The ROC Diagonal is Not Layperson's Chance: A New Baseline Shows the Useful Area

André M. Carrington[1(✉)], Paul W. Fieguth[2], Franz Mayr[3], Nick D. James[4], Andreas Holzinger[5], John W. Pickering[6], and Richard I. Aviv[1]

[1] Department of Radiology, Radiation Oncology and Medical Physics, Faculty of Medicine, University of Ottawa and the Ottawa Hospital, Ottawa, Canada
{acarrington,raviv}@toh.ca

[2] Department of Systems Design Engineering, University of Waterloo, Waterloo, Canada
pfieguth@uwaterloo.ca

[3] Faculty of Engineering, Universidad ORT Uruguay, Montevideo, Uruguay
mayr@ort.edu.uy

[4] Software Solutions, Systems Integration and Architecture, The Ottawa Hospital, Ottawa, Canada
njames@toh.ca

[5] University of Natural Resources and Life Sciences Vienna, Vienna, Austria
andreas.holzinger@human-centered.ai

[6] Christchurch Heart Institute, Department of Medicine, University of Otago, Christchurch, New Zealand
john.pickering@otago.ac.nz

Abstract. In many areas of our daily lives (e.g., healthcare), the performance of a binary diagnostic test or classification model is often represented as a curve in a Receiver Operating Characteristic (ROC) plot and a quantity known as the area under the ROC curve (AUC or AUROC). In ROC plots, the main diagonal is often referred to as "chance" or the "random line". In general, however, this does not correspond to the layperson's concept of chance or randomness for binary outcomes. Rather, this represents a special case of layperson's chance, or the ROC curve for a classifier that has the same distribution of scores for the positive class and negative class. Where the ROC curve of a model deviates from the main diagonal, there is information. However, not all information is "useful information" compared to chance, including some areas and points above the diagonal. We define the binary chance baseline to identify areas and points in a ROC plot that are more useful than chance. In this paper, we explain this novel contribution about the state-of-art and provide examples that classify benchmark data.

Keywords: ROC · AUC · C-statistic · Chance · Explainable AI · Classification · Diagnostic tests

Published by Springer Nature Switzerland AG 2022
A. Holzinger et al. (Eds.): CD-MAKE 2022, LNCS 13480, pp. 100–113, 2022.
https://doi.org/10.1007/978-3-031-14463-9_7

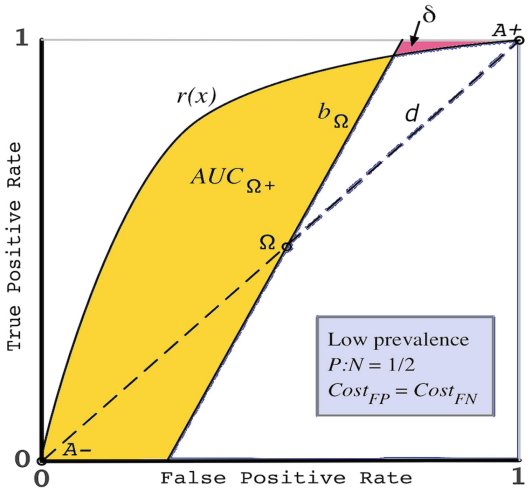

Fig. 1. Layperson's chance or 'binary chance', is a coin toss represented by the point $\Omega = (0.5, 0.5)$. Performance (cost-weighted accuracy) equal to Ω is depicted as the solid straight line b_Ω: the binary chance baseline. The ROC curve $r(x)$ has a useful area under the curve $AUC_{\Omega+}$ above and to the left of b_Ω—not the main diagonal d. There is negative utility in δ.

1 Introduction

In many fields which affect human life (e.g. health), the receiver operating characteristic (ROC) plot [3,19,23] depicts the performance of a binary diagnostic test or classification model as a ROC curve (Fig. 1, $r(x)$) which may be smooth (fitted) or staircase-like (empirical). Each point on the curve has a threshold value, often not displayed, where that threshold (e.g., t = 0.7) decides if a classifier's output (e.g., 0.6) is a positive outcome, if greater, or a negative outcome, if not.

In the ROC plot, a line drawn from the bottom left to the top right is called the main diagonal or chance diagonal (Fig. 1, dashed line) [12,27] because it is commonly said to represent chance [20,27], while others describe it as representing a random classifier [9].

The main diagonal is commonly used to interpret results in two ways. First, for the classifiers where a model's ROC curve is higher than the main diagonal, the model is said to be informative. Second, a model is thought to be better than chance where it is higher than the main diagonal—but we show that is **not** true for the most intuitive concept of chance for binary outcomes: a fair coin toss.

It is useful to compare any binary diagnostic test or classifier, to how a coin toss would decide the outcome for an instance or input, because if it performs worse, then we might as well use a coin toss. That is, we would compare a classifier against a black box classifier that had chance as a coin toss, as its internal mechanism.

A classifier starts to become useful when it performs better than a coin toss. However, let's posit the status quo, that a classifier starts to become useful at or just above the main diagonal.

The point (1, 1) on the main diagonal at the top-right of a ROC plot (Fig. 1, $A+$), is an all-positive classifier. It has the lowest possible threshold and predicts that all instances are positive. Predictions from $A+$ are mostly wrong for low prevalence data, yet $A+$ is also on the main diagonal which is also said to represent chance. Clearly, a coin toss performs better, being half right and half wrong, instead of mostly wrong. Hence the main diagonal does not represent chance as a coin toss, nor the point at which a classifier becomes useful.

Conversely, consider the point (0, 0) on the main diagonal at the bottom-left of a ROC plot (Fig. 1, $A-$). A classifier with a threshold at $A-$ classifies all inputs as negative. For data with low prevalence, i.e., few instances in the positive class, its predictions are mostly correct—more than the 50% correct predictions one obtains with a fair coin toss. Yet, $A-$ is on the main diagonal which is said to represent chance. Clearly, $A-$ performs better than chance as a coin toss—the layperson's concept of chance.

The main diagonal is commonly treated as a performance baseline, as in the examples above, but when it is used that way, we explain that what actually captures the user's intention and expectation is a line with performance equal to a fair coin toss.

This paper has four contributions, two that are clarifying and two that are novel. Our first contribution clarifies that the literature referring to the main diagonal as chance, is misleading, because it is not layperson's chance, nor any definition of chance found in a dictionary. Secondly, to evaluate and explain performance relative to layperson's chance (binary chance), we define a novel baseline. Iso-performance lines have not previously been applied for this purpose. Thirdly, we clarify that for realistic performance evaluation and explanation, one must express the prevalence and costs of error, which are either assumed implicitly or specified outright. These are relative costs incurred by the test subject or patient, not the health system—and these can be estimated without difficulty. Fourthly, we show the useful part of the area under the curve, as a novel contribution, demarcated by our novel baseline.

In the sections that follow we discuss: binary chance, the binary chance baseline, the main diagonal, the useful area under the ROC curve, counterbalancing effects on the slope of the binary baseline, examples in classification, related work and conclusions.

2 Binary Chance

We refer to "chance" in this manuscript as something which happens by chance, i.e., "luck" or "without any known cause or reason" [4]. Randomness is a synonym, and for binary outcomes, a fair coin toss produces a random outcome.

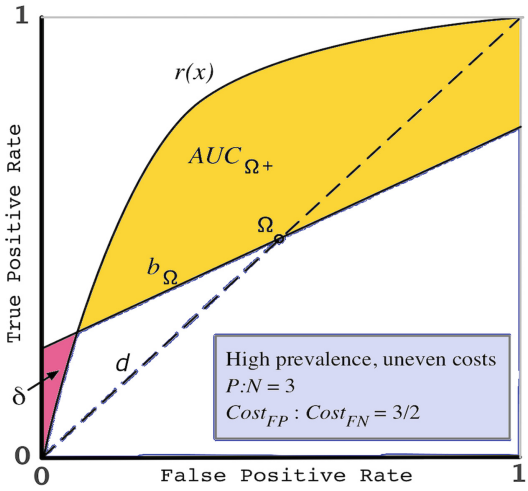

Fig. 2. In this example the binary chance baseline b_Ω has a low slope (lower than 1). This occurs for data with high prevalence and balanced costs, or when the class imbalance is greater than the cost imbalance, i.e., $3 > \frac{3}{2}$. For the ROC curve r(x), the useful area under the curve is $AUC_{\Omega+}$ (yellow) and negative utility is found in δ (red). (Color figure online)

We introduce the term "binary chance" for a fair coin toss. This is different from continuous chance as in the main diagonal (Sect. 4) and it is also distinguished from "a chance" of rain which implies a non-zero probability without fairness.

In this paper, we consider a coin toss, all-positive decisions and all-negative decisions to be classifiers with a hidden mechanism. The mechanism may consider the inputs or ignore them altogether when producing a corresponding output. Such classifiers are not good or useful by themselves—but they act as a baseline against which other classifiers are compared to determine if they are useful or not.

A fair coin toss (binary chance) is represented by the centre point $\Omega = (0.5, 0.5)$ [25] in a ROC plot (Fig. 2) because on average, the rate of heads (events) is 0.5 and the rate of tails (non-events) is 0.5.

To make comparison easy and visible between chance and points and areas in a ROC plot, we can draw a line through Ω that has performance equivalent to binary chance as established in the literature [16].

3 The Binary Chance Baseline

To find points with performance equivalent to binary chance, we draw an iso-performance line [9,10,16], denoted b_Ω through Ω (Fig. 2) and call it the binary chance baseline. It can represent cost-weighted accuracy, accuracy, or balanced accuracy—equivalent to chance.

We define the binary chance baseline $y = b_\Omega(x)$ or $x = b_\Omega^{-1}(y)$ for data $\mathcal{X}$ as a line with performance equal to chance $\Omega = (0.5, 0.5)$ with slope $m_\mathcal{X}$, except

where the line is bottom-coded (no less than 0) and top-coded (no greater than 1) to stay within the ROC plot by $[\cdot]_{01}$, as follows:

$$[\cdot]_{01} := \min(\max(\cdot, 0), 1) \tag{1}$$

$$b_{\Omega}(x) := [m_{\mathcal{X}}\,(x - \Omega_x) + \Omega_y]_{01} \tag{2}$$

$$b_{\Omega}^{-1}(y) := \left[\frac{(y - \Omega_y)}{m_{\mathcal{X}}} + \Omega_x\right]_{01} \tag{3}$$

where the literature defines the slope $m_{\mathcal{X}}$ (or skew) of iso-performance lines [10, 16,21] as a function of: P positives and N negatives (or prevalence π); and costs $C_{(.)}$ of false positives (FP), false negatives (FN), treatment for true positives (TP) and non-treatment for true negatives (TN) as follows.

$$m_{\mathcal{X}} := \frac{N}{P} \cdot \frac{C_{FP} - C_{TN}}{C_{FN} - C_{TP}} \tag{4}$$

$$= \frac{(1 - \pi)}{\pi} \cdot \frac{C_{FP} - C_{TN}}{C_{FN} - C_{TP}} \tag{5}$$

Notably, we are referring to the **relative cost** of a false negative versus a false positive, **incurred by a patient or subject**, such as reduced quality of life, risk of death, lost opportunity, or anxiety. We are **not** referring to system costs such as the test apparatus, labor, materials or administration.

We combine Eqs. (2) and (5), then (3) and (5) to obtain the explicit formula for the binary chance baseline, from a vertical and horizontal perspective, respectively:

$$b_{\Omega}(x) := \left[\frac{(1 - \pi)}{\pi} \cdot \frac{C_{FP} - C_{TN}}{C_{FN} - C_{TP}}(x - 0.5) + 0.5\right]_{01} \tag{6}$$

$$b_{\Omega}^{-1}(y) := \left[\frac{(y - 0.5)}{\frac{(1-\pi)}{\pi} \cdot \frac{C_{FP} - C_{TN}}{C_{FN} - C_{TP}}} + 0.5\right]_{01} \tag{7}$$

If we attempt to ignore costs, then we are *de facto* assuming equal costs which is worse than an estimate of costs. The estimate does not need to be perfect, just a better assumption. We usually know, for any given problem, if the cost of a false negative is worse than a false positive, or vice-versa, or about the same.

If we specify the prevalence and all of the cost parameters in the equation above (6) then the binary chance baseline represents **cost-weighted accuracy** or **average net benefit** equivalent to chance. This is the most realistic way to interpret performance relative to chance, and the baseline only aligns with the main diagonal in special cases when the class ratio and costs are balanced, together, or separately.

If we specify the prevalence but ignore costs, then the baseline represents **accuracy** equivalent to chance. When the prevalence is 50%, which is rarely the case, the baseline coincides with the main diagonal.

If we ignore prevalence and costs then the baseline represents **balanced accuracy** and coincides with the main diagonal. This approach and measure,

while unrealistic, can be useful to view performance without the majority class dominating. To compare performance between different data sets (with different class imbalances) as an abstract sanity check or benchmark of sorts. For example, 70% AUC or 20% area under the curve and above the main diagonal, is generally considered to be a reasonable classifier, in an abstract sense. However, realistically, for some applications such as detecting melanoma or fraud, such performance may be inadequate.

Hence, there is a choice between two different goals: unrealistic theoretic performance that can be compared between data sets (e.g., AUC and the main diagonal), versus realistic performance measures for decision-making the include prevalence and costs (cost-weighted accuracy and average net benefit).

Historically, use of the ROC plot, AUC and the main diagonal have ignored prevalence and costs. However, artificial intelligence (AI) is being increasingly applied to situations that affect people in everyday life, which requires real-world explanations.

The binary chance baseline is meaningful for explanation because a model starts to become useful only when it performs better than chance—otherwise, a coin toss performs just as well. The main diagonal can be misleading if we use it for the wrong reasons, which leads into the next section: what the main diagonal does represent, if not layperson's chance.

4 The Main Diagonal

The literature on ROC plots sometimes refers to the ROC main diagonal as chance [20,27]—a ROC curve produced by classification scores (or probabilities) drawn from *the same distribution* for events and non-events. The distribution does not need to be specified, as long as it is (almost everywhere[1]) the same. The classification scores (that underlie and precede the binary outcome) have an equal chance of being an event or non-event, hence we refer to this as "continuous chance" as distinguished from binary chance (layperson's chance).

When the distribution of scores is the same for events and non-events, the model is not informative, i.e., it has zero information to distinguish the two classes, resulting in a ROC curve along the main diagonal [13, Fig. 2]. The absence of information is demonstrated by divergence and distance measures which are zero in this case: the Jensen-Shannon (J-S) divergence [14,15,17] is zero, the Kullback-Leibler divergence [7] in either direction is zero, and the Hellinger distance [2] is zero.

The Hellinger distance fits the classification context nicely since it achieves a value of 1 when the two distributions are non-overlapping, or when two unimodel distributions are linearly separable—a situation where many classifiers

[1] Two probability density functions (PDF) are the same "almost everywhere" if they disagree on, at most, a set of isolated points (more formally, on a *set of measure zero*). This qualification is, admittedly, somewhat pedantic but necessary because any two such PDFs are effectively the same (and share the same cumulative distribution function). Changing a PDF at only individual points has no actual effect on the corresponding random variable it describes.

can perform perfectly with an AUC of 1 [13, Fig. 2]. That is, the [0, 1] range of the Hellinger distance correlates to the [0.5, 1] range of AUC achievable by most classifiers.

Points (and curves) on the main diagonal are said to be non-informative and of no predictive value [9,18,19]—however, that is typically not quite accurate. A model may not have any information along the diagonal, but if a threshold is chosen wisely based on prevalence as prior knowledge, then the resulting classifier may be useful at that threshold—informed by the threshold choice, not the model's own intelligence.

In all situations except those in which costs and prevalence exhibit a rare and unusual equilibrium, some points on the main diagonal if chosen with prior knowledge of prevalence, are more predictive than (binary) chance, as we have explained previously in the introduction for the points $A+$ and $A-$.

5 The Useful Area Under the ROC Curve

To explain performance in a ROC plot, we must consider both the binary chance baseline and the main diagonal. The area under the ROC curve and above the binary chance baseline (the yellow area denoted $AUC_{\Omega+}$ in Fig. 2), tells us where a model is useful, i.e., more useful than chance.

We define the useful area $AUC_{\Omega+}$ for an ROC curve $r(x)$ relative to binary chance Ω as follows, with a vertical and horizontal component, like the concordant partial AUC [3]. We use the notation $[\ \cdot\ ]_+$ for a function or filter that only passes positive values (bottom codes to zero). For a partial (or whole) ROC curve in the range $\theta_{xy} = \{x \in [x_1, x_2],\ y \in [y_1, y_2]\}$:

$$[\,\cdot\,]_+ := \min(\,\cdot\,,\ 0) \tag{8}$$

$$\begin{aligned} AUC_{\Omega+}(\theta_{xy}) := {} & \frac{1}{2}\int_{x_1}^{x_2} [\,r(x) - b_\Omega(x)\,]_+ \ dx \\ & + \frac{1}{2}\int_{y_1}^{y_2} \left[\left(1 - r^{-1}(y)\right) - \left(1 - b_\Omega^{-1}(x)\right)\right]_+ \ dy \end{aligned} \tag{9}$$

In the special case of a whole ROC curve, $\theta_{01} = \{x, y \in [0, 1]\}$, the expression simplifies, because the horizontal and vertical areas are the same:

$$AUC_{\Omega+}(\theta_{01}) := \int_0^1 [\,r(x) - b_\Omega(x)\,]_+ \ dx \tag{10}$$

Underneath the main diagonal and the binary chance baseline, the areas are both 0.5, but the location of those areas differ. Similarly, the areas between the curve and each baseline over the whole plot, are the same (Fig. 3), despite appearing to differ by δ (Figs. 1 and 2).

The difference is apparent in a range or region of interest (Fig. 4). Also, there are some points on a ROC curve that perform worse than chance—those which border δ (Fig. 2).

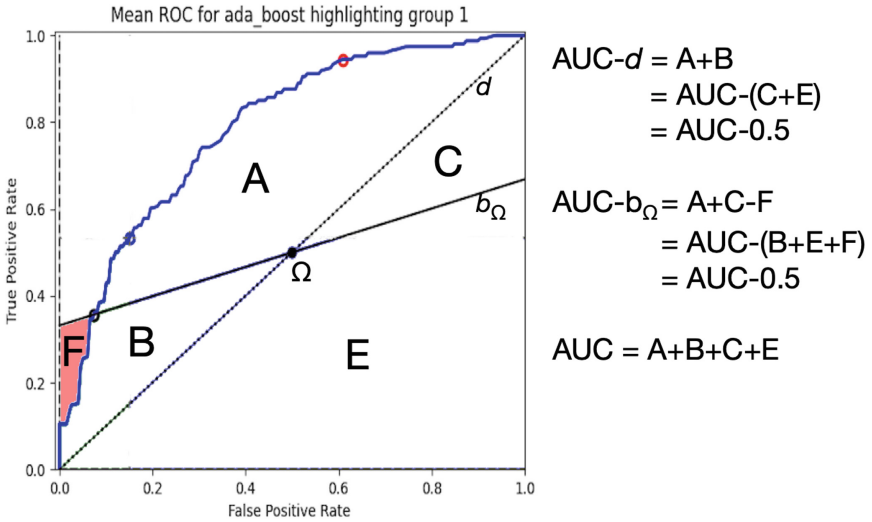

Fig. 3. We illustrate that AUC_d, i.e., AUC minus the diagonal and AUC_Ω, i.e., AUC minus the binary chance baseline, are the same, using a ROC curve from Adaboost used to classify breast cancer data. While there is no difference over a whole ROC curve, there are differences for part of a ROC curve (in subsequent figures).

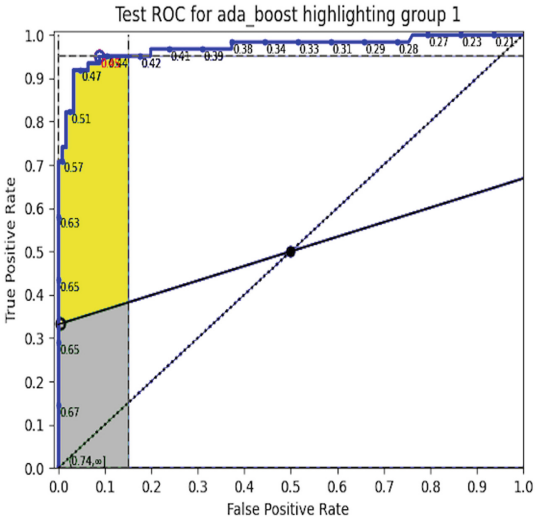

Fig. 4. A ROC plot for Adaboost applied to Wisconsin Breast Cancer recurrence data (size and texture). The vertical (sensitivity) aspect is highlighted for the region of interest $FPR = [0, 0.15]$, with yellow area better than binary chance. (Color figure online)

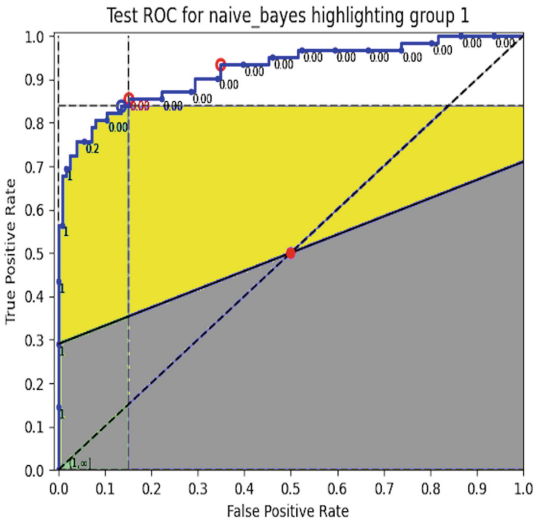

Fig. 5. A ROC plot for Naive Bayes applied to Wisconsin Breast Cancer recurrence data (size and texture). The horizontal (specificity) aspect is highlighted for the ROI $FPR = [0, 0.15]$.

The status quo approach to ROC plots, historically but unnecessarily, ignores prevalence and costs, limiting the use of ROC plots to abstract interpretations for initial model development, instead of interpreting a model in the context of real-life applications [11]. Halligan *et al.* [11] imply that this is an inherent limitation of ROC plots, but our work demonstrates that sometimes it is not. Others have also suggested additions to status quo ROC plots for better explanations [1].

The useful areas and points on ROC curves identified by the binary chance baseline in ROC plots, provide explanations that are complimentary to decision curve analysis. Furthermore, ROC plots can relate a variety of pre-test and post-test measures to each other, including predictive values and likelihood ratios.

6 Prevalence and Costs May Counteract Each Other

The previous section explains that ignoring prevalence and costs may cause errors in interpretation and choosing the best classifier. However, we may also incur errors by including prevalence while ignoring costs, or including costs while ignoring prevalence, because they often have counteracting effects in (2) as we explain in the following example.

Consider a medical condition, such as colon cancer. The cost C_{FN} of missing the disease in screening, a false negative, is much worse than a mistaken detection C_{FP}, a false positive, for which follow-up tests are conducted with some expense. The term $\frac{C_{FP}-C_{TN}}{C_{FN}-C_{TP}}$ causes a low slope, i.e., less than one.

However, as a condition with low prevalence, the term $\frac{N}{P}$ causes a high slope, i.e., it has an effect that counteracts the costs. These effects do not in general

balance each other, and in some cases they act in the same direction. However, more often than not, the minority class tends to be the class of interest, and false negatives tend to be more costly, resulting in counteracting effects.

7 Examples in Classification

We conduct an experiment using the Wisconsin Diagnostic Breast Cancer (WDBC) data set [26] from the UCI data repository [8]. The Wisconsin data set was curated by clinicians for the prediction of whether or not breast cancer recurs by a time endpoint.

Our testing examines size, texture and shape features of nuclei sampled by thin needle aspiration. We omit shape, or size and shape in some experiments to show a range of different ROC curves and results for analysis with our new baseline.

For diagnostic tests (as opposed to screening) the SpPin rule [22] recommends a focus on high specificity: the left side of a ROC plot. Hence, we define the region of interest (ROI) as 85% specificity or above: FPR = $[0, 0.15]$ (Fig. 4). AUC in the ROI is computed from the combination of a vertical perspective (Fig. 4) and a horizontal perspective (Fig. 5) for the same ROC curve.

We interpret points (Table 1) in the ROC plot (Fig. 4) for the Adaboost algorithm. As discussed previously, the endpoints of the main diagonal (0, 0) and (1, 1) do **not** perform the same as chance (Table 1).

We compute **average net benefit** according to Metz [16, pp. 295], as the difference in the cost of using a test and not using it:

$$\overline{NB} = -(\overline{C}_{\text{use}} - \overline{C}_{\text{not\_use}}) \tag{11}$$

where cost of using the test is [16, pp. 295]:

$$\begin{aligned} \overline{C}_{\text{use}} = &-(C_{FN} - C_{TP}) \cdot \pi \cdot TPR \ + (C_{FP} - C_{TN}) \cdot (1 - \pi) \cdot FPR \\ &+ C_{FN} \cdot \pi + C_{TN} \cdot (1 - \pi) + C_o \end{aligned} \tag{12}$$

and the fixed cost, $\overline{C}_{\text{not\_use}}$ is set to zero for a diagnostic test. The overhead cost C_o is set to zero, since we are not interested in return on investment.

Cost-weighted accuracy is average net benefit normalized to the range [0, 1] or [0%, 100%]. Note that all points along the binary chance baseline, such as (0.5, 0.5) and the baseline's intersection with the ROC at (0, 0.33), have the same average net benefit and cost-weighted accuracy as chance.

In Fig. 4 about one third of the vertical area in the ROI performs no better than binary chance (Table 2). Thresholds in the gray region below the binary chance baseline perform worse than chance and should not be used, contrary to analysis with the main diagonal. The binary chance baseline has a gradual or low slope in this example (Fig. 4). Low prevalence (30%) with no other factors would cause a high slope, however the cost of false negatives are specified as five times worse than a false positive (a hypothetical cost), causing a low slope.

Table 1. Validation of expected cost-weighted accuracy at various points in the ROC plot for Wisconsin Diagnostic Breast Cancer classified with Adaboost (Fig. 4).

Description	ROC point	Accuracy	Average net benefit	Cost-weighted accuracy	Expectation for Avg NB, CW-Acc
A perfect classifier at a perfect threshold t	(0, 1)	100%	0	100%	Best possible value
The worst classifier at the worst threshold t	(1, 0)	0%	−2.49	0%	Worst possible value
Binary chance	(0.5, 0.5)	50%	−1.25	50%	Value for chance
All negative classifier A−, t = ∞	(0, 0)	67%	−1.86	25%	Worse than chance
All positive classifier A+, t = −∞	(1, 1)	33%	−0.63	75%	Better than chance
An optimal point on ROC curve, t = 0.45	(0.09, 0.95)	93%	−0.14	94%	Best value on ROC
An optimal point on ROC curve, in ROI_1, t = 0.45	(0.09, 0.95)	93%	−0.14	94%	Less than or equal to the best ROC value
Intersection of ROC curve and binary chance baseline	(0, 0.33)	75%	−1.25	50%	Equal to value for chance

Table 2. Results with useful areas highlighted in green cells and useful sensitivity and specificity highlighted in orange cells. *the same value as AUC_Ω

Description	AUC in a part	AUC in whole or normalized	Average Sens	Average Spec	Expectation
ROC	-	AUC =97.2%	97.2%	97.2%	Visually, AUC is nearly 100%
ROC above main diagonal	AUC_d =47.2%	-	47.2% above	47.2% above	Over the whole curve, $AUC_d = AUC - 0.5$
ROC above binary chance	AUC_Ω =47.2%	-	47.2% above	47.2% above	Over the whole curve, $AUC_\Omega = AUC - 0.5$, different from the 0.5 above $\lvert AUC_\Omega \rvert = \lvert AUC_d \rvert$ (Figure 3)
ROC in ROI_1 $[0, 0.15]$	AUC_1 =54.0%	$AUCn_1$ =98.1%	90.5%	99.3%	Visually, the ROI is not much better or worse than the whole curve (98.1% ≈ 97.2%)
ROC in ROI_1 above the main diagonal	AUC_{d1} =28.5%	-	83.0% above	46.8% above	
ROC in ROI_1 above the binary chance baseline	$AUC_{\Omega 1}$ =26.3%	-	54.8% above	46.7% above	In ROI_1 the useful area is smaller than what the diagonal identifies: 26.3% < 28.5%. Avg Sens above is smaller too.

8 Related Work

Zhou *et al.* [27] provide a classic work on the interpretation of ROC curves and plots including discussion of the main diagonal and chance. Numerous other sources [12,20,27] discuss and describe the main diagonal as chance or a random classifier [9], and the main diagonal has long been used as a point of comparison for performance measures in ROC plots.

Aside from binary chance and continuous chance as concepts we discuss in this paper, or "a chance" of rain, there is also the concept of chance agreement. Kappa [5] describes the amount of agreement beyond chance agreement, between any two models or people that produce scores. Kappa includes prevalence but does not include costs unless modified [6]. The unmodified version is more commonly known.

While Kappa and other priors may provide alternative baselines from which to judge utility, our paper focuses on the misunderstanding of the main diagonal as chance, and the clarification of the binary chance baseline as the layperson's concept of chance for binary outcomes. Other baselines may be investigated in other work.

Iso performance lines were introduced by Metz [16], and later used by others, such as Provost and Fawcett [9] for the purpose of identifying an optimal ROC point on the ROC curve. Flach [10] then investigated the geometry of ROC plots with iso performance lines. Subtil and Rabilloud [24] were the first, to apply iso performance lines as a baseline from which to measure performance. They examine performance equivalent to the all-negative and all-positive classifiers we discuss and denote as $A-$ and $A+$ respectively.

9 Conclusions and Future Work

ROC plots that label the main diagonal as chance are misleading for interpretation—whereas a label such as "no skill" is accurate. We newly illustrate the baseline that represents performance equal to layperson's chance, which is more intuitive and explainable. We showed that explanations based on the main diagonal, about several ROC points, are faulty, whereas the binary chance baseline is congruent with our expectations and computed values of cost-weighted accuracy.

While ROC plots were originally applied to disregard prevalence and costs to compare performance between different data sets (e.g., different radar scenarios with different prevalence)—that does not serve the need for realistic performance evaluation and explanation. We explained that prevalence and costs are always present—they are either assumed and implicit or they are explicit and more correct if estimated. We posit that the new methods in our paper will provide more insight into performance and ROC plots in past and present results.

Availability of Code

The measures and plots in this paper were created with the bayesianROC toolkit in Python, which can be installed at a command line, as follows. It requires the deepROC toolkit as well.

```
pip install bayesianroc
pip install deeproc
```

The associated links are:
https://pypi.org/project/bayesianroc/
https://github.com/DR3AM-Hub/BayesianROC
https://pypi.org/project/deeproc/
https://github.com/Big-Life-Lab/deepROC

Acknowledgements. Parts of this work has received funding by the Austrian Science Fund (FWF), Project: P-32554 "A reference model for explainable Artificial Intelligence in the medical domain".

Contributions. All authors contributed in writing this article. AC conceived the main ideas initially. In consultation with PF, JP, FM, and NJ various ideas were further developed and refined, with AH and RA providing guidance. Experiments were conducted and coded by AC and FM. All authors reviewed and provided edits to the article.

References

1. Althouse, A.D.: Statistical graphics in action: making better sense of the ROC curve. Int. J. Cardiol. **100**(215), 9–10 (2016)
2. Beran, R.: Minimum Hellinger distance estimates for parametric models. Ann. Stat. **5**(3), 445–463 (1977)
3. Carrington, A.M., et al.: A new concordant partial AUC and partial C statistic for imbalanced data in the evaluation of machine learning algorithms. BMC Med. Inform. Decis. Making **20**(1), 1–12 (2020)
4. Chance Noun: In the Cambridge Dictionary. Cambridge University Press. https://dictionary.cambridge.org/dictionary/english/chance
5. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Measur. **20**(1), 37–46 (1960)
6. Cohen, J.: Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol. Bull. **70**(4), 213 (1968)
7. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, Hoboken (2012)
8. Dua, D., Graff, C.: UCI machine learning repository (2017). http://archive.ics.uci.edu/ml
9. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
10. Flach, P.A.: The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In: Proceedings of the Twentieth International Conference on Machine Learning (2003)

11. Halligan, S., Altman, D.G., Mallett, S.: Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach. Eur. Radiol. **25**(4), 932–939 (2015). https://doi.org/10.1007/s00330-014-3487-0
12. Hand, D.J.: Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach. Learn. **77**(1), 103–123 (2009). https://doi.org/10.1007/s10994-009-5119-5
13. Inácio, V., Rodríguez-Álvarez, M.X., Gayoso-Diz, P.: Statistical evaluation of medical tests. Ann. Rev. Stat. Appl. **8**, 41–67 (2021)
14. Lin, J.: Divergence measures based on the Shannon entropy. IEEE Trans. Inf. Theory **37**(1), 145–151 (1991). https://doi.org/10.1109/18.61115. http://ieeexplore.ieee.org/document/61115/
15. Menéndez, M., Pardo, J., Pardo, L., Pardo, M.: The Jensen-Shannon divergence. J. Franklin Inst. **334**(2), 307–318 (1997). Publisher: Elsevier
16. Metz, C.E.: Basic principles of ROC analysis. In: Seminars in Nuclear Medicine, vol. 8, pp. 283–298. Elsevier (1978)
17. Nielsen, F.: On a variational definition for the Jensen-Shannon symmetrization of distances based on the information radius. Entropy **23**(4), 464 (2021)
18. Obuchowski, N.A.: Receiver operating characteristic curves and their use in radiology. Radiology **229**(1), 3–8 (2003). https://doi.org/10.1148/radiol.2291010898
19. Obuchowski, N.A., Bullen, J.A.: Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. Phys. Med. Biol. **63**(7), 07TR01 (2018)
20. Powers, D.M.W.: Evaluation: from precision, recall and F-factor to ROC, informedness, markedness & correlation. Technical report, Flinders University, December 2007
21. Provost, F., Fawcett, T.: Robust classification for imprecise environments. Mach. Learn. **42**, 203–231 (2001). https://doi.org/10.1023/A:1007601015854
22. Sackett, D.L., Straus, S.: On some clinically useful measures of the accuracy of diagnostic tests. BMJ Evid.-Based Med. **3**(3), 68 (1998)
23. Streiner, D.L., Cairney, J.: What's under the ROC? An introduction to receiver operating characteristics curves. Can. J. Psychiatry **52**(2), 121–128 (2007)
24. Subtil, F., Rabilloud, M.: An enhancement of ROC curves made them clinically relevant for diagnostic-test comparison and optimal-threshold determination. J. Clin. Epidemiol. **68**(7), 752–759 (2015)
25. Van den Hout, W.B.: The area under an ROC curve with limited information. Med. Decis. Making **23**(2), 160–166 (2003). https://doi.org/10.1177/0272989X03251246
26. Wolberg, W.H., Mangasarian, O.L.: Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proc. Natl. Acad. Sci. **87**(23), 9193–9196 (1990)
27. Zhou, X.H., McClish, D.K., Obuchowski, N.A.: Statistical Methods in Diagnostic Medicine, vol. 569. Wiley, Hoboken (2002)