theta) &= \frac{TN(\theta)}{N_0} = 1 - \widehat{\mathrm{FPR}}(\theta) && \text{(True Negative Rate / Specificity)}. \end{aligned} \tag{2}$$

Because $\widehat{\mathrm{FNR}} = 1 - \widehat{\mathrm{TPR}}$ and $\widehat{\mathrm{TNR}} = 1 - \widehat{\mathrm{FPR}}$, reporting all four rates is usually redundant; one member of each complementary pair is sufficient. ROC displays conventionally use $(\widehat{\mathrm{FPR}}, \widehat{\mathrm{TPR}})$, whereas DET displays use $(\widehat{\mathrm{FPR}}, \widehat{\mathrm{FNR}})$. Some practitioners prefer the latter because both quantities are error rates, but both conventions are widely used.

A small worked example. Suppose the evaluation set contains $N_1 = 100$ targets and $N_0 = 900$ nontargets. At a particular threshold, 80 targets raise an alarm and 45 nontargets raise an alarm. Then

$$TP = 80, \qquad FN = 20, \qquad FP = 45, \qquad TN = 855,$$

and therefore

$$\widehat{\mathrm{TPR}} = \frac{80}{100} = 0.80, \qquad \widehat{\mathrm{FPR}} = \frac{45}{900} = 0.05.$$

The pair $(\widehat{\mathrm{FPR}}, \widehat{\mathrm{TPR}}) = (0.05, 0.80)$ is one operating point of the detector. Changing the threshold changes these four counts and therefore moves the operating point. In Section 2, we will see that sweeping the threshold over all possible values traces out the Receiver Operating Characteristic (ROC) curve.

2 Discrimination: Ranking Ability

The central discrimination question is simple: *does the detector tend to assign larger scores to targets than to nontargets?* This question is about the *ordering* of scores. It does not ask whether a score of 0.9 means a 90 % probability, and it does not ask which threshold should be used in deployment.

Terminology. ROC stands for *receiver operating characteristic*. In this guide, *ROC* refers to the threshold-dependent relationship between false-positive rate and true-positive rate, while *ROC curve* refers specifically to its graphical display. Similarly, *DET* refers to the detection-error tradeoff and *DET curve* to its graphical display. The word *curve* is conventional in both cases, even though an empirical ROC or DET display based on finite data is typically stepwise rather than smooth.

ROC-based discrimination measures depend on the ordering of the scores, not on their numerical scale. If every score is transformed by the same **strictly increasing** function g, then the ordering is unchanged and so are the ROC curve, DET curve, AUC, and EER. Examples include

$$g(s) = 3s + 7, \qquad g(s) = e^s, \qquad g(s) = \frac{1}{1 + e^{-s}}.$$

This invariance is one reason discrimination must be kept conceptually separate from calibration or literal score meaning.

2.1 The Receiver Operating Characteristic (ROC)

The receiver operating characteristic (ROC) curve is the fundamental visual summary of discrimination (Fawcett, 2006; Pepe, 2003). It is important to distinguish the *population ROC curve* from the *empirical ROC curve* computed from a finite evaluation set.

Let $S = s(X)$ denote the detector score for a randomly drawn trial from a specified population. At threshold θ, the population true-positive and false-positive rates are

$$\mathrm{TPR}(\theta) = P(S \geq \theta \mid Y = 1), \qquad \mathrm{FPR}(\theta) = P(S \geq \theta \mid Y = 0). \tag{3}$$

As θ varies over all possible thresholds, the collection of points

$$\big\{\big(\mathrm{FPR}(\theta), \mathrm{TPR}(\theta)\big) : \theta \in \mathbb{R}\big\} \tag{4}$$

forms the **population ROC curve**. It describes the detector's discrimination performance under the specified class-conditional score distributions. The population ROC curve is generally unknown because those distributions are unknown.

From a finite labeled evaluation set, we replace the population probabilities by the sample proportions defined in Section 1. Thus every threshold produces the empirical operating point

$$\big(\widehat{\mathrm{FPR}}(\theta), \widehat{\mathrm{TPR}}(\theta)\big),$$

and sweeping the threshold through the observed scores produces the **empirical ROC curve**. The empirical ROC curve is an estimate of the unknown population ROC curve.

At an extremely high threshold, no examples are declared targets, so

$$\theta \to +\infty \quad \Longrightarrow \quad (\widehat{\mathrm{FPR}}, \widehat{\mathrm{TPR}}) = (0, 0).$$

As the threshold is lowered, more examples trigger alarms, and both rates can increase. At an extremely low threshold, every example is declared a target:

$$\theta \to -\infty \quad \Longrightarrow \quad (\widehat{\mathrm{FPR}}, \widehat{\mathrm{TPR}}) = (1, 1).$$

On a finite evaluation set, the empirical ROC is a **step curve**. Moving the threshold between two adjacent observed score values changes no classifications; the ROC moves only when the threshold crosses one or more observed scores. Figure 2 illustrates this threshold sweep.

Why prevalence does not enter the ROC curve. At any threshold,

$$\widehat{\mathrm{TPR}} = \frac{TP}{N_1}, \qquad \widehat{\mathrm{FPR}} = \frac{FP}{N_0}.$$

Each rate is normalized within its own class. Equivalently, TPR and FPR are conditional probabilities given the true class:

$$\mathrm{TPR}(\theta) = P(S \geq \theta \mid Y = 1), \qquad \mathrm{FPR}(\theta) = P(S \geq \theta \mid Y = 0).$$

Because the conditioning fixes the class, the overall prevalence $P(Y = 1)$ does not enter either rate. Consequently, changing only the target prevalence—while leaving the class-conditional score distributions unchanged—does not change the population ROC curve. Class prevalence does matter for quantities such as precision/PPV and operational cost, which we discuss later. The class-specific sample sizes N_1 and N_0, however, directly affect how precisely the ROC curve can be estimated.

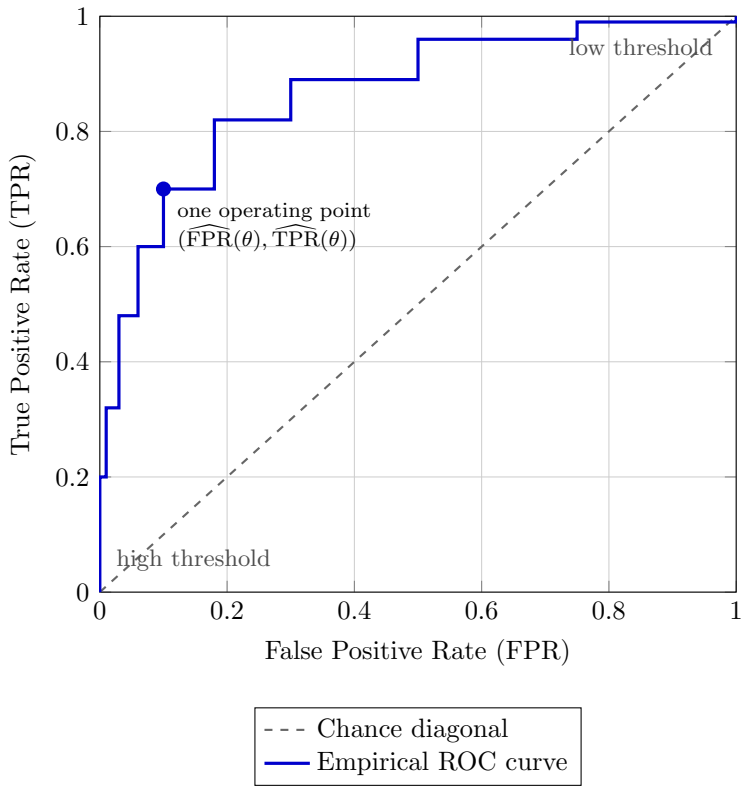


Figure 2: Visual anatomy of an empirical ROC curve. Every point corresponds to a threshold. Lowering the threshold generally moves the operating point toward larger true-positive and false-positive rates. The empirical curve is a staircase because classifications change only when the threshold crosses observed scores.

2.2 The Detection Error Tradeoff (DET)

The ROC curve can be visually compressed in high-performance regimes because good systems crowd near the upper-left corner $(0, 1)$. For comparisons among high-performing detectors, especially when the operational region involves very small error rates, a DET display can therefore be substantially more informative visually than a conventional ROC display. An equivalent error-tradeoff display replaces TPR by $\mathrm{FNR} = 1 - \mathrm{TPR}$, so that both axes represent error probabilities. The underlying threshold-indexed operating points have not changed; only their graphical representation has. ROC remains useful for displaying the full operating tradeoff over the unit square.

A detection error tradeoff (DET) display may apply monotone transformations to the false-positive and false-negative error-rate axes to improve visualization. Logarithmic scales and inverse-Gaussian (probit) scales are both used in practice (Grother and Ngan, 2014; Egan, 1975). These transformations do not change the relationship between FPR and FNR as the threshold varies; they change only how that relationship is displayed.

The conventional probit DET representation of Martin et al. (1997) applies the inverse standard-normal CDF to both error rates. The probit transformation of a probability p is

$$\mathrm{probit}(p) \equiv \Phi^{-1}(p),$$

where Φ is the cumulative distribution function (CDF) of a standard normal random variable,

$N(0,1)$. The probit DET therefore plots

$$\Phi^{-1}\big(\mathrm{FNR}(\theta)\big) \quad \text{versus} \quad \Phi^{-1}\big(\mathrm{FPR}(\theta)\big). \tag{5}$$

Probit scaling spreads small error probabilities over a larger portion of the plotting area and can make differences among high-performing detectors much easier to see. No Gaussian assumption about the detector scores is required to construct or interpret this display. Although the coordinates are transformed by Φ^{-1}, probit DET plots are conventionally labeled with the corresponding error probabilities—for example, 0.1 %, 1 %, 5 %, and 10 %—rather than with their standard-normal quantiles.

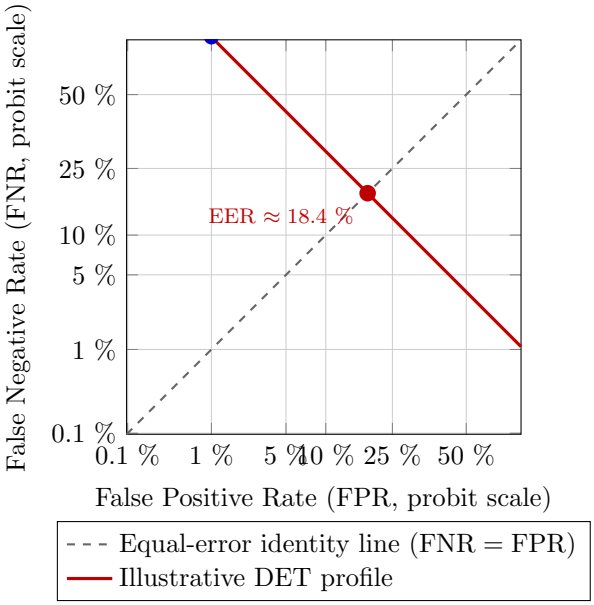


Figure 3: Visual anatomy of a probit DET curve. Probit coordinates spread out small error probabilities, which can make high-performance operating regions easier to compare. The equal-error rate occurs where the detector trace crosses the identity line. The straight illustrative profile shown here is consistent with an equal-variance Gaussian score model, but Gaussianity is not required for using the display.

Gaussian score distributions provide a useful special case. If the target and nontarget score distributions for a detector are Gaussian, its probit DET relation is a straight line. If the two class-specific variances are equal, the line has slope -1; consequently, detectors satisfying the equal-variance Gaussian model produce parallel straight lines on the probit DET display. If the variances are unequal, the relation remains linear, but its slope depends on the ratio of the class-specific standard deviations (Martin et al., 1997; Egan, 1975). This linearity is a consequence of the Gaussian model, not an assumption required for using the DET display.

Two practical advantages are especially useful:

1. **Tail resolution:** operating regions such as FPR ≤ 0.01 or 0.001 occupy much more visual space than they do on an ordinary ROC plot.

2. **Symmetric benchmarking:** the equal-error point lies on the identity line

$$\Phi^{-1}(\mathrm{FNR}) = \Phi^{-1}(\mathrm{FPR}),$$

so EER is easy to identify graphically.

Because $\Phi^{-1}(0) = -\infty$ and $\Phi^{-1}(1) = +\infty$, empirical error rates of exactly 0 or 1 cannot appear at finite DET coordinates. Plotting software therefore clips the displayed range or uses an explicit boundary convention. This is a plotting issue, not new statistical information.

Figure 4 illustrates why this change of coordinates can be useful for two high-performing detectors.

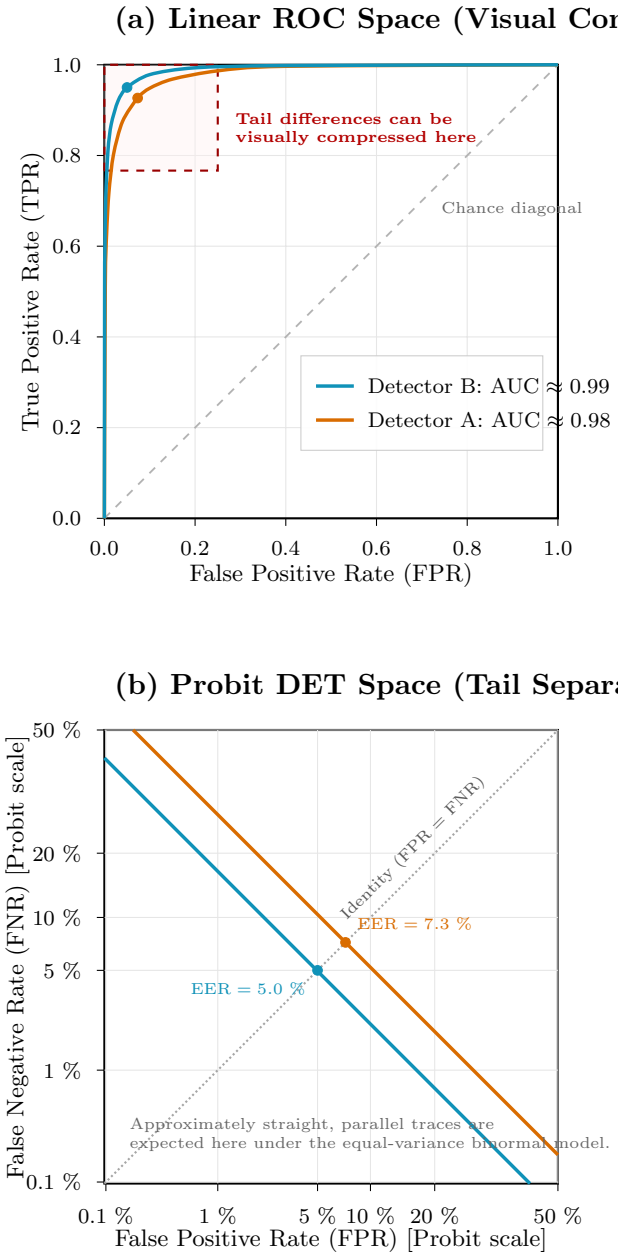


Figure 4: The same pair of high-performing detectors viewed in two coordinate systems. **Top (a):** In ordinary ROC space both traces crowd near the upper-left corner, so differences in the low-error region are visually compressed. **Bottom (b):** Probit DET coordinates expand the tails. In this equal-variance binormal illustration, the DET traces are straight and parallel, and the EER difference is easier to see.

2.3 Common Summaries and Operating-Point Measures

The full ROC curve displays the discrimination trade-off across thresholds. In practice, however, analysts often report one or more scalar summaries or application-specific operating points.

Area Under the Curve (AUC). The area under the ROC curve is a global summary of ranking ability. On a finite sample it is computed by the Mann–Whitney pairwise statistic

$$\widehat{\mathrm{AUC}} = \frac{1}{N_1 N_0} \sum_{i:\, y_i=1} \sum_{j:\, y_j=0} \left[\mathbf{1}\{s_i > s_j\} + \tfrac{1}{2}\mathbf{1}\{s_i = s_j\} \right]. \tag{6}$$

The population interpretation is especially intuitive. Draw one target score S_1 and one nontarget score S_0 independently from their respective populations. Then

$$\mathrm{AUC} = P(S_1 > S_0) + \frac{1}{2} P(S_1 = S_0).$$

Thus AUC is the probability that a randomly drawn target outranks a randomly drawn nontarget, with half credit for a tie (Hanley and McNeil, 1982). The empirical AUC estimates this probability by averaging over all $N_1 N_0$ target–nontarget pairs in the evaluation set.

An uninformative random-ranking system has expected AUC $= 0.5$, whereas AUC $= 1$ represents perfect separation. But an observed AUC $= 0.5$ does *not* by itself prove that the detector is useless at every operating point: a ROC curve can lie above the chance diagonal in one region and below it in another, with the areas canceling. This is one reason the full ROC curve should be inspected rather than relying on AUC alone.

AUC also averages performance over the entire false-positive range $0 \le \mathrm{FPR} \le 1$. If deployment only permits very small false-positive rates, overall AUC may give substantial weight to operating regions that are irrelevant to the application.

Equal Error Rate (EER). The equal error rate is the operating point at which false positives and false negatives are equal:

$$\mathrm{FPR}(\theta_{\mathrm{EER}}) = \mathrm{FNR}(\theta_{\mathrm{EER}}). \tag{7}$$

Equivalently, in ROC coordinates the equal-error point satisfies $\mathrm{TPR} = 1 - \mathrm{FPR}$ (Martin et al., 1997). In a continuous idealization, the crossing point minimizes

$$\max\{\mathrm{FPR}(\theta), \mathrm{FNR}(\theta)\},$$

which gives EER a minimax interpretation. This does *not* make EER a universally optimal deployment threshold; it is appropriate only when that symmetric criterion is relevant to the application.

On a finite empirical ROC, there may be no observed threshold for which $\widehat{\mathrm{FPR}} = \widehat{\mathrm{FNR}}$ exactly. Software therefore typically reports an interpolated EER or uses a stated nearest-point convention. The convention should be documented when exact reproducibility matters.

Performance under a false-positive constraint. Many applications specify a maximum tolerable false-positive rate α rather than symmetric error costs. For example, an operational requirement might demand $\mathrm{FPR} \le 0.001$. On the evaluation sample, a natural operating-point summary is obtained by choosing a threshold that maximizes detection subject to that constraint:

$$\theta^*_\alpha \in \operatorname*{arg\,max}_{\theta:\, \widehat{\mathrm{FPR}}(\theta) \le \alpha} \widehat{\mathrm{TPR}}(\theta), \tag{8}$$

and then reporting the achieved $\widehat{\text{TPR}}$. This is often denoted informally as TPR@FPR $\leq \alpha$. Exact equality $\widehat{\text{FPR}} = \alpha$ need not be attainable on a finite sample.

Resolution is not certainty. With N_0 nontargets and distinct scores, the smallest possible positive increment in empirical false-positive rate is $1/N_0$; similarly, the smallest positive TPR increment is $1/N_1$. For example, if $N_0 = 500$, the empirical FPR moves in increments no smaller than $1/500 = 0.002$ (ties can produce larger jumps). Such a sample cannot finely resolve performance around an FPR of 10^{-4}. Moreover, observing $FP = 0$ does not prove that the population FPR is zero; it only means that no false positives were observed in this finite sample. Section 5 returns to this distinction formally.

2.4 Comparing Detectors: ROC Dominance and Crossing Curves

AUC is useful as a global ranking summary, but detector selection should not automatically be reduced to "pick the larger AUC." The full ROC curves can contain operational information that a single area cannot preserve.

Detector A **ROC-dominates** Detector B if A's ROC curve lies at or above B's curve for every false-positive rate, and lies strictly above it somewhere. In that case, A can achieve at least as much detection power as B at every false-alarm rate represented by the curves.

If the ROC curves cross, neither detector ROC-dominates the other. One may be better in the low-FPR region while the other is better when larger false-positive rates are acceptable. Which detector is preferable then depends on the deployment region or on an explicit decision criterion.

A larger overall AUC does not imply that a detector is better at every operating point. When ROC curves cross, inspect the region that matters for the application rather than using AUC alone to declare a universal winner.

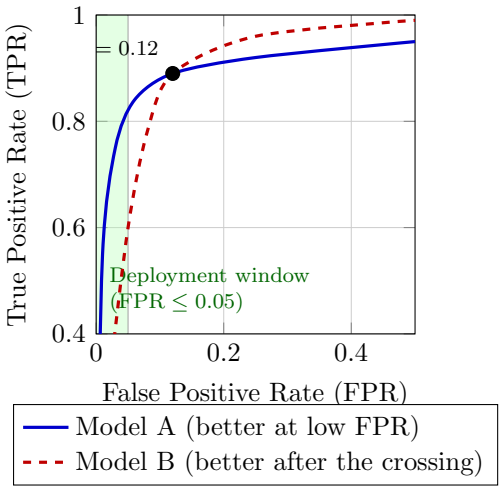


Figure 5: The crossing-curves trap. In this constructed example, Model A is substantially better in the deployment window FPR ≤ 0.05, while Model B is better at larger false-positive rates and may have the larger overall AUC. Neither model ROC-dominates the other.

2.5 Comparing the Same Detector Across Datasets or Conditions

ROC and DET summarize how the available error tradeoff changes as the threshold varies, and they are useful when comparing a detector across datasets, demographic groups, sensors, image-quality conditions, or other evaluation strata. A different operational question arises when the *same numerical threshold* will be used in every condition: at that fixed threshold, how do the two error rates change?

For that question, ROC or DET alone can obscure whether an observed operational difference is driven primarily by false positives, false negatives, or both. A useful complementary display plots the two class-conditional error rates separately as functions of threshold,

$$\mathrm{FPR}_g(\theta) \quad \text{and} \quad \mathrm{FNR}_g(\theta),$$

with one trace for each dataset, condition, or group g. The common vertical slice at a deployed threshold then shows directly how the two error types differ across conditions; Figure 6 gives a schematic example.

Such a fixed-threshold comparison presumes that the detector score scale and the numerical threshold have the same operational meaning across the conditions being compared. If scores are separately recalibrated, normalized, or otherwise transformed by condition, equality of the numerical thresholds need not represent equality of the operating rule. As elsewhere in this guide, observed differences should be accompanied by uncertainty appropriate to the sampling design.

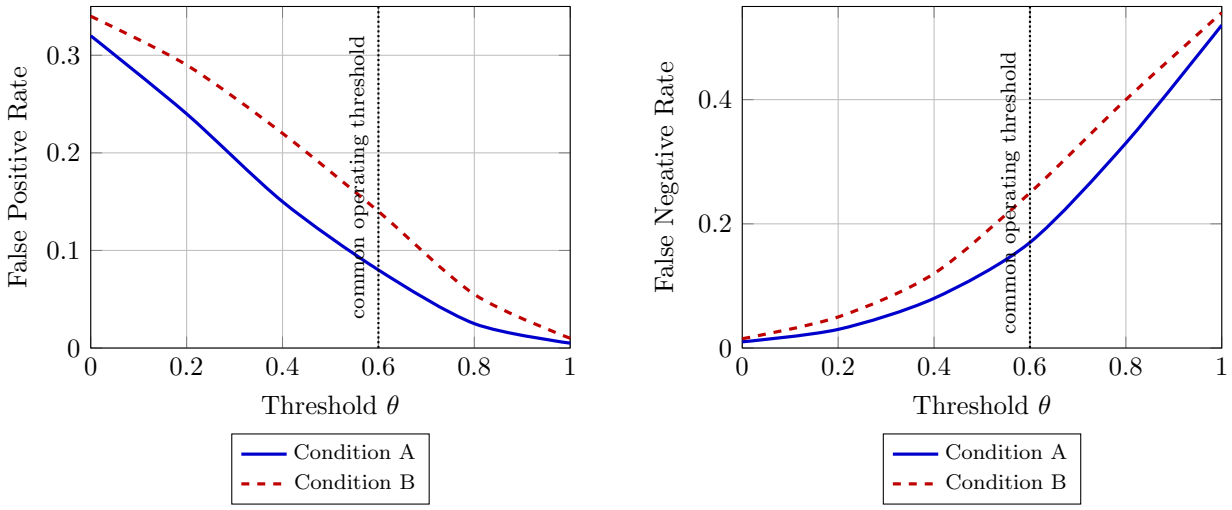


Figure 6: Schematic fixed-threshold comparison of the same detector under two conditions. Plotting $\mathrm{FPR}(\theta)$ and $\mathrm{FNR}(\theta)$ separately shows whether operational differences at a common threshold arise on the nontarget side, the target side, or both. The curves are illustrative rather than measured data.

3 Calibration and Score Meaning

Section 2 asked whether targets tend to receive larger scores than nontargets. Calibration asks a different question: *what do the numerical score values themselves mean?*

That question only makes sense relative to a claimed interpretation of the score. A detector may output an arbitrary ranking score, an estimated posterior probability, a likelihood ratio (LR), or a log-likelihood ratio (LLR). These objects have different meanings and therefore require different calibration checks.

Suppose, for example, that an airport screening detector outputs the value $s = 0.91$. If s is merely an internal ranking score, there is no reason to interpret 0.91 as a 91 % probability that the baggage contains contraband. If, however, the system explicitly reports s as a posterior probability, then that probability claim can be evaluated empirically. Modern machine-learning systems can have excellent discrimination while producing poorly calibrated probability estimates (Guo et al., 2017).

Calibration is claim-dependent. An arbitrary detector score need not be a probability. Probability calibration asks whether a score that is *intended* to mean, say, "80 % probability of target" actually has that meaning. Likelihood-ratio calibration asks a different question: whether the reported LR or LLR has the evidentiary meaning claimed for it.

3.1 When Does Probability Calibration Make Sense?

Let S denote a detector output that is intended to represent the posterior probability of the target class. Define the population calibration function

$$\eta(s) \triangleq P(Y = 1 \mid S = s). \tag{9}$$

A perfectly calibrated probability forecast satisfies

$$\eta(s) = s, \tag{10}$$

for score values that occur in the population. Equivalently, and more generally,

$$\mathbb{E}[Y \mid S] = S. \tag{11}$$

Thus, among cases assigned probability near 0.80, roughly 80 % should actually be targets in the population for which the probability interpretation is intended.

This definition depends on the *joint population distribution* of (S, Y). In particular, posterior probabilities depend on target prevalence. If an evaluation set has been deliberately rebalanced—for example, to contain 50 % targets even though the deployment population contains only 1 %—then a naive calibration analysis on that case–control sample does not by itself establish probability calibration for deployment. The evaluation must represent, or be appropriately adjusted to, the population in which the probability statements will be interpreted.

3.2 Reliability Diagrams and Binned Expected Calibration Error

A reliability diagram is a simple empirical diagnostic for probability calibration. Recall that, for observation i, s_i denotes the reported score and $y_i \in \{0, 1\}$ the corresponding ground-truth label. Partition the probability range $[0, 1]$ into M bins. For bin m, containing n_m observations, compute

$$\bar{s}_m = \frac{1}{n_m} \sum_{i \in m} s_i, \qquad \text{(mean predicted probability)}, \tag{12}$$

$$\widehat{r}_m = \frac{1}{n_m} \sum_{i \in m} y_i, \qquad \text{(observed target fraction)}. \tag{13}$$

Plotting $\widehat{r}_m$ against $\bar{s}_m$ yields the empirical reliability diagram (Figure 7). The population calibration relationship for a perfectly calibrated probability forecast lies on the 45° line. A finite empirical

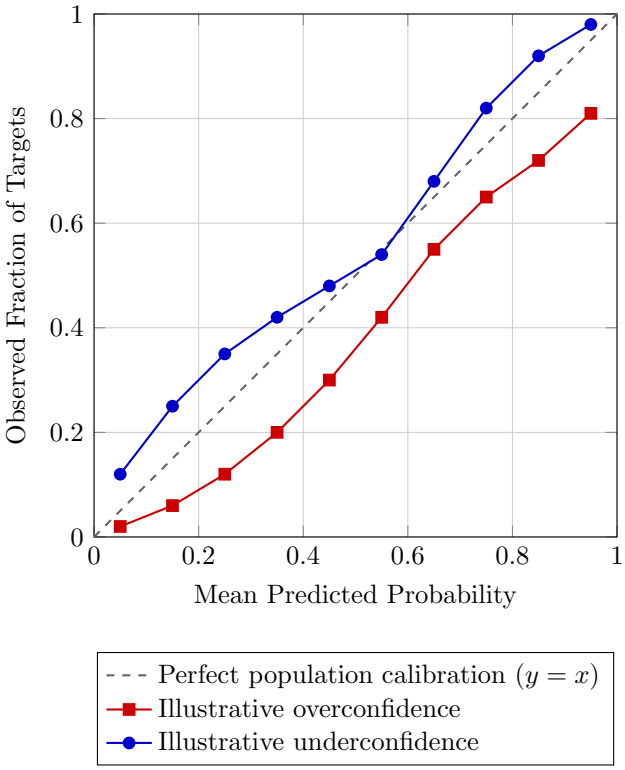


Figure 7: Reliability diagram. A point below the diagonal indicates overconfidence: for example, a bin whose mean predicted probability is 0.80 but whose observed target fraction is 0.60. The plotted empirical points are estimates, not population truths; they are affected by sampling variability and by the choice of bins.

reliability diagram will fluctuate around that line because of sampling variability, even when the underlying probabilities are perfectly calibrated.

A common scalar summary of the same binned discrepancies is the **binned Expected Calibration Error**, often abbreviated ECE in the machine-learning literature (Naeini et al., 2015; Guo et al., 2017):

$$\mathrm{ECE}_{L_1} = \sum_{m=1}^{M} \frac{n_m}{N} \left| \widehat{r}_m - \bar{s}_m \right|. \tag{14}$$

The subscript L_1 emphasizes that the binwise discrepancy is measured by absolute difference. Other variants replace the absolute gap by a squared gap, or report a root-mean-square calibration error. Because these conventions are not identical, the precise definition should be stated whenever an "ECE" value is reported.

Binned ECE_{L_1} is easy to understand, but it should not be treated as a definitive measure of calibration. Its value can change substantially with the number of bins and with equal-width versus equal-frequency binning, and it is also subject to ordinary sampling variation (Kumar et al., 2019). It summarizes a particular empirical reliability diagram; it does not remove the need to inspect that diagram or quantify uncertainty.

Do not confuse two different "ECE" acronyms. In much of machine learning, ECE means *Expected Calibration Error*, usually a binned absolute calibration discrepancy. In forensic evidence evaluation, ECE often means *Empirical Cross-Entropy*, an information-theoretic loss for likelihood-ratio systems. They are different mathematical objects. We introduce the forensic quantity in §3.5.

3.3 Brier Score and Class-Conditional Brier Components

When $s_i \in [0, 1]$ is intended as a probability for $Y_i = 1$, the Brier score is the mean squared probability error:

$$\mathrm{Brier} = \frac{1}{N} \sum_{i=1}^{N} (s_i - y_i)^2. \tag{15}$$

The Brier score is a *strictly proper scoring rule*: in expectation, it is uniquely minimized when the reported probability equals the true conditional probability of the target event given the information available to the forecaster (Gneiting and Raftery, 2007). A scoring rule is *proper* if truthful probability reporting minimizes expected loss, and *strictly proper* if truthful reporting is the unique minimizer. Lower Brier values are better, and 0 is perfect.

For diagnostics it is sometimes useful to separate the contributions of targets and nontargets:

$$\mathrm{BrierT} = \frac{1}{N_1} \sum_{i:\, y_i = 1} (s_i - 1)^2, \qquad \mathrm{BrierN} = \frac{1}{N_0} \sum_{i:\, y_i = 0} (s_i - 0)^2. \tag{16}$$

They combine to give the ordinary sample Brier score:

$$\mathrm{Brier} = \frac{N_1}{N} \mathrm{BrierT} + \frac{N_0}{N} \mathrm{BrierN}. \tag{17}$$

Ideal performance is 0 on both class-conditional components. However, BrierT and BrierN are diagnostic components, not proper scoring rules when considered separately.

Consider an evaluation sample containing one target and one nontarget. Model A assigns scores 0.99 and 0.01, respectively, while Model B assigns 0.51 and 0.49. Both models rank the two observations correctly, so both have the same *empirical* AUC,

$$\widehat{\mathrm{AUC}} = 1,$$

on this sample.

Their empirical Brier scores, however, differ substantially because the Brier score uses the numerical values of probability forecasts, not merely their ordering. For Model A,

$$\widehat{\mathrm{Brier}} = \frac{(0.99-1)^2 + (0.01-0)^2}{2} = 0.0001,$$

whereas for Model B,

$$\widehat{\mathrm{Brier}} = \frac{(0.51-1)^2 + (0.49-0)^2}{2} \approx 0.24.$$

This toy example does *not* establish the population AUC, population Brier score, or calibration of either model. It illustrates only that, on a given finite sample, empirical ROC/AUC depends on the ordering of the scores, whereas the empirical Brier score also depends on the numerical values of probability forecasts.

3.4 The Murphy Decomposition: Reliability, Resolution, and Uncertainty

Although miscalibration contributes to the Brier score, the Brier score is not a measure of calibration alone. The following decomposition makes this transparent. At the population level, let

$$\eta(S) \triangleq P(Y=1 \mid S), \qquad \pi \triangleq P(Y=1).$$

The Murphy decomposition (Murphy, 1973; DeGroot and Fienberg, 1983) gives the exact identity

$$\underbrace{\mathbb{E}\big[(S-Y)^2\big]}_{\text{Brier score}} = \underbrace{\mathbb{E}\big[(S-\eta(S))^2\big]}_{\text{Reliability}} - \underbrace{\mathbb{E}\big[(\eta(S)-\pi)^2\big]}_{\text{Resolution}} + \underbrace{\pi(1-\pi)}_{\text{Uncertainty}}. \tag{18}$$

The three terms have distinct meanings:

- **Reliability (miscalibration):** How far the reported probabilities deviate from the true conditional target probabilities. In this decomposition, the term called *reliability* is a miscalibration penalty, so lower is better and 0 is perfect calibration.
- **Resolution:** How much the conditional target probability varies across forecast values. Higher resolution means that the forecasts separate groups having meaningfully different target frequencies. Resolution is related conceptually to discrimination, but it is not equivalent to AUC and is not determined by rank ordering alone.
- **Uncertainty:** The intrinsic Bernoulli variance of the outcome, $\pi(1-\pi)$, determined by the population prevalence.

Because Resolution is nonnegative and $\pi(1-\pi) \le 1/4$, Equation (18) implies the population bound

$$\text{Reliability} \ge \text{Brier score} - \frac{1}{4}. \tag{19}$$

If a probability-producing detector has a population Brier score of 0.85, then its reliability term must be at least $0.85 - 0.25 = 0.60$. Thus it cannot be well calibrated, regardless of prevalence. If that same detector also has very high AUC, the two statements are not contradictory: the detector can rank targets extremely well while reporting badly distorted probabilities.

Empirical estimates of the Murphy components are themselves sample-dependent and require estimation of the calibration relationship $\eta(S)$. As elsewhere in this guide, apparent differences should therefore be interpreted together with sampling uncertainty rather than as exact population facts.

3.5 Likelihood-Ratio Scores: LR, LLR, C_{llr}, and Empirical Cross-Entropy

Probability calibration is not the only meaningful form of score calibration. In forensic science and biometric verification, a system may instead report the strength of evidence as a **likelihood ratio**. For a continuous raw score s, define

$$\mathrm{LR}(s) \triangleq \frac{f(s \mid Y = 1)}{f(s \mid Y = 0)}, \tag{20}$$

where $f(s \mid Y = 1)$ and $f(s \mid Y = 0)$ are the class-conditional score densities. An LR greater than 1 supports the target hypothesis, an LR less than 1 supports the nontarget hypothesis, and $\mathrm{LR} = 1$ is neutral evidence.

The corresponding natural log-likelihood ratio is

$$\ell(s) \triangleq \ln \mathrm{LR}(s) = \ln \left[\frac{f(s \mid Y = 1)}{f(s \mid Y = 0)} \right]. \tag{21}$$

Thus $\ell = 0$ is neutral evidence, positive LLRs support the target hypothesis, and negative LLRs support the nontarget hypothesis (Brümmer and Du Preez, 2006; Hannig and Iyer, 2022).

The key distinction from a posterior probability is that an LR is an *evidence multiplier*. If the prior target odds are

$$O(\pi) \triangleq \frac{\pi}{1 - \pi}, \tag{22}$$

then, for an observed score s, Bayes' rule gives

$$\text{posterior odds} = \mathrm{LR}(s) \times O(\pi). \tag{23}$$

The likelihood ratio $\mathrm{LR}(s)$ represents the evidentiary contribution of the observed score, while the prior odds incorporate the target prevalence. The LR itself does not contain the deployment prevalence; prevalence enters when the evidence is combined with prior odds.

The cost of log-likelihood ratios. When a system reports likelihood ratios or log-likelihood ratios, we want to assess not only whether target scores tend to exceed nontarget scores, but also whether the reported numerical strength of evidence is appropriate. The metric C_{llr} uses logarithmic loss to evaluate both aspects together: discrimination and calibration of the reported LLR scale. For reported natural LLRs ℓ_i (Brümmer and Du Preez, 2006):

$$C_{\text{llr}} = \frac{1}{2} \left[\frac{1}{N_1} \sum_{i:\, y_i = 1} \log_2\left(1 + e^{-\ell_i}\right) + \frac{1}{N_0} \sum_{i:\, y_i = 0} \log_2\left(1 + e^{\ell_i}\right) \right]. \tag{24}$$

An uninformative system that always reports $\ell = 0$ receives $C_{\text{llr}} = 1$. Lower values are better. Logarithmic loss penalizes confidently wrong evidence especially strongly; unlike the Brier score, it is unbounded above.

Like the ordinary Brier score, C_{llr} is not a measure of discrimination alone. Its observed value reflects both how well the classes can be separated and how well the numerical LLR scale represents evidentiary strength. In empirical LR evaluation, pool-adjacent-violators (PAV) isotonic calibration is often used to determine the minimum log-likelihood-ratio cost attainable without changing the score ordering. The resulting $C_{\text{llr}}^{\text{min}}$ summarizes discrimination potential under optimal monotone calibration, while

$$C_{\text{llr}}^{\text{cal}} \triangleq C_{\text{llr}} - C_{\text{llr}}^{\text{min}} \tag{25}$$

quantifies the empirical penalty associated with the reported calibration.

Forensic Empirical Cross-Entropy (ECE). A related diagnostic used in forensic evidence evaluation examines logarithmic loss over a range of prior odds (Ramos et al., 2013; Brümmer and Du Preez, 2006). This quantity is also abbreviated **ECE**, but here the letters mean *Empirical Cross-Entropy*, not Expected Calibration Error.

For target prevalence π, define prior odds $O(\pi) = \pi/(1-\pi)$. Bayes' rule gives

$$P(Y = 1 \mid \text{LR}, \pi) = \frac{\text{LR}\, O(\pi)}{1 + \text{LR}\, O(\pi)}, \qquad P(Y = 0 \mid \text{LR}, \pi) = \frac{1}{1 + \text{LR}\, O(\pi)}. \tag{26}$$

Here LR_i denotes the likelihood ratio reported by the system for evaluated trial i; if the system reports a likelihood-ratio function $\text{LR}(s)$ applied to an underlying score s_i, then $\text{LR}_i = \text{LR}(s_i)$. The empirical cross-entropy at prior π is

$$\text{ECE}(\pi) = -\frac{\pi}{N_1} \sum_{i:\, y_i = 1} \log_2 P(Y = 1 \mid \text{LR}_i, \pi) - \frac{1-\pi}{N_0} \sum_{j:\, y_j = 0} \log_2 P(Y = 0 \mid \text{LR}_j, \pi). \tag{27}$$

Equivalently,

$$\text{ECE}(\pi) = \frac{\pi}{N_1} \sum_{i:\, y_i = 1} \log_2 \left(1 + \frac{1}{\text{LR}_i\, O(\pi)}\right) + \frac{1-\pi}{N_0} \sum_{j:\, y_j = 0} \log_2 (1 + \text{LR}_j\, O(\pi))\,. \tag{28}$$

The target term in Equation (28) contains the product $\text{LR}_i O(\pi)$ in the denominator; this follows directly from the posterior-odds form of Bayes' rule.

Because logarithmic loss is strictly proper, empirical cross-entropy jointly reflects discrimination and calibration of the reported likelihood ratios. An ECE plot, illustrated in Figure 8, evaluates this loss over a chosen range of operational prior log-odds, often written

$$\mathcal{O} = \log_{10} O(\pi). \tag{29}$$

Three curves are commonly useful:

- **Null reference (ECE_{null}):** the loss for neutral evidence, $\text{LR} \equiv 1$. It equals the binary entropy of the prior,
$$\text{ECE}_{\text{null}}(\pi) = -\pi \log_2 \pi - (1-\pi) \log_2 (1-\pi). \tag{30}$$
- **Observed curve (ECE_{obs}):** the loss computed from the likelihood ratios actually reported by the system using Equation (28).

- **Monotone-calibrated lower envelope ($\mathrm{ECE}_{\mathrm{min}}$):** the empirical loss after optimal PAV monotone calibration. The gap $\mathrm{ECE}_{\mathrm{obs}} - \mathrm{ECE}_{\mathrm{min}}$ measures the empirical calibration penalty relative to that monotone optimum; the improvement from the null reference to $\mathrm{ECE}_{\mathrm{min}}$ reflects the useful discrimination available in the score ordering.

The vertical reduction from $\mathrm{ECE}_{\mathrm{null}}$ to $\mathrm{ECE}_{\mathrm{min}}$ represents the performance gain attributable to the score ordering after optimal monotone calibration; this is what Figure 8 labels the *discrimination benefit*. The reader is referred to Zadora et al. (2014) for further discussion.

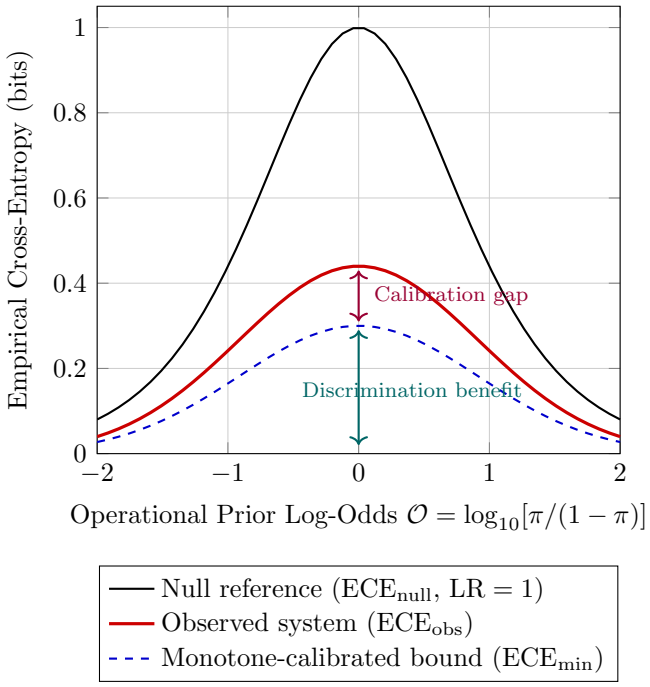


Figure 8: Illustrative empirical cross-entropy diagnostic profiles. The null reference is the loss from neutral evidence. The gap between the observed curve and the monotone-calibrated lower envelope represents calibration loss relative to the PAV optimum, while the reduction from the null reference to the lower envelope reflects useful discrimination in the score ordering.

4 Decision and Operations: Choosing a Threshold

Discrimination tells us what combinations of true-positive rate and false-positive rate are available as the threshold changes. It does *not*, by itself, tell us which operating point should be used in deployment. Choosing a threshold requires an operational objective.

In this guide we consider two common operating formulations. The first specifies the deployment prevalence together with the relative consequences of misses and false alarms and chooses a threshold to minimize expected cost. The second specifies a maximum acceptable false-positive rate and chooses the threshold that gives the largest possible true-positive rate while respecting that requirement. More complicated operational constraints, such as multiple simultaneous constraints or resource-allocation limits, are beyond the scope of this introductory guide.

4.1 Expected Cost and the Population Operating Target

Let C_{FA} denote the cost of a false alarm, C_{Miss} the cost of a miss, and let

$$\pi = P(Y = 1)$$

denote the *deployment* prevalence of targets. The corresponding nontarget prevalence is $1 - \pi$. At threshold θ, the population expected detection cost is

$$C_{det}(\theta) = C_{Miss}\,\pi\,\mathrm{FNR}(\theta) + C_{FA}\,(1-\pi)\,\mathrm{FPR}(\theta). \tag{31}$$

The quantity $C_{det}(\theta)$ is commonly called the *Detection Cost Function* (DCF), particularly in the NIST speaker-recognition evaluation literature (NIST, 2018, 2024; Wu et al., 2011).

On a finite evaluation set, the corresponding empirical estimate is

$$\widehat{C}_{det}(\theta) = C_{Miss}\,\pi\,\widehat{\mathrm{FNR}}(\theta) + C_{FA}\,(1-\pi)\,\widehat{\mathrm{FPR}}(\theta). \tag{32}$$

Importantly, the π appearing here is the prevalence relevant to deployment. It need not equal the target fraction N_1/N in an artificially balanced or case-control evaluation set.

If the population error rates were known exactly, the population-optimal expected cost would be

$$C^*_{det} = \inf_{\theta} C_{det}(\theta). \tag{33}$$

A threshold attaining this infimum, when one exists, is a population cost-minimizing threshold. In practice, however, the population error rates are unknown and must be estimated from finite data. Consequently, neither C^*_{det} nor its minimizing threshold is known exactly.

Iso-cost lines in ROC space. Using $\mathrm{FNR}(\theta) = 1 - \mathrm{TPR}(\theta)$, a line of constant population cost in ROC coordinates has slope

$$m = \frac{C_{FA}(1-\pi)}{C_{Miss}\pi}. \tag{34}$$

Thus the operational trade-off depends jointly on prevalence and relative error cost.

For example, if false alarms and misses have equal cost, a slope of $m = 1.25$ corresponds to

$$\pi = \frac{1}{1+1.25} \approx 0.444.$$

The same slope can arise in a rare-target setting. If $\pi = 0.01$, then $m = 1.25$ corresponds to

$$\frac{C_{Miss}}{C_{FA}} = \frac{1-\pi}{m\pi} = \frac{0.99}{1.25(0.01)} \approx 79.2.$$

In other words, the same geometric slope can represent very different combinations of prevalence and cost.

For fixed prevalence and error costs, the iso-cost contours form a family of parallel lines. For a smooth population ROC, a cost-minimizing operating point occurs where one of these iso-cost lines supports the ROC curve. Figure 9 illustrates this geometry. With finite evaluation data, the empirical ROC and empirical cost curve are only estimates of their population counterparts; the apparent sample minimum should therefore not be treated as an exactly known deployment threshold.

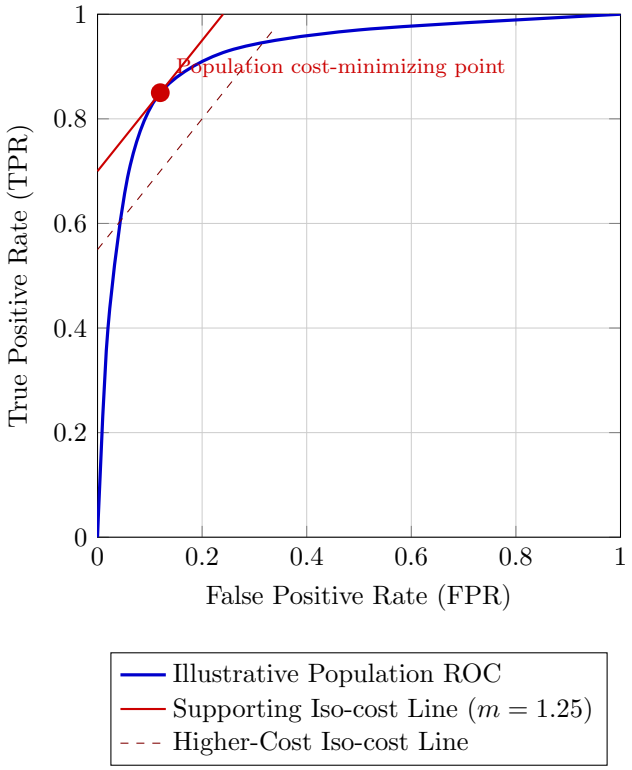


Figure 9: Expected-cost threshold selection in ROC space. Parallel lines represent equal-cost contours for a fixed prevalence and cost ratio. For a smooth population ROC, a population cost-minimizing operating point can be visualized where a supporting iso-cost line reaches the ROC curve (Fawcett, 2006). A finite empirical ROC estimates this population geometry and does not identify the population-optimal threshold without uncertainty.

If the score is a calibrated posterior probability. Suppose the detector outputs

$$p(x) = P(Y = 1 \mid X = x)$$

for the deployment population. Here $p(x)$ denotes the posterior probability based on the detector input x. This differs from the calibration relation $P(Y = 1 \mid S = s)$, which asks whether a reported score s has the probabilistic meaning claimed for it. A *Bayes decision rule* chooses the action having the smaller expected loss conditional on the available information. Declaring "target" incurs expected conditional loss $C_{FA}[1 - p(x)]$, whereas declaring "nontarget" incurs expected conditional loss $C_{Miss}p(x)$. Therefore the Bayes decision rule declares target when

$$p(x) \geq \frac{C_{FA}}{C_{FA} + C_{Miss}}. \tag{35}$$

If the two error costs are equal, the threshold is 0.5. Notice that prevalence does not appear explicitly in Equation (35): when $p(x)$ is a genuine deployment posterior probability, the prior prevalence has already been incorporated into that probability.

If the score is a likelihood ratio. For a likelihood ratio $LR(x)$, Bayes' rule gives

$$\text{posterior odds} = LR(x) \times \frac{\pi}{1 - \pi}.$$

The Bayes decision rule is therefore to declare the target class whenever

$$LR(x) \geq \frac{C_{FA}(1-\pi)}{C_{Miss}\pi}. \tag{36}$$

Thus prevalence appears explicitly when the score represents prior-independent evidentiary strength rather than a posterior probability.

4.2 Finite Evaluation Data: A Range of Defensible Operating Thresholds

The population quantity in Equation (33) is what we ultimately care about, but a finite evaluation set gives only estimates of $\mathrm{FNR}(\theta)$, $\mathrm{FPR}(\theta)$, and therefore $C_{det}(\theta)$. Sampling variation can move the apparent minimum and can make several nearby thresholds look essentially indistinguishable.

For that reason, a useful evaluation should not present a single sample-minimizing threshold as though it were known exactly. Instead, it should aim to report two things:

1. an uncertainty interval for the population minimum expected cost C^*_{det}; and
2. a *range of defensible operating thresholds*: thresholds whose population expected costs cannot be distinguished, given the evaluation uncertainty, from the minimum or from a practically negligible neighborhood of the minimum.

If a small excess cost $\varepsilon \geq 0$ is considered operationally negligible, the corresponding population near-optimal set is

$$\Theta_\varepsilon = \{\theta : C_{det}(\theta) \leq C^*_{det} + \varepsilon\}. \tag{37}$$

When $\varepsilon = 0$, this reduces to the population-optimal threshold set; with $\varepsilon > 0$, it describes thresholds that are practically near-optimal. In a finite evaluation, this set must itself be inferred with uncertainty. We return to the construction of uncertainty bands and their use for this purpose in Section 5.

Do not confuse the lowest cost observed on one finite evaluation set with the population optimum. The practical goal is to estimate the population minimum expected cost and to identify a defensible range of operating thresholds, with uncertainty made explicit.

Terminology you may encounter: minDCF and actDCF. In speaker-recognition evaluation, two sample summaries are commonly reported (NIST, 2018, 2024). The term *minDCF* usually denotes

$$\widehat{\mathrm{minDCF}} = \min_\theta \widehat{C}_{det}(\theta),$$

the smallest empirical DCF obtained after searching over thresholds on a particular labeled evaluation set. It is a commonly reported descriptive sample quantity and a natural plug-in estimate of C^*_{det}, but it is *not* the known population minimum. Moreover, because the same finite evaluation data are used both to search over thresholds and to evaluate the resulting minimum, the empirical minimum can be optimistic as an estimate of the population minimum. It should therefore not be described as "best attainable performance."

The term *actDCF* usually denotes the empirical DCF at a threshold fixed independently of the evaluation labels. This can be useful when the purpose is to audit a pre-specified operating rule. It is still a finite-sample estimate and therefore has sampling uncertainty. For threshold selection,

however, the more informative goal is the population cost curve together with an uncertainty interval for C^*_{det} and a defensible threshold range.

Within the *same* evaluation corpus, minDCF is necessarily no larger than the empirical DCF at any other threshold because it is defined as the empirical minimum. That ordering does not carry across different corpora: an actDCF measured on one corpus can be smaller or larger than a minDCF measured on another.

4.3 A Fixed False-Positive-Rate Requirement

Sometimes the operational requirement is simpler than specifying costs. For example, an application may require that the population false-positive rate not exceed a specified value τ:

$$\mathrm{FPR}(\theta) \le \tau.$$

The natural operating objective is then to obtain the largest possible true-positive rate while satisfying that requirement:

$$\theta_\tau \in \arg\max_\theta \mathrm{TPR}(\theta) \qquad \text{subject to} \qquad \mathrm{FPR}(\theta) \le \tau. \tag{38}$$

This is the Neyman–Pearson style of operating-point selection.

Because larger scores indicate stronger target evidence, lowering the threshold generally increases both TPR and FPR. Thus the desired operating threshold is the least restrictive threshold that still satisfies the false-positive-rate requirement. In the idealized continuous-score case, if F_0 is the population CDF of nontarget scores, this threshold is the corresponding upper-tail quantile,

$$\theta_\tau = F_0^{-1}(1-\tau). \tag{39}$$

With discrete scores or ties, an exact false-positive rate of τ may not be attainable; the relevant threshold is then one that satisfies $\mathrm{FPR}(\theta) \le \tau$ while giving the largest attainable TPR.

As before, the population quantities are unknown. A finite evaluation set provides an estimate $\widehat{\theta}_\tau$ of the required threshold and an estimate of the resulting detection rate,

$$\widehat{\mathrm{TPR}}\left(\widehat{\theta}_\tau\right).$$

Both should be accompanied by uncertainty. In particular, uncertainty in the achieved TPR reflects not only the finite number of target trials but also uncertainty in the threshold estimated from the nontarget trials.

There is also an important distinction between observing

$$\widehat{\mathrm{FPR}}(\theta) \le \tau$$

on the evaluation sample and having statistical evidence that the *population* false-positive rate satisfies the operational requirement. If the requirement truly applies to the population, a more defensible rule is to require the upper confidence bound for the population false-positive rate to satisfy

$$U_{\mathrm{FPR}}(\theta) \le \tau.$$

The corresponding threshold and the resulting TPR, with their uncertainties, then provide the quantities of operational interest. Methods for constructing the required confidence bounds and simultaneous ROC uncertainty bands are discussed in Section 5.

When the operational requirement is $\mathrm{FPR} \leq \tau$, the goal is not merely to find a sample threshold having $\widehat{\mathrm{FPR}} \leq \tau$. The goal is to estimate a population operating threshold that satisfies the false-positive constraint, and to report uncertainty in both that threshold and the resulting true-positive rate.

4.4 Common Thresholding Shortcuts and Their Assumptions

Several common thresholding rules are legitimate only under specific assumptions.

- **The "$s = 0.5$" default.** A threshold of 0.5 is Bayes-optimal under equal false-alarm and miss costs when the score is a posterior probability calibrated for the deployment population. Rarity by itself does *not* invalidate the 0.5 threshold if the prevalence is already incorporated correctly into that posterior probability. The rule becomes unjustified when s is merely an arbitrary detector score, when the probability calibration corresponds to a different population, or when the two errors have unequal costs.

- **Maximizing Youden's $J = \mathrm{TPR} - \mathrm{FPR}$.** Maximizing J selects the point on the ROC curve corresponding to an iso-cost slope of $m = 1$. In the Bayes formulation this is equivalent to the operating assumption
$$C_{FA}(1 - \pi) = C_{Miss}\pi.$$
That may be appropriate in a particular application, but it is a substantive assumption and should not be adopted merely because J is convenient to compute.

- **Using the EER threshold by default.** The equal-error-rate threshold enforces $\mathrm{FPR} = \mathrm{FNR}$ in the population idealization. As discussed in Section 2, EER has a useful symmetric minimax interpretation, but equality of the two error rates is itself an operating criterion. It should not be treated as universally optimal when costs, prevalence, or operational constraints are asymmetric.

4.5 Precision, Recall, and Prevalence

Recall is another name for the true-positive rate:

$$\mathrm{Recall}(\theta) = \mathrm{TPR}(\theta).$$

It is class-conditional and does not depend on prevalence.

Precision, also called the **positive predictive value (PPV)**, asks a different question: among all trials declared target, what fraction are truly targets? At population prevalence π,

$$\mathrm{PPV}(\theta) = P(Y = 1 \mid S \geq \theta) = \frac{\pi\,\mathrm{TPR}(\theta)}{\pi\,\mathrm{TPR}(\theta) + (1 - \pi)\,\mathrm{FPR}(\theta)}. \tag{40}$$

Thus PPV is not a property of the ranking model alone. It is a property of the detector, threshold, and deployment population together.

If the evaluation sample is representative of the deployment population, empirical precision is

$$\widehat{\mathrm{Precision}}(\theta) = \frac{TP(\theta)}{TP(\theta) + FP(\theta)},$$

whenever the denominator is nonzero. If the evaluation set was deliberately sampled or rebalanced so that its target fraction N_1/N differs from the deployment prevalence π, a deployment-prevalence-adjusted estimate can instead be formed from the class-conditional rate estimates:

$$\widehat{\mathrm{PPV}}_\pi(\theta) = \frac{\pi\,\widehat{\mathrm{TPR}}(\theta)}{\pi\,\widehat{\mathrm{TPR}}(\theta) + (1-\pi)\,\widehat{\mathrm{FPR}}(\theta)}. \tag{41}$$

The Benchmark Portability Trap. Suppose a threshold has TPR $= 0.70$ and FPR $= 0.07$. On a balanced population with $\pi = 0.50$, its precision is

$$\mathrm{PPV} = \frac{0.50(0.70)}{0.50(0.70) + 0.50(0.07)} \approx 0.91.$$

If the same class-conditional operating performance is transported to a population with only 1 % targets, then

$$\mathrm{PPV} = \frac{0.01(0.70)}{0.01(0.70) + 0.99(0.07)} \approx 0.092.$$

Nothing about TPR or FPR changed; the precision fell from about 91 % to about 9.2 % because the deployment prevalence changed.

What happens when no alarms are raised? At a threshold for which $TP = FP = 0$, empirical precision is the undefined ratio 0/0. Many software packages append the point

$$(\mathrm{Recall}, \mathrm{Precision}) = (0, 1)$$

to the plotted precision–recall curve as a convenient boundary convention. This point should not be interpreted as an observed empirical precision of 100 %, nor as a universal population limit as the threshold becomes infinitely strict. The limiting PPV depends on the relative upper tails of the target and nontarget score distributions.

Figure 10 illustrates how the same underlying (FPR, TPR) operating points produce very different precision–recall curves when only the prevalence changes.

4.6 F_β: A Composite Summary, Not an Imbalance Cure

The F_β family is a collection of scalar predictive-performance measures that combine precision and recall into a single number:

$$F_\beta = (1+\beta^2)\frac{\mathrm{Precision} \times \mathrm{Recall}}{\beta^2\,\mathrm{Precision} + \mathrm{Recall}}. \tag{42}$$

The special case $\beta = 1$ gives the familiar harmonic mean

$$F_1 = 2\frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}.$$

Choosing $\beta > 1$ gives recall more weight, while $\beta < 1$ gives precision more weight. This can be useful when that composite objective genuinely matches the application. It should not, however, be adopted merely because the class proportions are imbalanced.

There are three points to keep in mind.

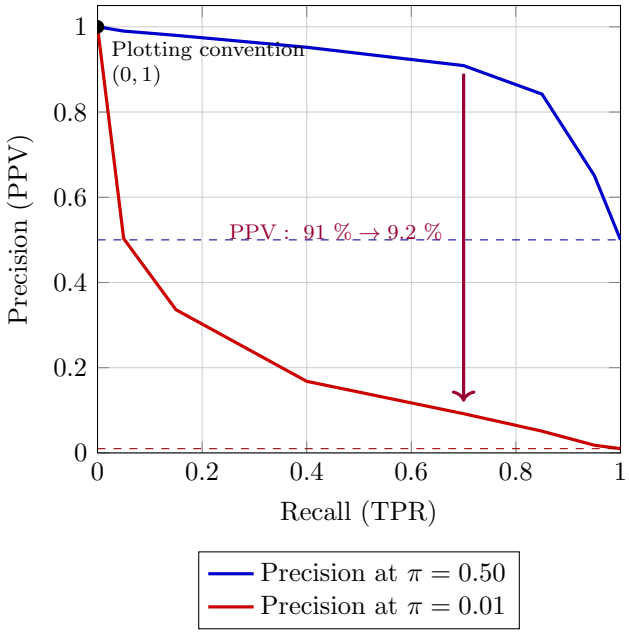


Figure 10: Precision depends on prevalence even when the detector's class-conditional operating characteristics are unchanged. Each curve is computed from the same illustrative set of (FPR, TPR) points, once using $\pi = 0.50$ and once using $\pi = 0.01$. At recall 0.70 with FPR $= 0.07$, precision changes from about 91 % to about 9.2 %. The displayed point $(0, 1)$ is a plotting convention for the no-positive-predictions boundary, not an observed precision value or a universal population limit.

- **Prevalence dependence.** Because precision depends on π, so does F_β. For example, keeping TPR $= 0.95$ and FPR $= 0.01$ fixed gives approximately $F_1 = 0.97$ at $\pi = 0.50$, but only $F_1 \approx 0.16$ at $\pi = 10^{-3}$. The detector's class-conditional discrimination has not changed; the population in which the decision rule is being evaluated has.

- **Non-additivity over trials.** F_β is a nonlinear function of aggregate counts rather than an additive per-trial loss. Consequently, maximizing F_β is not the same problem as minimizing an externally specified expected cost such as C_{det} (Buja et al., 2005).

- **An implicit trade-off is still being chosen.** Optimizing F_1 may appear to avoid specifying the relative consequences of false positives and false negatives, but it does not make that trade-off disappear. Instead, the trade-off is built into the geometry of the F_1 objective. For F_1, the population expression is

$$F_1 = \frac{2\pi\,\mathrm{TPR}}{\pi(1+\mathrm{TPR}) + (1-\pi)\mathrm{FPR}}. \tag{43}$$

  For a fixed F_1 value $k \in [0, 1]$, the corresponding contour has ROC-space slope

$$\frac{k(1-\pi)}{\pi(2-k)}.$$

  At a smooth optimum, matching this contour to a Bayes iso-cost line gives the locally equivalent cost ratio

$$\frac{C_{FA}}{C_{Miss}} = \frac{k}{2-k}.$$

Thus maximizing F_1 implicitly selects a particular local trade-off between the two error types. That trade-off is induced by the metric and the achieved operating point, rather than supplied externally from application-specific costs or consequences.

> **F_β is not invalid.** It is appropriate when the desired objective really is a specified balance of precision and recall, as in many retrieval or triage settings. The mistake is to treat F_1 as an automatic replacement for ROC-based discrimination measures simply because the classes are imbalanced.

4.7 ROC or Precision–Recall? Ask Which Question You Need Answered

ROC and precision–recall curves answer different questions. ROC summarizes the class-conditional trade-off between TPR and FPR. Holding the class-conditional score distributions fixed, the population ROC curve is invariant to prevalence. Precision, by contrast, is a posterior quantity and changes with π according to Equation (40).

Recent work has emphasized that class imbalance does not mathematically invalidate ROC analysis (Richardson et al., 2024; McDermott et al., 2024). Precision–recall analysis is especially useful when the operational question concerns the *yield of alerts*: among the cases sent for review or action, what fraction are truly targets? That question is inherently prevalence-dependent.

> Use ROC or DET when the question is about class-conditional discrimination. Use PPV or precision–recall when the fraction of positive decisions that are truly positive is itself operationally important. Use expected cost when the question is where to set the threshold under specified prevalence and error costs. If the operational requirement is a maximum allowable false-positive rate, use the fixed-FPR operating-point formulation described in Section 4.3. More complicated operational constraints are beyond the scope of this introductory guide. Do not switch from ROC to precision–recall merely because the observed class ratio is skewed.

5 Uncertainty Quantification: What the Evaluation Data Can Support

Every quantity computed from a finite evaluation corpus is an estimate of an unknown population quantity. A point estimate without uncertainty can therefore create a false sense of precision. The appropriate uncertainty calculation depends on three things: *what quantity is being estimated*, *how the evaluation observations were sampled*, and *whether one operating point or an entire curve was inspected*.

5.1 Imbalance Is Not the Same as Sparsity

Class imbalance describes a *ratio*; sampling precision depends on the *absolute amount of information* available in each class. Consider two evaluation sets with the same target-to-nontarget ratio:

$$(N_1, N_0) = (10, 1000) \qquad \text{and} \qquad (N_1, N_0) = (100{,}000, 10{,}000{,}000).$$

Both have a ratio of $1 : 100$, yet they provide radically different precision.

At a fixed, pre-specified threshold and under independent sampling within each class,

$$\widehat{\mathrm{TPR}}(\theta) = \frac{TP(\theta)}{N_1}, \qquad \widehat{\mathrm{FPR}}(\theta) = \frac{FP(\theta)}{N_0}$$

are binomial proportions. Their approximate standard errors are

$$\mathrm{SE}\Big\{\widehat{\mathrm{TPR}}(\theta)\Big\} \approx \sqrt{\frac{\mathrm{TPR}(\theta)\{1-\mathrm{TPR}(\theta)\}}{N_1}}, \tag{44}$$

$$\mathrm{SE}\Big\{\widehat{\mathrm{FPR}}(\theta)\Big\} \approx \sqrt{\frac{\mathrm{FPR}(\theta)\{1-\mathrm{FPR}(\theta)\}}{N_0}}. \tag{45}$$

Thus uncertainty in the target-side rate is governed primarily by N_1, whereas uncertainty in the nontarget-side rate is governed primarily by N_0. The ROC curve depends on both.

For example, if the population TPR at a fixed threshold is 0.80, then with $N_1 = 10$,

$$\mathrm{SE}(\widehat{\mathrm{TPR}}) \approx \sqrt{\frac{0.8(0.2)}{10}} \approx 0.126.$$

The rough margins above are based on the usual Wald normal approximation (Wald, 1939; Brown et al., 2001). They are useful for intuition, but Wald intervals can have poor coverage, particularly for small samples or when the estimated probability is near 0 or 1 (Brown et al., 2001).

A commonly recommended alternative is the Wilson score interval (Wilson, 1927; Brown et al., 2001). An exact binomial interval, such as the Clopper–Pearson interval (Clopper and Pearson, 1934), is another option when conservative coverage is acceptable. Clopper–Pearson intervals typically have coverage at least as large as the nominal level, but can be wider than necessary.

Class imbalance and sample sparsity are different problems. A highly imbalanced evaluation can still estimate class-conditional rates very precisely if both class-specific sample sizes are large. Conversely, a mildly imbalanced study can be very uncertain if one class contains only a handful of independent observations.

5.2 The Sampling Unit Matters: Clustered and Repeated Data

The binomial formulas above assume independent observations within each class. That assumption can fail badly in modern detector evaluations. For example:

- many face images may come from the same person;
- many medical scans may come from the same patient;
- many adversarial perturbations may be generated from the same source item; and
- repeated model runs may share the same prompt, episode, or underlying event.

Observations from the same subject or source are typically correlated. Treating all rows as independent can therefore make confidence intervals much too narrow.

The practical rule is simple: **the resampling or variance calculation must respect the independent sampling unit**. If subjects are independently sampled but each subject contributes several observations, a cluster bootstrap should resample subjects and carry all of each selected subject's observations with them. More elaborate hierarchical models are possible, but the raw row count should not be substituted for the number of independent sampling units.

5.3 Sampling Uncertainty Does Not Account for Population Shift

The uncertainty procedures in this guide quantify variation arising from finite sampling under the population and sampling design represented by the evaluation data. They do *not* by themselves protect against systematic differences between that evaluation population and a later deployment population. Changes in demographics, sensors, image quality, acquisition environment, or other operating conditions can alter the class-conditional score distributions and therefore change both error rates and the threshold required to satisfy an operational constraint.

This distinction is important when interpreting familiar procedures. A Wilson or other binomial interval quantifies sampling uncertainty for a class-conditional rate at a fixed, prespecified threshold, under its sampling assumptions. A bootstrap that re-estimates the threshold in every replicate can additionally propagate sampling uncertainty caused by estimating that threshold from finite data. Neither procedure, by itself, accounts for a deployment shift that was not represented in the evaluation sample.

In operational use, a threshold may therefore need to be reassessed as representative deployment data become available. Such iterative adjustment can be entirely appropriate, but it constitutes a new threshold-selection step. If the same data are used both to adjust the threshold and to report its apparent performance, the resulting assessment is data-adaptive rather than an evaluation of a prespecified operating rule. A cleaner assessment uses independent or subsequently collected data for the final evaluation, or an inferential procedure that explicitly incorporates the threshold-selection step.

Sampling uncertainty and population shift are different problems. Confidence intervals and bootstrap bands describe what finite evaluation data can support for the population they represent. They do not guarantee that the same error rates or operating threshold will transfer to a different deployment population.

5.4 Pointwise Intervals and Simultaneous Confidence Bands Answer Different Questions

Suppose a threshold θ_0 was fixed *before* inspecting the evaluation labels. A confidence interval for $\mathrm{TPR}(\theta_0)$ or $\mathrm{FPR}(\theta_0)$ is then a *pointwise* interval: it quantifies uncertainty at that one operating point.

A different question arises when the entire ROC curve is inspected, perhaps to choose a threshold after looking at many candidate operating points. A collection of ordinary 95 % pointwise intervals does not form a 95 % confidence band for the whole curve. The more locations we inspect, the easier it becomes for at least one pointwise interval to miss its target.

A *simultaneous confidence band* is designed to cover the population ROC curve over an entire range of false-positive rates with one stated confidence level (Macskassy et al., 2005; Luckett et al., 2021). This is the appropriate object when conclusions are drawn after examining the curve as a whole.

A fixed-threshold confidence interval and an interval at a threshold *estimated from the same data* are not the same problem. For quantities such as EER or TPR at a specified population FPR, the defining threshold is itself estimated. The uncertainty procedure should re-estimate that threshold, rather than pretending it was known in advance.

5.5 A Distribution-Free ROC Band: DKW–Massart

One conservative way to obtain simultaneous ROC uncertainty starts from uniform confidence bands for the two class-conditional score distributions. The Dvoretzky–Kiefer–Wolfowitz inequality, with Massart's sharp constant, provides such a band without assuming a parametric score distribution (Massart, 1990).

Let $1-\alpha$ be the desired joint confidence level. Allocate failure probabilities $\delta_0 > 0$ and $\delta_1 > 0$ to the nontarget and target samples, with

$$\delta_0 + \delta_1 = \alpha.$$

For class-specific sample sizes N_0 and N_1, define

$$\varepsilon_0 = \sqrt{\frac{\ln(2/\delta_0)}{2N_0}}, \qquad \varepsilon_1 = \sqrt{\frac{\ln(2/\delta_1)}{2N_1}}. \tag{46}$$

Any positive values δ_0 and δ_1 satisfying $\delta_0 + \delta_1 = \alpha$ give the stated Bonferroni coverage. In the absence of a specific operational reason to favor one component, a simple default is the equal allocation

$$\delta_0 = \delta_1 = \alpha/2.$$

An unequal allocation may be specified in advance when one component of the ROC uncertainty is more important. For example, low-false-positive-rate applications may allocate a larger share of the failure probability to δ_0, thereby reducing the horizontal margin ε_0, whereas a small target sample may motivate a larger δ_1 to reduce the vertical margin ε_1. Because the allocation affects the confidence band, it should not be chosen after inspecting the observed ROC curve merely to obtain a narrower-looking band.

With probability at least $1-\alpha$, both class-conditional empirical CDFs are then uniformly within their respective margins of the population CDFs.

To avoid pretending that an empirical ROC can attain every exact false-positive rate, define the population ROC function as the best TPR attainable without exceeding a false-positive-rate budget u:

$$R(u) = \sup_{\theta:\,\mathrm{FPR}(\theta)\le u} \mathrm{TPR}(\theta), \qquad 0 \le u \le 1, \tag{47}$$

and define its empirical counterpart by

$$\widehat{R}(u) = \max_{\theta:\,\widehat{\mathrm{FPR}}(\theta)\le u} \widehat{\mathrm{TPR}}(\theta). \tag{48}$$

This definition naturally handles the stepwise empirical ROC and does not require an observed threshold with $\widehat{\mathrm{FPR}} = u$ exactly.

A conservative simultaneous envelope is then

$$R_L(u) = \begin{cases} 0, & u \le \varepsilon_0, \\ \max\left\{0, \widehat{R}(u-\varepsilon_0) - \varepsilon_1\right\}, & u > \varepsilon_0, \end{cases} \tag{49}$$

and

$$R_U(u) = \begin{cases} \min\left\{1, \widehat{R}(u+\varepsilon_0) + \varepsilon_1\right\}, & u < 1-\varepsilon_0, \\ 1, & u \ge 1-\varepsilon_0. \end{cases} \tag{50}$$

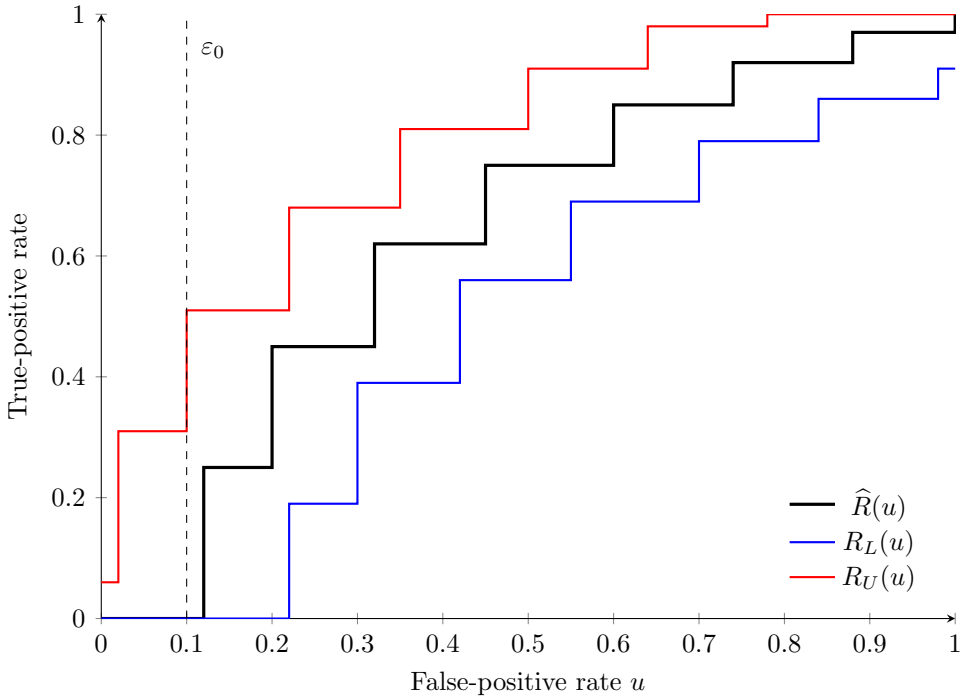


Figure 11: Illustration of the simultaneous DKW–Massart ROC envelope. Here $\varepsilon_0 = 0.10$ and $\varepsilon_1 = 0.06$. The lower envelope is obtained by shifting the empirical ROC $\widehat{R}(u)$ to the right by ε_0 and downward by ε_1, with truncation at zero. The upper envelope is obtained by shifting it to the left by ε_0 and upward by ε_1, with truncation at one. The vertical dashed line marks $u = \varepsilon_0$. The resulting envelope holds simultaneously over the ROC curve on the joint DKW event.

Define the joint DKW event

$$\mathcal{E} = \left\{ \sup_t \left| \widehat{F}_0(t) - F_0(t) \right| \le \varepsilon_0 \right\} \cap \left\{ \sup_t \left| \widehat{F}_1(t) - F_1(t) \right| \le \varepsilon_1 \right\}.$$

By the DKW–Massart inequality and the allocation $\delta_0 + \delta_1 = \alpha$,

$$P(\mathcal{E}) \ge 1 - \alpha.$$

On this event,

$$R_L(u) \le R(u) \le R_U(u) \qquad \text{for every } u \in [0, 1].$$

Figure 11 illustrates the geometry of this distribution-free simultaneous envelope.

The band is deliberately conservative. In particular, when u is smaller than the horizontal uncertainty margin ε_0, the data may be unable to provide a useful positive lower bound on the detection rate. That loss of precision is not a defect of ROC analysis; it is a statement about how little nontarget-tail information the evaluation contains.

5.6 A Practical Alternative: Recompute the Entire ROC in a Stratified Bootstrap

A stratified bootstrap is often a more data-adaptive practical alternative to the distribution-free DKW band. The key is to bootstrap the *entire ROC construction*, not only the target-side detection proportions.

For independent observations, one bootstrap replicate proceeds as follows:

1. resample the N_1 target observations with replacement;
2. separately resample the N_0 nontarget observations with replacement;
3. recompute the complete empirical ROC curve on a common FPR grid; and
4. retain the deviation of the bootstrap ROC from the original empirical ROC over the whole grid.

Repeating these steps many times gives an empirical distribution for the entire ROC process. A simultaneous band can then be based on a bootstrap quantile of the largest deviation over the chosen FPR range. For example, a studentized construction uses

$$T_b^* = \sup_{u \in \mathcal{U}} \frac{\left| \widehat{R}_b^*(u) - \widehat{R}(u) \right|}{\widehat{\sigma}(u)}, \tag{51}$$

where $\widehat{\sigma}(u)$ is a pointwise standard-error estimate and $\mathcal{U}$ is the FPR range to be covered. If $c_{1-\alpha}$ is the $(1-\alpha)$ bootstrap quantile of T_b^*, an approximate simultaneous band has the form

$$\left[\max\{0, \widehat{R}(u) - c_{1-\alpha}\widehat{\sigma}(u)\},\ \min\{1, \widehat{R}(u) + c_{1-\alpha}\widehat{\sigma}(u)\} \right]. \tag{52}$$

Other bootstrap constructions are possible; the important principle is that the threshold-selection step and both class-specific samples are recomputed in each replicate.

This point matters because an ROC ordinate at a specified FPR is *not* just a binomial proportion from the target sample. The threshold corresponding to that FPR is estimated from the nontarget sample. Consequently, uncertainty in an ROC ordinate generally depends on both N_1 and N_0.

If the original data are clustered, the bootstrap should resample independent clusters rather than individual rows. Bootstrap bands are approximate rather than finite-sample distribution-free guarantees, so the resampling design and the FPR range $\mathcal{U}$ should be stated explicitly.

Figure 12 compares the geometry of the conservative distribution-free DKW–Massart envelope with a data-adaptive stratified-bootstrap simultaneous band for the same empirical ROC curve.

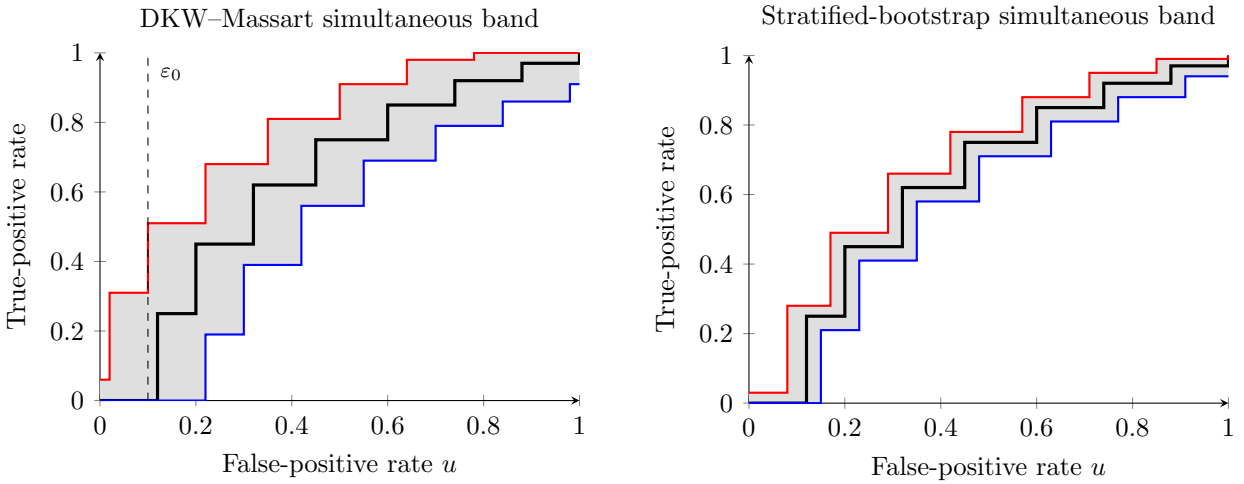


Figure 12: Illustrative comparison of two simultaneous confidence-band constructions for the same empirical ROC curve $\widehat{R}(u)$ (black). Left: a DKW–Massart simultaneous band, which is distribution-free and conservative. Right: a stratified-bootstrap simultaneous band, which is data-adaptive and often narrower in practice. The comparison is illustrative only; no universal nesting relationship is implied between the two methods.

The DKW–Massart envelope is useful primarily as a conservative, distribution-free benchmark. In practice, a properly constructed stratified or cluster bootstrap can provide a more data-adaptive and substantially narrower simultaneous band, although its coverage is approximate rather than a finite-sample distribution-free guarantee.

5.7 Uncertainty for Operating Thresholds and Expected Cost

Section 4 introduced two operational questions that require more than a point estimate of the ROC curve. The same uncertainty machinery can be used to answer them.

5.7.1 A fixed population false-positive-rate requirement

Suppose the operational requirement is

$$\mathrm{FPR}(\theta) \le \tau,$$

as in Section 4.3. The population threshold θ_τ and the resulting $\mathrm{TPR}(\theta_\tau)$ are both unknown. A useful uncertainty analysis should therefore report uncertainty for *both* quantities.

A stratified bootstrap can do this directly. In every bootstrap replicate, re-estimate the nontarget threshold satisfying the FPR requirement and then recompute the target detection rate at that re-estimated threshold. The resulting bootstrap distributions describe uncertainty in

$$\widehat{\theta}_\tau \qquad \text{and} \qquad \widehat{\mathrm{TPR}}\left(\widehat{\theta}_\tau\right),$$

while propagating uncertainty from both classes.

If the operational statement is genuinely that the *population* FPR must not exceed τ, merely observing $\widehat{\mathrm{FPR}}(\theta) \le \tau$ is not enough. A more defensible selection rule requires an upper confidence bound for the population FPR to satisfy

$$U_{\mathrm{FPR}}(\theta) \le \tau. \tag{53}$$

When the threshold is selected after scanning many candidates, the upper bound should come from a procedure that accounts for that selection, such as a simultaneous one-sided band for the nontarget tail probability.

At very small τ, the nontarget sample size becomes decisive. With N_0 independent nontarget observations, the ordinary empirical FPR moves in steps of $1/N_0$. If the required τ is much smaller than $1/N_0$, the data alone contain little or no nonparametric information about that tail. More data or an explicit modeling assumption is then required.

5.7.2 The population minimum expected cost

For the expected-cost formulation in Equation (31), define

$$a = C_{Miss}\pi, \qquad b = C_{FA}(1-\pi).$$

Writing the population ROC as $R(u)$ gives

$$C^*_{det} = \inf_{0\le u\le 1} \left[a\{1 - R(u)\} + bu\right]. \tag{54}$$

Suppose a simultaneous $(1-\alpha)$ ROC band satisfies

$$R_L(u) \le R(u) \le R_U(u) \qquad \text{for every } u$$

with the stated joint coverage. Because larger TPR means smaller expected cost, the same band immediately yields

$$C_L^* = \inf_{0\le u\le 1} \left[a\{1 - R_U(u)\} + bu\right], \tag{55}$$

$$C_U^* = \inf_{0\le u\le 1} \left[a\{1 - R_L(u)\} + bu\right]. \tag{56}$$

The infimum appears in both bounds because the population quantity of interest is itself the minimum expected cost over all operating points. The upper ROC envelope gives the optimistic cost curve and hence the lower bound on that minimum, while the lower ROC envelope gives the pessimistic cost curve and hence the upper bound.

Because the ROC band has simultaneous coverage of at least $1 - \alpha$, the induced bounds satisfy

$$\Pr(C_L^* \le C_{\text{det}}^* \le C_U^*) \ge 1 - \alpha. \tag{57}$$

Thus $[C_L^*, C_U^*]$ is a possibly conservative $100(1-\alpha)$ % confidence interval for the population minimum expected cost, provided that the prevalence and error costs are treated as fixed.

This is different from simply reporting the smallest empirical cost observed on the evaluation sample. The latter is a sample-optimized quantity; the interval above makes an inferential statement about the unknown population minimum expected cost.

The same idea supports the *range of defensible operating thresholds* introduced in Section 4. Let

$$C_L(\theta) \le C_{det}(\theta) \le C_U(\theta)$$

be a simultaneous confidence band for the population cost curve, and let

$$C_U^* = \inf_\theta C_U(\theta).$$

Any threshold satisfying

$$C_L(\theta) \le C_U^* \tag{58}$$

cannot be ruled out, at the band confidence level, as attaining the population minimum. More generally, if an excess cost no larger than ε is considered practically negligible, a conservative set of thresholds not ruled out as ε-near-optimal is

$$\widehat{\Theta}_\varepsilon^{\text{def}} = \{\theta : C_L(\theta) \le C_U^* + \varepsilon\}. \tag{59}$$

This set may be one interval or several disjoint intervals. Figure 13 illustrates how either geometry can arise. The phrase "range of defensible operating thresholds" is therefore shorthand for the set of thresholds the evaluation cannot distinguish from the optimum, or from a pre-specified practically negligible neighborhood of it.

The construction above treats π, C_{FA}, and C_{Miss} as externally specified constants. If any of them are themselves estimated, their uncertainty must also be propagated.

For threshold selection, the quantities of interest are population quantities, such as the minimum expected cost C_{det}^* or the true-positive rate at a specified false-positive-rate requirement, $\text{TPR}(\theta_\tau)$. Because these quantities are estimated from finite data, uncertainty should be reported. The evaluation should also indicate which operating thresholds are reasonably supported by the data.

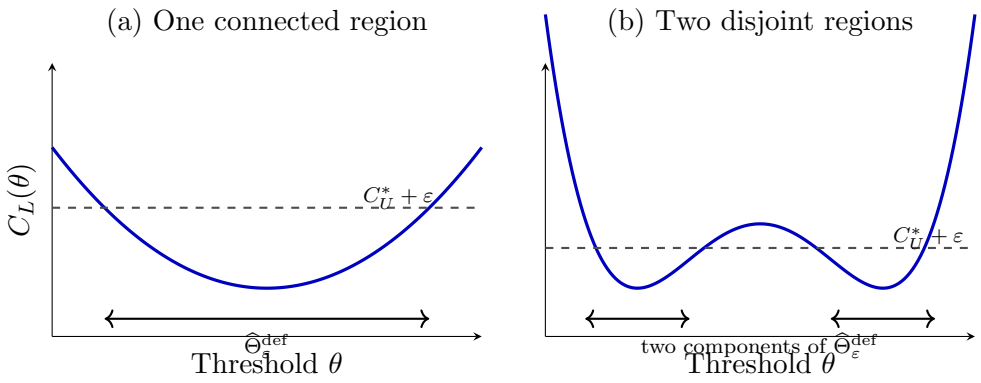


Figure 13: Schematic examples of defensible-threshold sets. A threshold belongs to $\widehat{\Theta}_\varepsilon^{\text{def}}$ when $C_L(\theta) \le C_U^* + \varepsilon$. Left: one broad low-cost region yields a connected interval. Right: two separated low-cost regions yield two disjoint intervals. The curves are illustrative.

5.8 Uncertainty for Other Common Metrics

The same principle applies beyond the ROC curve.

AUC. The empirical AUC is a finite-sample estimate of a population pairwise-ranking probability. Established large-sample methods and bootstrap procedures can be used to construct confidence intervals (Hanley and McNeil, 1982). When two AUCs are computed on the same evaluation items, their estimates are correlated; the comparison procedure should preserve that pairing.

EER. EER is defined through a crossing of two estimated error-rate functions. Its uncertainty should therefore account for uncertainty in the crossing threshold, for example by recomputing the EER inside each bootstrap replicate. Attaching a binomial interval to the empirical EER value while treating its threshold as fixed understates the uncertainty.

PPV / precision. PPV depends on TPR, FPR, and deployment prevalence through Equation (40). If prevalence is externally estimated rather than fixed, its uncertainty should be propagated along with uncertainty in the two class-conditional rates. A precision estimate from an artificially balanced or case–control evaluation should not be interpreted as the precision expected in deployment unless the deployment prevalence is properly accounted for.

Calibration quantities. Reliability diagrams, binned Expected Calibration Error, Brier score, and likelihood-ratio calibration summaries are also finite-sample quantities. Reliability diagrams have sampling variability in the observed target fraction, and binned summaries additionally depend on the binning scheme. If calibration is claimed over a range of scores, simultaneous uncertainty over that range is more informative than isolated pointwise error bars. Specialized procedures are available for uncertainty-aware likelihood-ratio calibration claims (Hannig and Iyer, 2022).

Key Takeaway for Practitioners

There is no universal "uncertainty bar" that can be attached mechanically to every detector metric. Match the uncertainty procedure to the population quantity, the threshold-selection rule, and the sampling design. Pointwise rates, full ROC curves, operating thresholds, minimum

expected cost, PPV, AUC, and calibration summaries are different inferential objects.

6 Evaluating Adversarial Attack Efficacy

Adversarial evaluation asks a different question from ordinary detector benchmarking. The object of interest is not only how well the detector performs, but how much a deliberate manipulation changes its behavior. For binary detectors, this distinction is especially important because an attack commonly modifies examples from only one class. Attack efficacy should therefore be assessed first on the class that was actually manipulated. Detector-level quantities such as ROC curves and AUC can then be used as secondary summaries of the resulting two-class system.

Whenever possible, the same source items should be evaluated both before and after attack. For item i, write

$$s_i^{\text{clean}} \qquad \text{and} \qquad s_i^{\text{adv}}$$

for the clean and adversarially attacked scores. This paired design lets the evaluator ask what the attack changed for each item rather than comparing two unrelated samples.

6.1 Two Common Class-Conditional Attack Objectives

Two common attack objectives are:

- **Evasion (Targets → Nontargets).** The attacker modifies true targets so that their scores move downward. At a threshold θ, successful evasion causes additional misses.
- **False-alarm induction or spoofing (Nontargets → Targets).** The attacker modifies true nontargets so that their scores move upward. At a threshold θ, successful spoofing causes additional false alarms.

The first principle for attack evaluation is therefore simple:

Evaluate the attack first on the class it actually modified.

For an evasion attack, the primary data are the clean and attacked target scores. For a spoofing attack, the primary data are the clean and attacked nontarget scores. Quantities that additionally involve the unattacked class answer a broader detector-level question and should be identified as such.

Constructing the attacked corpus. The evaluation unit is the original source item, not an individual attack attempt. The same source items should remain represented before and after attack. An attack procedure may make several admissible attempts on a source item, but those attempts are part of the attack strategy rather than separate evaluation trials. For each source item, the procedure therefore produces one final attacked outcome under the stated attack budget. This preserves the source item as the evaluation unit and ensures that unsuccessful attacks remain represented in the corpus.

For a corpus-level evaluation, the attacker may try several allowable perturbations of an item and retain the one that best advances the attack objective, subject to a prespecified attack budget.

For an evasion attack on target item i, a natural construction is

$$s_i^{\text{adv}} = \min\left\{ s_i^{\text{clean}}, s_{i1}^{\text{attack}}, \ldots, s_{iK}^{\text{attack}} \right\}.$$

For a spoofing attack on nontarget item j, the analogous construction is

$$s_j^{\text{adv}} = \max\left\{ s_j^{\text{clean}}, s_{j1}^{\text{attack}}, \ldots, s_{jK}^{\text{attack}} \right\}.$$

Thus a candidate perturbation that moves the score in the wrong direction is not retained, but the corresponding source item is not removed from the evaluation. If no allowed perturbation improves the attack objective, the clean score is retained. Excluding unsuccessful source items would condition the evaluation on attack success and would overstate corpus-level attack efficacy. The number of attempts, query limit, perturbation limit, or other search restriction is part of the attack budget and should be reported.

6.2 Attack Success at a User-Facing Threshold

Attack Success Rate (ASR) is widely used in adversarial evaluation (Biggio et al., 2013; Carlini and Wagner, 2017). Terminology for ASR is not completely uniform across the literature. In this guide, $\text{ASR}(\theta)$ denotes the fraction of all attacked items that end up on the wrong side of threshold θ. We separately report the conditional flip rate among items that were correctly handled before attack, calling that quantity the *induced ASR*.

At a threshold θ, the population attack success rate for an evasion attack is

$$\text{ASR}_{\text{evasion}}(\theta) = P\Big(S^{\text{adv}} < \theta \mid Y = 1\Big) = \text{FNR}^{\text{adv}}(\theta), \tag{60}$$

and for a spoofing attack it is

$$\text{ASR}_{\text{spoof}}(\theta) = P\Big(S^{\text{adv}} \geq \theta \mid Y = 0\Big) = \text{FPR}^{\text{adv}}(\theta). \tag{61}$$

On finite evaluation samples, the corresponding estimates are

$$\widehat{\text{ASR}}_{\text{evasion}}(\theta) = \frac{1}{N_1} \sum_{i:y_i=1} \mathbf{1}\Big\{ s_i^{\text{adv}} < \theta \Big\}, \tag{62}$$

$$\widehat{\text{ASR}}_{\text{spoof}}(\theta) = \frac{1}{N_0} \sum_{j:y_j=0} \mathbf{1}\Big\{ s_j^{\text{adv}} \geq \theta \Big\}. \tag{63}$$

Why $\theta = 0.5$ can still be worth reporting. Many detectors return a score in $[0, 1]$ and present it to users as a probability, confidence, or probability-like quantity. In practice, users may naturally interpret 0.5 as the boundary between the two classes, especially when no ROC curve or application-specific threshold is supplied. In that setting, attack success at $\theta = 0.5$ answers a legitimate user-facing question: *does the attack change the decision made under the detector's natural default interpretation?*

This does not imply that 0.5 is a calibrated Bayes-optimal threshold. It should be described as a *nominal* or *user-facing* threshold unless the probability and decision assumptions needed to justify 0.5 have actually been established.

Separate post-attack error from attack-induced change. The post-attack ASR in Equations (62) and (63) includes items that the detector may already have classified incorrectly before the attack. For an evasion attack, the population increase in miss rate is

$$\Delta_{\text{FNR}}(\theta) = \text{FNR}^{\text{adv}}(\theta) - \text{FNR}^{\text{clean}}(\theta). \tag{64}$$

For a spoofing attack, the analogous increase is

$$\Delta_{\mathrm{FPR}}(\theta) = \mathrm{FPR}^{\mathrm{adv}}(\theta) - \mathrm{FPR}^{\mathrm{clean}}(\theta). \tag{65}$$

These quantities measure the net change in class-conditional error at the specified threshold.

A second useful quantity asks a different question: among items that the clean detector handled correctly, what fraction did the attack flip to the wrong side of the threshold? For evasion,

$$\mathrm{ASR}_{\mathrm{ind,evasion}}(\theta) = P\Big(S^{\mathrm{adv}} < \theta \mid S^{\mathrm{clean}} \geq \theta,\, Y = 1\Big), \tag{66}$$

and for spoofing,

$$\mathrm{ASR}_{\mathrm{ind,spoof}}(\theta) = P\Big(S^{\mathrm{adv}} \geq \theta \mid S^{\mathrm{clean}} < \theta,\, Y = 0\Big). \tag{67}$$

Their empirical versions are obtained by replacing the probabilities with the corresponding paired sample fractions.

The net error increase in Equations (64) and (65) and the conditional flip rates in Equations (66) and (67) are not the same quantity. Both can be informative: the first measures the change in the overall class-conditional error rate, while the second focuses on previously successful clean decisions.

Uncertainty should preserve the pairing. If the same items are evaluated clean and attacked, uncertainty calculations should use that paired structure. For example, for target item i define

$$D_i(\theta) = \mathbf{1}\Big\{s_i^{\mathrm{adv}} < \theta\Big\} - \mathbf{1}\Big\{s_i^{\mathrm{clean}} < \theta\Big\}.$$

The sample mean of $D_i(\theta)$ estimates $\Delta_{\mathrm{FNR}}(\theta)$. A paired bootstrap, or another method for paired binary outcomes, can be used to obtain a confidence interval. If items are themselves clustered, the resampling should occur at the independent cluster level as discussed in Section 5.2.

At a threshold used by the detector's users, report the post-attack error rate, the change from clean performance, and, when useful, the fraction of clean successes that the attack flips. These quantities answer different questions and should be accompanied by uncertainty.

6.3 How Far Did the Attacked Scores Move?

A threshold-based success rate records whether a score crossed a particular boundary, but it does not show how far the score moved. The score values can therefore provide useful supplementary information.

Probability or probability-like outputs. If the detector output is intended to be a probability for $Y = 1$, the class-conditional Brier components introduced in Section 3 give simple score-scale diagnostics. For an evasion attack,

$$\widehat{\mathrm{BrierT}}_{\mathrm{clean}} = \frac{1}{N_1} \sum_{i:y_i=1} \left(s_i^{\mathrm{clean}} - 1\right)^2, \tag{68}$$

$$\widehat{\mathrm{BrierT}}_{\mathrm{adv}} = \frac{1}{N_1} \sum_{i:y_i=1} \left(s_i^{\mathrm{adv}} - 1\right)^2, \tag{69}$$

and the attack-associated change is

$$\Delta\widehat{\mathrm{BrierT}} = \widehat{\mathrm{BrierT}}_{\mathrm{adv}} - \widehat{\mathrm{BrierT}}_{\mathrm{clean}}. \tag{70}$$

For spoofing, the analogous quantities use

$$\widehat{\mathrm{BrierN}} = \frac{1}{N_0} \sum_{j:y_j=0} s_j^2.$$

An increase in BrierT means that attacked target scores have moved, on average, farther from the target endpoint 1; an increase in BrierN means that attacked nontarget scores have moved farther from the nontarget endpoint 0.

The formal probability-forecast interpretation of the Brier score requires the scores to represent probabilities. Some detectors, however, present a number in $[0, 1]$ as a confidence or probability-like score without establishing calibration. In that situation, BrierT and BrierN can still describe movement on the numerical scale that the user sees, but they should not be interpreted as proving probability calibration or probabilistic accuracy.

For probability-like detector scores, 0.25 provides a useful reference value for the class-conditional Brier components. A detector that reports the neutral value 0.5 for every trial has

$$\mathrm{BrierT} = (0.5 - 1)^2 = 0.25$$

and

$$\mathrm{BrierN} = (0.5 - 0)^2 = 0.25.$$

Thus 0.25 corresponds to a detector that expresses no preference between the two classes. Values below 0.25 indicate better class-conditional squared-error performance than this neutral reference for that class, whereas values above 0.25 indicate worse performance.

If the deployment prevalence π is known and is used as the constant no-feature probability forecast, then the corresponding reference values are instead

$$\mathrm{BrierT} = (1 - \pi)^2, \qquad \mathrm{BrierN} = \pi^2.$$

Thus the appropriate reference depends on what information is assumed available. When prevalence is unknown, or when a symmetric probability reference is desired, 0.5 and the corresponding Brier value 0.25 provide a natural benchmark.

At the other extreme, BrierT = 1 occurs when all evaluated target scores are 0, and BrierN = 1 occurs when all evaluated nontarget scores are 1. These values indicate maximal class-conditional squared error on the 0-to-1 score scale, but they do not imply failure at every mathematically possible decision threshold.

General score scales. If the detector score is not a probability-like quantity, a direct paired score shift is easier to interpret. For evasion define

$$\Delta S_i^{\mathrm{evasion}} = s_i^{\mathrm{clean}} - s_i^{\mathrm{adv}}, \tag{71}$$

so positive values represent downward movement of target scores. For spoofing, define

$$\Delta S_j^{\mathrm{spoof}} = s_j^{\mathrm{adv}} - s_j^{\mathrm{clean}}, \tag{72}$$

so positive values represent upward movement of nontarget scores. The mean, median, quantiles, or full distribution of these paired shifts can be reported with uncertainty. These summaries are detector-scale dependent, so they are most useful when comparing attacks applied to the same detector and score scale.

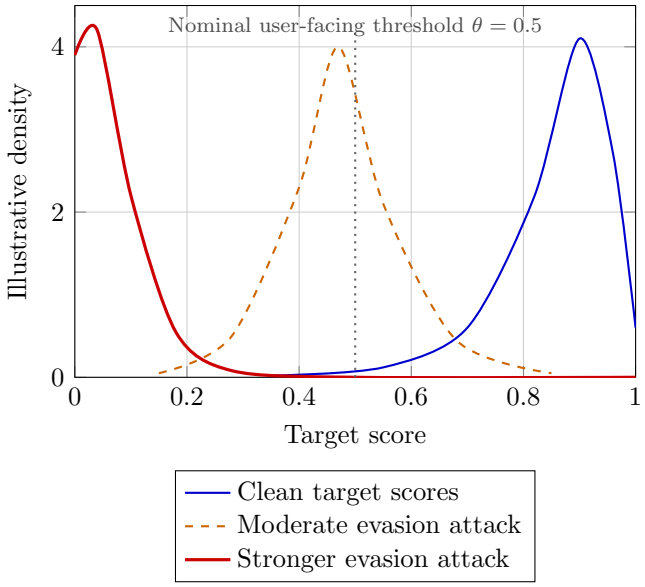


Figure 14: A class-conditional view of an evasion attack. The figure uses only clean and attacked target scores. A threshold-based ASR records how many attacked scores cross the nominal decision boundary, while paired score-shift or probability-scale diagnostics describe how far the target scores moved.

6.4 Operational Consequence at a Specified System Threshold

The primary attack-efficacy measures above concern the class that was attacked. A separate question is how that class-conditional change affects the detector at an operating threshold chosen for deployment.

For example, suppose an evasion attack modifies targets while the nontarget population is not attacked. If the clean operating policy requires FPR $\leq \tau$, let θ_τ denote the population threshold determined by the clean nontarget distribution, as in Section 4.3. The same threshold can then be applied to clean and attacked targets. The operational loss in detection rate is

$$L_{\mathrm{TPR}}(\tau) = \mathrm{TPR}^{\mathrm{clean}}(\theta_\tau) - \mathrm{TPR}^{\mathrm{adv}}(\theta_\tau). \tag{73}$$

This has a direct interpretation: it is the reduction in target detection under the specified attack procedure and attack budget at the operating point established from the nontarget population.

For a spoofing attack on nontargets, the analogous detector-level consequence at a pre-specified threshold is the increase

$$\mathrm{FPR}^{\mathrm{adv}}(\theta) - \mathrm{FPR}^{\mathrm{clean}}(\theta). \tag{74}$$

If the threshold itself was estimated from finite reference data, its uncertainty should be propagated as described in Section 5.7. If the threshold is fixed in advance, the clean-versus-attacked comparison on the manipulated class remains paired.

Changing the threshold after seeing the attacked data answers a different question: how well can the defender adapt after learning about the attack? That may be important in a separate robustness analysis, but it should not be confused with the direct effect of the attack on the originally specified operating rule.

6.5 ROC and AUC as Secondary Detector-Level Summaries

ROC curves and AUC involve both classes. They are therefore not intrinsic measures of the success of an attack that modifies only one class.

For an evasion attack, for example, the attacked AUC relative to a fixed nontarget reference population is

$$\mathrm{AUC}_{\mathrm{adv}} = P\left(S_1^{\mathrm{adv}} > S_0\right) + \frac{1}{2}P\left(S_1^{\mathrm{adv}} = S_0\right), \tag{75}$$

where S_1^{adv} is an attacked target score and S_0 is a score from the nontarget reference population. The clean AUC uses S_1^{clean} with the same nontarget reference.

Consequently,

$$\Delta\mathrm{AUC} = \mathrm{AUC}_{\mathrm{adv}} - \mathrm{AUC}_{\mathrm{clean}}$$

describes the effect of the attack on two-class discrimination *relative to that particular nontarget population*. If the nontarget reference population changes, the attacked AUC can change even when the attack on the targets is unchanged. The analogous statement applies to spoofing attacks, where the nontarget distribution is attacked and the target distribution forms the reference class.

ROC curves can still be useful as secondary diagnostics when the same unattacked reference population is used for the clean and attacked comparisons. They show how the class-conditional score shift changes the operating trade-off across thresholds. In many applications, the change at a relevant operating point—for example the loss in TPR at a specified clean-reference FPR—is more directly interpretable than the change in whole-curve AUC.

Uncertainty calculations for clean-versus-attacked ROC quantities should preserve both the clean/attacked pairing on the manipulated class and the sampling variation in the reference class. A paired, stratified bootstrap is a natural implementation of this principle; see Section 5.6.

An attacked ROC curve is a property of the attacked class distribution *relative to a reference distribution from the other class*. It should not be presented as a pure measure of attacker success when only one class was modified.

6.6 Attack Efficacy Versus Attack Budget

Attack efficacy should be interpreted together with the effort, perturbation, or distortion required to achieve it. Let ϵ denote an attack budget or attack-strength parameter. Depending on the application, ϵ might represent a norm-bounded image perturbation, a perceptual distortion measure, a number of allowed edits or queries, a time or compute budget, or another application-specific notion of attack cost.

At each value of ϵ, the evaluation can report one or more quantities from the preceding subsections, for example

- user-facing ASR at a stated nominal threshold;
- net increase in FNR or FPR;
- induced ASR among clean successes;
- paired score displacement;
- ΔBrierT or ΔBrierN when the score scale supports that interpretation; and
- loss in TPR or increase in FPR at a pre-specified operational threshold.

Each reported trajectory should include uncertainty, and the clean/attacked pairing should be preserved in the analysis.

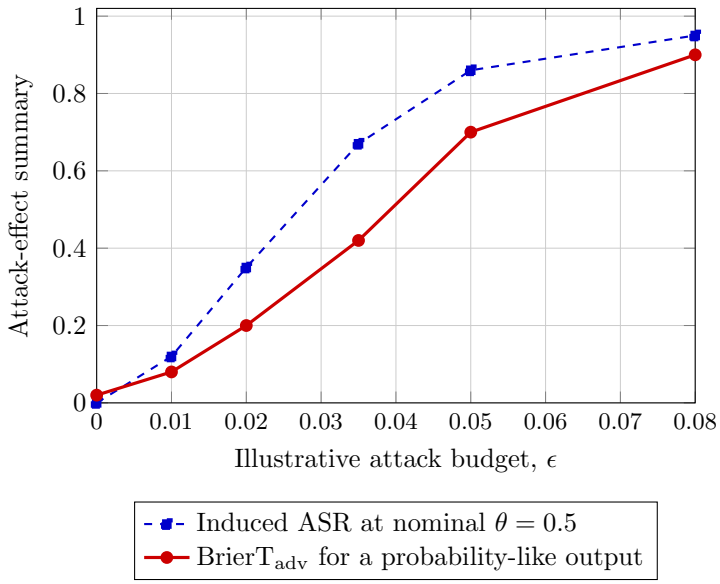


Figure 15: Illustrative attack-efficacy trajectories. Different summaries answer different questions: the induced ASR records decision flips at the stated user-facing threshold, while BrierT$_{\text{adv}}$ describes movement on a probability-like target score scale. In an actual evaluation, confidence intervals or bands should accompany the trajectories.

A complete adversarial evaluation should identify which class was attacked, describe how the attacked corpus was constructed, compare clean and attacked outcomes on the same source items when possible, and report uncertainty. Threshold-based ASR describes post-attack error at a stated threshold; changes in class-conditional error and induced ASR describe attack-induced changes in decisions. Paired score summaries describe how far the manipulated class moved on the detector's score scale, while ROC/AUC describe the resulting two-class system relative to a specified reference population. Attack efficacy should also be reported as a function of the relevant attack budget or cost.

7 Practitioner's Reference Matrix and Red Flags

The purpose of this section is not to identify a single "best" metric. Different quantities answer different questions. A useful evaluation begins by stating the question of interest and then choosing the quantity, display, and uncertainty analysis that address that question.

7.1 Matching the Question to the Quantity

Table 1: A question-first reference matrix for binary-detector evaluation.

Question	**Primary quantity or display**	**Interpretation and important caution**
How does the detector trade target detection against false alarms as the threshold changes?	ROC curve or DET curve	Shows the class-conditional operating trade-off across thresholds. ROC and DET do not require a deployment prevalence, but finite N_1 and N_0 determine the uncertainty in the estimated curves. See §2.1 and §2.2.
How well does the detector rank a random target above a random nontarget overall?	AUC	A rank-based scalar summary. It can hide ROC-curve crossings and may give substantial weight to false-positive-rate regions that are operationally irrelevant. See §2.3.
Is a conventional single-number symmetric error summary desired?	EER	The error rate at an operating point where false-positive and false-negative rates are equal, when such a point is defined or suitably interpolated. EER is useful for benchmark reporting, but equal error rates are not automatically the Bayes-optimal operating rule under equal costs. See §2.3.
How well can the detector operate subject to a false-positive-rate requirement FPR $\leq \tau$?	TPR at the fixed-FPR operating requirement; estimated threshold $\widehat{\theta}_\tau$ and its uncertainty	The threshold is itself estimated from the nontarget sample, so uncertainty in both the threshold and the resulting TPR should be reported. If the population constraint must be protected, one may require an upper confidence bound for FPR(θ) to be no larger than τ. See §4.3 and §5.7.
Do scores intended as probabilities have the numerical meaning they claim?	Reliability diagram; Brier score	Reliability diagrams estimate the relationship between reported probability and observed target frequency. Results depend on sampling design and, for binned displays, on the binning choice. Probability calibration for deployment also depends on the relevant deployment prevalence. See §3.2–§3.4.
How do probability-like scores behave separately on targets and nontargets?	BrierT and BrierN	For probability-like scores, 0.25 is a useful reference value: a detector that reports 0.5 for every trial has BrierT = BrierN = 0.25. If the deployment prevalence π is known, a natural no-feature reference is instead the constant forecast $S = \pi$, which gives BrierT = $(1-\pi)^2$ and BrierN = π^2. See §3.3.
Do likelihood-ratio or log-likelihood-ratio scores provide useful evidentiary information?	C_{llr}; forensic empirical cross-entropy; calibrated/PAV lower envelope	These quantities evaluate LR/LLR evidence on a logarithmic scoring-rule scale. The gap between the observed score and an appropriately calibrated lower envelope can diagnose calibration loss; raw C_{llr} alone does not isolate calibration from discrimination. See §3.5.

Question	Primary quantity or display	Interpretation and important caution
Among the alerts produced at a chosen threshold, what fraction are actually targets?	PPV / precision; precision–recall display	PPV depends on deployment prevalence as well as TPR and FPR. Precision measured on an artificially balanced or case–control evaluation should not be transported to deployment without accounting for the deployment prevalence. See §4.5.
Which threshold is preferred when deployment prevalence and error costs are specified?	$C_{\mathrm{det}}(\theta)$ and the population minimum $C^*_{\mathrm{det}} = \inf_\theta C_{\mathrm{det}}(\theta)$	Report the expected-cost curve or relevant operating region, uncertainty for C^*_{det}, and the range of thresholds reasonably supported by the data. The smallest empirical cost on one evaluation corpus is a sample-optimized quantity and is generally optimistic for the population minimum. Conventional empirical summaries minDCF and actDCF are discussed in §4.2.
What did an adversarial attack do to the class it actually modified?	Class-conditional ASR; induced ASR; ΔFNR or ΔFPR; paired score displacement; class-conditional Brier component when appropriate	Construct one final attacked outcome for every source item under a stated attack budget and preserve clean–attacked pairing. Do not condition the corpus on attack success. See §6.1–§6.3.
What consequence did the attack have for the detector as a two-class system?	Clean-versus-attacked performance at a fixed operating threshold; secondarily, ROC/AUC relative to a specified unattacked reference population	ROC and AUC are not intrinsic one-class attack-efficacy measures: they also depend on the reference distribution of the class that was not attacked. See §6.4–§6.5.
How certain are any of the reported quantities?	Pointwise confidence intervals, simultaneous confidence bands, or a design-appropriate bootstrap	Uncertainty is part of the reported result. The resampling or interval construction should respect the effective independent sampling unit; clustered or repeated observations should not be treated as independent trials. See §5.

Two conventional terms that readers may encounter are *minDCF* and *actDCF*. On a finite evaluation corpus, empirical minDCF is the smallest empirical detection cost obtained by searching over thresholds on that corpus. Because the same data are used to choose and evaluate the threshold, it is generally optimistic as an estimate of the population minimum expected cost. Empirical actDCF is the observed cost at a threshold that was specified independently of the evaluation sample; it is useful for auditing an already-fixed operating rule. Neither quantity replaces an uncertainty analysis for the corresponding population performance.

7.2 Six Evaluation Red Flags

Before signing off on a binary-detector evaluation, check for the following problems.

☐ **Red Flag 1: Treating 0.5 as automatically justified.** A detector whose interface presents a probability-like score may naturally be used with a nominal threshold of 0.5, and performance at that threshold can be worth reporting. But 0.5 should not automatically be interpreted as a calibrated or decision-optimal threshold. Its operational justification depends on the meaning of

the score and, when an optimal decision rule is claimed, on the relevant costs and deployment conditions.

☐ **Red Flag 2: Choosing a detector by AUC alone.** A higher AUC does not imply better performance at every operating point. ROC curves can cross, so competing detectors should also be compared in the false-positive-rate region that matters for the application, with uncertainty reported there.

☐ **Red Flag 3: Transporting precision from the evaluation sample to deployment.** Precision is a posterior quantity and depends on prevalence. Precision measured on an artificial 50/50 test set, a case–control sample, or any other evaluation sample with a different class mixture should not be reported as deployment precision without accounting for the deployment prevalence.

☐ **Red Flag 4: Ignoring sampling uncertainty or dependence.** A point estimate is not enough when the relevant class-specific sample size is small. Likewise, repeated measurements from the same patient, speaker, device, document, or other source are not independent observations. Confidence intervals, confidence bands, and bootstrap procedures should respect the sampling design and the effective independent unit.

☐ **Red Flag 5: Confusing class imbalance with sample sparsity.** A highly unequal target-to-nontarget ratio does not by itself invalidate ROC or AUC. The precision of class-conditional estimates is governed primarily by the corresponding class-specific sample sizes. A study can be severely imbalanced yet have very precise ROC estimates, or nearly balanced yet have large uncertainty because both classes are small. Report PPV (precision) when you need to know how many positive decisions are expected to be correct; calculate it using the deployment prevalence.

☐ **Red Flag 6: Reporting only successful adversarial attacks.** For a corpus-level attack evaluation, every source item in the attacked class should contribute a final outcome under the stated attack procedure and attack budget. If an admissible attack fails to improve the attack objective, that source item remains part of the evaluation. Removing unsuccessful attacks conditions the analysis on success and overstates corpus-level attack efficacy.

Start with the evaluation question, not with a favorite metric. Report the quantity that answers that question, state the assumptions needed to interpret it, and report uncertainty appropriate to the sampling design. Discrimination, calibration, operational decision performance, and adversarial efficacy are related but distinct aspects of detector evaluation; no single scalar summary answers all of them.

A Formal Mathematical Definitions (Population Quantities)

This appendix collects the population quantities used throughout the guide. The corresponding empirical quantities are obtained from a finite labeled sample and are therefore estimates of the quantities defined here. Throughout, $Y \in \{0, 1\}$ denotes the true class, with $Y = 1$ the target class and $Y = 0$ the nontarget class. Let

$$S = s(X)$$

be the detector score for a randomly drawn trial, with larger values indicating stronger evidence for the target class. Unless a probability interpretation is stated explicitly, S may take values on any ordered numerical scale.

A.1 Class-Conditional Score Distributions and Threshold Rates

Define the class-conditional cumulative distribution functions

$$F_1(t) = P(S \le t \mid Y = 1), \qquad F_0(t) = P(S \le t \mid Y = 0). \tag{76}$$

Because the decision rule used in this guide declares a target when $S \ge \theta$, atoms or tied score values require the left limits

$$F_k(\theta^-) \triangleq P(S < \theta \mid Y = k), \qquad k \in \{0, 1\}.$$

The population operating rates are therefore

$$\mathrm{TPR}(\theta) = P(S \ge \theta \mid Y = 1) = 1 - F_1(\theta^-), \tag{77}$$
$$\mathrm{FPR}(\theta) = P(S \ge \theta \mid Y = 0) = 1 - F_0(\theta^-), \tag{78}$$
$$\mathrm{FNR}(\theta) = P(S < \theta \mid Y = 1) = F_1(\theta^-) = 1 - \mathrm{TPR}(\theta), \tag{79}$$
$$\mathrm{TNR}(\theta) = P(S < \theta \mid Y = 0) = F_0(\theta^-) = 1 - \mathrm{FPR}(\theta). \tag{80}$$

If the class-conditional score distributions are continuous at θ, then $F_k(\theta^-) = F_k(\theta)$ and the familiar expressions $\mathrm{TPR}(\theta) = 1 - F_1(\theta)$ and $\mathrm{FPR}(\theta) = 1 - F_0(\theta)$ follow.

The population ROC operating set is

$$\mathcal{R} = \left\{ \big(\mathrm{FPR}(\theta), \mathrm{TPR}(\theta)\big) : \theta \in \mathbb{R} \right\}, \tag{81}$$

with the limiting endpoints $(0, 0)$ and $(1, 1)$ obtained by thresholds above or below the score range. For a false-positive-rate budget $u \in [0, 1]$, it is also useful to define the best attainable true-positive rate

$$R(u) = \sup_{\theta:\, \mathrm{FPR}(\theta) \le u} \mathrm{TPR}(\theta). \tag{82}$$

This definition remains meaningful with discrete scores and ties. In a continuous-score idealization, with generalized quantile

$$F_0^{-1}(p) = \inf\{t : F_0(t) \ge p\},$$

one obtains the familiar representation

$$R(u) = 1 - F_1\big(F_0^{-1}(1 - u)\big), \tag{83}$$

subject to the usual regularity conditions that make the desired false-positive rate attainable exactly.

A.2 AUC and Equal Error Rate

Let S_1 and S_0 be independent scores drawn from the target and nontarget populations, respectively. The population area under the ROC curve is most cleanly defined by the pairwise ranking probability

$$\mathrm{AUC} = P(S_1 > S_0) + \frac{1}{2} P(S_1 = S_0). \tag{84}$$

The half-credit term makes the definition explicit when ties occur. Under the usual linear interpolation of tied ROC steps, Equation (84) equals the area under the corresponding ROC curve.

An *exact equal-error threshold* is any threshold θ_{EER} satisfying

$$\mathrm{FPR}(\theta_{\mathrm{EER}}) = \mathrm{FNR}(\theta_{\mathrm{EER}}). \tag{85}$$

When such a threshold exists, the common value is the population EER. An exact equality need not be attainable for discrete score distributions, and a unique threshold is not guaranteed. In that case an interpolated EER or another stated convention may be reported. In the continuous case, when the two error functions cross, the equal-error point also minimizes

$$\max\{\mathrm{FPR}(\theta), \mathrm{FNR}(\theta)\}. \tag{86}$$

Thus EER is a symmetric operating summary, not a generally optimal deployment threshold.

A.3 Precision, Recall, and Prevalence

Let

$$\pi = P(Y = 1)$$

denote the prevalence of targets in the population of interest. Recall is simply

$$\mathrm{Recall}(\theta) = \mathrm{TPR}(\theta). \tag{87}$$

Provided that the detector declares the target class with positive probability, precision or positive predictive value is

$$\mathrm{PPV}(\theta) = P(Y = 1 \mid S \geq \theta) \tag{88}$$

$$= \frac{\pi\,\mathrm{TPR}(\theta)}{\pi\,\mathrm{TPR}(\theta) + (1-\pi)\,\mathrm{FPR}(\theta)}. \tag{89}$$

If

$$P(S \geq \theta) = \pi\mathrm{TPR}(\theta) + (1-\pi)\mathrm{FPR}(\theta) = 0,$$

then $\mathrm{PPV}(\theta)$ is undefined. In particular, the plotted precision–recall boundary point $(0, 1)$ that some software adds when no positive decisions are made is a plotting convention, not a universal population limit.

A.4 Probability Scores, Calibration, and Brier Quantities

This subsection assumes that $S \in [0, 1]$ is intended to represent the posterior probability of the target class. Define

$$\eta(S) \triangleq P(Y = 1 \mid S). \tag{90}$$

The probability score is perfectly calibrated when

$$P(Y = 1 \mid S) = S \qquad \text{almost surely}, \tag{91}$$

or equivalently $E[Y \mid S] = S$ almost surely.

The population Brier score is

$$\mathrm{Brier} = E\big[(S - Y)^2\big]. \tag{92}$$

Its class-conditional components are

$$\mathrm{BrierT} = E\big[(S-1)^2 \mid Y = 1\big], \tag{93}$$

$$\mathrm{BrierN} = E\big[S^2 \mid Y = 0\big]. \tag{94}$$

They satisfy

$$\text{Brier} = \pi\,\text{BrierT} + (1-\pi)\,\text{BrierN}. \tag{95}$$

The class-conditional components are useful diagnostics, but considered separately they are not proper scoring rules.

For a constant probability forecast $S \equiv c$,

$$\text{BrierT} = (1-c)^2, \qquad \text{BrierN} = c^2. \tag{96}$$

Hence the symmetric neutral forecast $c = 0.5$ gives BrierT = BrierN = 0.25. If the deployment prevalence π is known and is used as the constant no-feature forecast, then

$$\text{BrierT} = (1-\pi)^2, \qquad \text{BrierN} = \pi^2. \tag{97}$$

The population Murphy decomposition is

$$E[(S-Y)^2] = \underbrace{E[(S-\eta(S))^2]}_{\text{Reliability}} - \underbrace{E[(\eta(S)-\pi)^2]}_{\text{Resolution}} + \underbrace{\pi(1-\pi)}_{\text{Uncertainty}}. \tag{98}$$

A.5 Likelihood Ratios, Log-Likelihood Ratios, and Logarithmic Loss

Suppose that the score S has class-conditional densities or probability mass functions $f_1(s)$ and $f_0(s)$ under the target and nontarget populations, respectively. The likelihood ratio is

$$\text{LR}(s) = \frac{f_1(s)}{f_0(s)}. \tag{99}$$

The likelihood ratio measures how much more strongly the observed score supports the target population than the nontarget population. We allow the extended values

$$0 \le \text{LR}(s) \le \infty.$$

Thus $\text{LR}(s) = 0$ when the score can occur under the nontarget population but not under the target population, and $\text{LR}(s) = \infty$ in the converse situation. Points at which both $f_1(s)$ and $f_0(s)$ are zero are outside the support of both populations and need not be assigned a likelihood ratio.

For $0 < \text{LR}(s) < \infty$, the natural log-likelihood ratio is

$$\ell(s) = \ln \text{LR}(s). \tag{100}$$

The extended conventions

$$\ell(s) = -\infty \quad \text{when} \quad \text{LR}(s) = 0, \qquad \ell(s) = +\infty \quad \text{when} \quad \text{LR}(s) = \infty$$

are useful when discussing logarithmic loss.

Let

$$O(\pi) = \frac{\pi}{1-\pi}, \qquad 0 < \pi < 1,$$

denote the prior odds for target prevalence π. Bayes' rule in odds form gives

$$\frac{P(Y=1 \mid S=s)}{P(Y=0 \mid S=s)} = \text{LR}(s)\,O(\pi). \tag{101}$$

Consequently,

$$P(Y = 1 \mid S = s) = \frac{\mathrm{LR}(s)\, O(\pi)}{1 + \mathrm{LR}(s)\, O(\pi)}, \tag{102}$$

and

$$P(Y = 0 \mid S = s) = \frac{1}{1 + \mathrm{LR}(s)\, O(\pi)}. \tag{103}$$

For a specified prior π, the corresponding population cross-entropy in bits is

$$\begin{aligned} \mathrm{CE}(\pi) = \pi\, E\left[\log_2\left(1 + \frac{1}{\mathrm{LR}(S)\, O(\pi)}\right) \middle| \, Y = 1\right] \\ + (1 - \pi)\, E[\log_2(1 + \mathrm{LR}(S)\, O(\pi)) | \, Y = 0]\,. \end{aligned} \tag{104}$$

The expressions in Equation (104) are interpreted by continuity at the endpoints. A target observation assigned $\mathrm{LR} = 0$ incurs infinite logarithmic loss, whereas a nontarget observation assigned $\mathrm{LR} = 0$ incurs zero loss. Conversely, a target observation assigned $\mathrm{LR} = \infty$ incurs zero loss, whereas a nontarget observation assigned $\mathrm{LR} = \infty$ incurs infinite loss. Accordingly, $\mathrm{CE}(\pi)$ may take the value $+\infty$.

If the likelihood ratio carries no information, so that $\mathrm{LR}(S) \equiv 1$, then

$$\mathrm{CE}_{\mathrm{null}}(\pi) = -\pi \log_2 \pi - (1 - \pi) \log_2(1 - \pi), \tag{105}$$

which is the binary entropy of the prior.

At the balanced prior $\pi = 1/2$, the prior odds are one and $\mathrm{CE}(1/2)$ reduces to the commonly used likelihood-ratio cost

$$C_{\mathrm{llr}} = \frac{1}{2}\left[E\Big\{\log_2(1 + e^{-\ell(S)}) \Big| \, Y = 1\Big\} + E\Big\{\log_2(1 + e^{\ell(S)}) \Big| \, Y = 0\Big\}\right]. \tag{106}$$

Equivalently,

$$C_{\mathrm{llr}} = \frac{1}{2}\left[E\left\{\log_2\left(1 + \frac{1}{\mathrm{LR}(S)}\right) \middle| \, Y = 1\right\} + E\{\log_2(1 + \mathrm{LR}(S)) | \, Y = 0\}\right]. \tag{107}$$

An uninformative system with $\mathrm{LR}(S) \equiv 1$, equivalently $\ell(S) \equiv 0$, has $C_{\mathrm{llr}} = 1$.

For an evaluation sample containing N_1 targets and N_0 nontargets, the forensic empirical cross-entropy is the sample analogue of $\mathrm{CE}(\pi)$:

$$\begin{aligned} \widehat{\mathrm{ECE}}(\pi) = \frac{\pi}{N_1} \sum_{i:Y_i=1} \log_2\left(1 + \frac{1}{\mathrm{LR}_i\, O(\pi)}\right) \\ + \frac{1 - \pi}{N_0} \sum_{j:Y_j=0} \log_2(1 + \mathrm{LR}_j\, O(\pi))\,. \end{aligned} \tag{108}$$

Here LR_i denotes the likelihood ratio reported for the ith evaluated trial. The same limiting conventions apply when an evaluated likelihood ratio is 0 or ∞.

A.6 Decision-Theoretic Population Quantities

Let $C_{Miss} \geq 0$ be the cost of a missed target and $C_{FA} \geq 0$ the cost of a false alarm. At deployment prevalence π, the population expected detection cost at threshold θ is

$$C_{det}(\theta) = C_{Miss}\,\pi\,\mathrm{FNR}(\theta) + C_{FA}\,(1-\pi)\,\mathrm{FPR}(\theta). \tag{109}$$

The population minimum expected cost is

$$C^*_{det} = \inf_{\theta} C_{det}(\theta). \tag{110}$$

The set of population cost-minimizing thresholds is

$$\Theta_0 = \arg\min_{\theta} C_{det}(\theta), \tag{111}$$

when the minimum is attained. More generally, for an operational tolerance $\varepsilon \geq 0$, the population near-optimal set is

$$\Theta_\varepsilon = \{\theta : C_{det}(\theta) \leq C^*_{det} + \varepsilon\}\,. \tag{112}$$

These are population quantities; an empirical minimum obtained by searching a finite evaluation set is not identical to C^*_{det}.

If the detector output is the calibrated deployment posterior probability $p(x) = P(Y = 1 \mid x)$, the Bayes decision rule declares the target class when

$$p(x) \geq \frac{C_{FA}}{C_{FA} + C_{Miss}}, \tag{113}$$

assuming $C_{FA} + C_{Miss} > 0$. If instead the score is a likelihood ratio, the Bayes rule declares the target class when

$$\mathrm{LR}(x) \geq \frac{C_{FA}(1-\pi)}{C_{Miss}\pi}, \tag{114}$$

for $0 < \pi < 1$ and positive error costs.

For an operational false-positive-rate requirement $\mathrm{FPR} \leq \tau$, define a population operating threshold by

$$\theta_\tau \in \arg\max_{\theta} \mathrm{TPR}(\theta) \qquad \text{subject to} \qquad \mathrm{FPR}(\theta) \leq \tau. \tag{115}$$

If the nontarget score distribution is continuous and the required tail probability is attainable, this becomes

$$\theta_\tau = F_0^{-1}(1-\tau). \tag{116}$$

With discrete scores or ties, exact equality $\mathrm{FPR}(\theta) = \tau$ may be impossible; Equation (115) is the more general definition.

A.7 Population Quantities for Adversarial Evaluation

Let S^{clean} and S^{adv} denote the clean and attacked scores for the same source item under a specified attack procedure and attack budget. If attack strength varies, the latter may be written $S^{\mathrm{adv}}_\epsilon$. The joint distribution of the paired scores is part of the population object of interest.

For an evasion attack on targets, the population post-attack success rate at threshold θ is

$$\mathrm{ASR}_{\mathrm{evasion}}(\theta) = P(S^{\mathrm{adv}} < \theta \mid Y = 1) = \mathrm{FNR}^{\mathrm{adv}}(\theta). \tag{117}$$

The net increase in miss rate is

$$\Delta_{\mathrm{FNR}}(\theta) = \mathrm{FNR}^{\mathrm{adv}}(\theta) - \mathrm{FNR}^{\mathrm{clean}}(\theta), \tag{118}$$

and the induced ASR among targets that were correctly detected before attack is

$$\mathrm{ASR}_{\mathrm{ind,evasion}}(\theta) = P\Big(S^{\mathrm{adv}} < \theta \mid S^{\mathrm{clean}} \geq \theta,\, Y = 1\Big), \tag{119}$$

provided that the conditioning event has positive probability.

For a spoofing attack on nontargets,

$$\mathrm{ASR}_{\mathrm{spoof}}(\theta) = P(S^{\mathrm{adv}} \geq \theta \mid Y = 0) = \mathrm{FPR}^{\mathrm{adv}}(\theta), \tag{120}$$

$$\Delta_{\mathrm{FPR}}(\theta) = \mathrm{FPR}^{\mathrm{adv}}(\theta) - \mathrm{FPR}^{\mathrm{clean}}(\theta), \tag{121}$$

$$\mathrm{ASR}_{\mathrm{ind,spoof}}(\theta) = P\Big(S^{\mathrm{adv}} \geq \theta \mid S^{\mathrm{clean}} < \theta,\, Y = 0\Big), \tag{122}$$

again provided that the conditioning event has positive probability.

If an evasion attack is evaluated at a population threshold θ_τ established from the clean nontarget distribution, the loss in target detection is

$$L_{\mathrm{TPR}}(\tau) = \mathrm{TPR}^{\mathrm{clean}}(\theta_\tau) - \mathrm{TPR}^{\mathrm{adv}}(\theta_\tau). \tag{123}$$

This is a detector-level consequence at a specified operating point, while the ASR quantities above are class-conditional attack measures.

Finally, for an evasion attack, the attacked AUC relative to a fixed nontarget reference population is

$$\mathrm{AUC}_{\mathrm{adv}} = P(S_1^{\mathrm{adv}} > S_0) + \frac{1}{2}P(S_1^{\mathrm{adv}} = S_0), \tag{124}$$

where S_0 is drawn from the specified nontarget reference population. This quantity depends on both the attacked target distribution and the chosen nontarget reference distribution; it is therefore not an intrinsic one-class measure of attack efficacy.

References

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning in adversarial settings. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2013)*, volume 8190 of *Lecture Notes in Computer Science*, pages 387–402. Springer, 2013. doi: 10.1007/978-3-642-40994-3_25.

Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–133, 2001. doi: 10.1214/ss/1009213286. URL https://doi.org/10.1214/ss/1009213286.

Niko Brümmer and Johan Du Preez. Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2-3):230–275, 2006.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. *Working Draft, University of Pennsylvania*, 2005.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. doi: 10.1109/SP.2017.49.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404. URL https://doi.org/10.1093/biomet/26.4.404.

Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

James P. Egan. *Signal Detection Theory and ROC-Analysis*. Academic Press, New York, 1975. ISBN 9780122328503.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Patrick J. Grother and Mei Lee Ngan. Face recognition vendor test (FRVT): Performance of face identification algorithms. NIST Interagency/Internal Report 8009, National Institute of Standards and Technology, 2014. URL https://doi.org/10.6028/NIST.IR.8009.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

Jan Hannig and Hari Iyer. Testing for calibration discrepancy of reported likelihood ratios in forensic science. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(1):267–301, 2022. doi: 10.1111/rssa.12747. URL https://doi.org/10.1111/rssa.12747.

Ananya Kumar, Percy S. Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Daniel J. Luckett, Eric B. Laber, Samer S. El-Kamary, Cheng Fan, Ravi Jhaveri, Charles M. Perou, Fatma M. Shebl, and Michael R. Kosorok. Receiver operating characteristic curves and confidence bands for support vector machines. *Biometrics*, 77(4):1422–1430, 2021.

Sofus A. Macskassy, Foster Provost, and Saharon Rosset. ROC confidence bands: An empirical evaluation. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pages 537–544, 2005.

Alvin Martin, George Doddington, Terri Kamm, Mark Ordowski, and Mark Przybocki. The DET curve in assessment of detection task performance. In *Proceedings of Eurospeech 1997*, pages 1895–1898, 1997. doi: 10.21437/Eurospeech.1997-504. URL https://doi.org/10.21437/Eurospeech.1997-504.

Pascal Massart. The tight constant in the dvoretzky–kiefer–wolfowitz inequality. *Annals of Probability*, 18(3):1269–1283, 1990.

Matthew B. A. McDermott, Lasse Hansen, Haoran Zhang, Giovanni Angelotti, and Jack Gallifant. A closer look at AUROC and AUPRC under class imbalance. *arXiv preprint arXiv:2401.06091*, 2024.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

NIST. NIST 2018 speaker recognition evaluation plan. Technical report, National Institute of Standards and Technology, 2018.

NIST. NIST 2024 speaker recognition evaluation plan. Technical report, National Institute of Standards and Technology, 2024.

Margaret S. Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, 2003.

Daniel Ramos, Joaquín González-Rodríguez, Grzegorz Zadora, and Colin G. G. Aitken. Information-theoretical assessment of the performance of likelihood ratio computation methods. *Science & Justice*, 53(4):442–452, 2013. doi: 10.1016/j.scijus.2013.07.004.

Eve Richardson, Raphael Trevizani, Jason A. Greenbaum, Hannah Carter, Morten Nielsen, and Bjoern Peters. The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns*, 5(6):100994, 2024.

Henry Swofford, Steven Lund, Hariharan Iyer, John Butler, Johannes Soons, Robert Thompson, Vincent Desiderio, John P. Jones, and Robert Ramotowski. Inconclusive decisions and error rates in forensic science. *Forensic Science International: Synergy*, 8:100472, 2024. doi: 10.1016/j.fsisyn.2024.100472. URL https://doi.org/10.1016/j.fsisyn.2024.100472.

Abraham Wald. Contributions to the theory of statistical estimation and testing hypotheses. *The Annals of Mathematical Statistics*, 10(4):299–326, 1939. doi: 10.1214/aoms/1177732144. URL https://doi.org/10.1214/aoms/1177732144.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Jin Chu Wu, Alvin F. Martin, Craig S. Greenberg, and Raghu N. Kacker. Uncertainties of measures in speaker recognition evaluation. In *Proceedings of SPIE*, volume 8040, page 804008, 2011. doi: 10.1117/12.883252. URL https://doi.org/10.1117/12.883252.

Grzegorz Zadora, Agnieszka Martyna, Daniel Ramos, and Colin Aitken. *Statistical Analysis in Forensic Science: Evidential Values of Multivariate Physicochemical Data*. Wiley-Blackwell Series in Forensic Science. John Wiley & Sons, Ltd, Chichester, UK, 2014. ISBN 978-0-470-75910-3. doi: 10.1002/9781118744031.