REVIEW PAPER

JMS Journal of Medical Science

The Toolbox for Rating Diagnostic Tests: A Guide to Classification Metrics

Wiktoria Zasada
Department of Computer Science and Statistics, Poznan University of Medical Sciences, Poznań, Poland
https://orcid.org/0000-0001-9329-3495

Corresponding author: 81576@student.ump.edu.pl

Przemysław Guzik
Department of Cardiology—Intensive Therapy, Poznan University of Medical Sciences, Poznań, Poland
University Centre for Sports and Medical Studies, Poznan University of Medical Sciences, Poznań, Poland
https://orcid.org/0000-0001-9052-5027

Katarzyna B. Kubiak
Department of Computer Science and Statistics, Poznan University of Medical Sciences, Poznań, Poland
https://orcid.org/0000-0002-1467-4853

Barbara Więckowska
Department of Computer Science and Statistics, Poznan University of Medical Sciences, Poznań, Poland
https://orcid.org/0000-0002-1811-2583

https://doi.org/10.20883/medical.e1474

Keywords: statistical analysis, binary classification, prediction model, diagnostic test, ROC curve, model evaluation

Received 2025-12-09
Accepted 2025-12-30
Published 2025-12-31

How to Cite: Zasada W, Guzik P, Kubiak KB, Więckowska B. The Toolbox for Rating Diagnostic Tests: A Guide to Classification Metrics. Journal of Medical Science. 2025 December;94(4);e1474. https://doi.org/10.20883/medical.e1474

ABSTRACT

Evaluating a classifier's performance is critical for its successful application. This paper explores various metrics used for binary classification tasks, highlighting their strengths and limitations.

Simple threshold metrics, such as Accuracy and Sensitivity, are efficient for binary data and a single cutoff point. However, their reliance on a single threshold and sensitivity to imbalanced data can be drawbacks.

For more robust evaluation, ranking metrics such as Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves provide a threshold-agnostic approach, enabling comparison across different cutoff points. Additionally, probabilistic metrics like Brier Score and Log Loss assess the model's ability to predict class probabilities.

The choice of metric depends on the specific classification problem and the characteristics of the data. When dealing with imbalanced data or complex decision-making processes, using multiple metrics is recommended to gain a comprehensive understanding of the model's performance.

This paper emphasises the importance of understanding metric limitations and of selecting appropriate metrics for a specific classification task. By doing so, researchers and practitioners can ensure a more accurate and informative evaluation of their models, ultimately leading to the development of reliable tools for various applications.

Introduction

In today's data-driven world, accurately distinguishing between healthy and sick individuals is crucial across various sectors. Diagnostic tests play a vital role in medicine, public health, and research by enabling objective evaluation of patients and the diagnosis of conditions [1].

With challenges like the COVID-19 pandemic, there's a growing awareness of the need for accurate diagnostic tests and continual improvement [2,3].

There is a vast array of diagnostic tests (classifiers), each with its own set of quality metrics. Choosing the appropriate metric depends on several factors.

First, the disease prevalence must be considered. Is the disease common or rare in the population being tested? Screening tests for diseases with low prevalence, like mammograms for breast cancer, use different metrics than tests for suspected cases. Secondly, the type of variable measured by the test is crucial. Can the test result be a number (e.g., body temperature), an ordered category (e.g., mild pain), or something else (e.g., gender or smoking status) [4]? Thirdly, the purpose of the test must be determined. Are we simply classifying someone as healthy or sick, or are we trying to predict future health outcomes, like dead or alive?

Classification vs. Prediction: Two Sides of the Same Coin

Classification and prediction are closely linked. While we often think of classifying things in the present and predicting future events, the technical difference is not always clear-cut. Although it is possible to study attachments in many categories, we will focus on just two. We can classify both present ("sick" vs. "healthy") and future ("will get sick" vs. "will stay healthy") states. We can also estimate the likelihood of an event occurring now (sick vs. healthy) or in the future (becoming sick vs. remaining healthy). Studies use a "training set" of data to determine how to classify or predict future situations and then test this method on new data sets.

A perfect classifier would flawlessly assign people to the sick or healthy groups. A useless classifier would not distinguish between groups and would always guess randomly.

This review explores various tools scientists use to assess test quality, like Sensitivity, Specificity, Matthews Correlation Coefficient (MCC), Area Under the Receiver Operating Characteristic – AUC (ROC) curve, Brier Score, and more. We will discuss their strengths, weaknesses, and limitations for classifying and predicting health outcomes. We will also show how to interpret these matrices and compute them using widely used software such as R and Python.

The Power of Metrics: Assessing Classification Quality in Diverse Fields

Accurately evaluating diagnostic tests and predictive models enables a wide range of applications in healthcare. These tools play a crucial role in forecasting disease outbreaks [5,6], patient admissions, and treatment outcomes across various fields like epidemiology, general healthcare, and therapeutic interventions [7,8]. They also contribute significantly to clinical trials by aiding in participant selection, identifying patients at higher risk of complications, and assessing individual risk for chronic diseases [9,10]. Beyond trials, they assist in clinical practice by supporting disease diagnosis through patient data and biomarkers [11,12], personalising treatment plans and predicting their effectiveness [13,14], evaluating genetic disease risk and susceptibility to adverse drug reactions [15,16], and guiding preventive interventions [17,18].

While numerous metrics exist to evaluate the predictive capabilities of variables, algorithms, and models, choosing the most appropriate set can be challenging. Mathematicians and statisticians continually develop new metrics to better characterise the nuances of various tasks and their outcomes [19-21]. This review introduces and explains the most commonly used metrics in the health sciences, with particular emphasis on those designed explicitly for recent machine learning techniques.

Metrics to assess the quality of classification

A classifier categorises objects based on their characteristics. For example, it can classify patients into specific classes, categories, or groups. Classifiers can be simple, relying on a single variable for categorisation. Conversely, they can be complex, derived from models that consider multiple variables.

Evaluating classifiers and prediction models often involves multiple metrics. With dozens of indicators available, selecting the appropriate one can be challenging. Ferri et al. categorise classification metrics into three groups: Threshold, Rank, and Probability Metrics (see **Figure 1**) [22].

Threshold metrics evaluate classification performance at a fixed threshold, such as the commonly used 6.5% cutoff for diagnosing diabetes using glycated haemoglobin A1c (HbA1c). If a patient's HbA1c is above this, they are classified as having diabetes.

Rank metrics assess how well a classifier ranks predictions. They consider the ordering of all scores, rather than a single cutoff, for example, in diagnosing diabetes using HbA1c, where sorted values of this variable can serve as consecutive cutoff points. Among the possible values, those above the cutoff point (6.5%) are generally considered indicative of diabetes [23]. Similarly, on the BIRADS (Breast Imaging Reporting and Data System) scale, sorted cut-off points range from 1 (negative) to 5 (high cancer probability) [24]. Conversely, probability metrics are used in models that calculate the exact likelihood of a disease or event, focusing on the specific value rather than a single cut-off or order.

In a nutshell, Threshold metrics are like on/off switches, ignoring prediction order. Rank metrics are like assigning grades (A, B, C), focusing on the order of predictions. Probability metrics are like percentages, considering both order and closeness to reality.

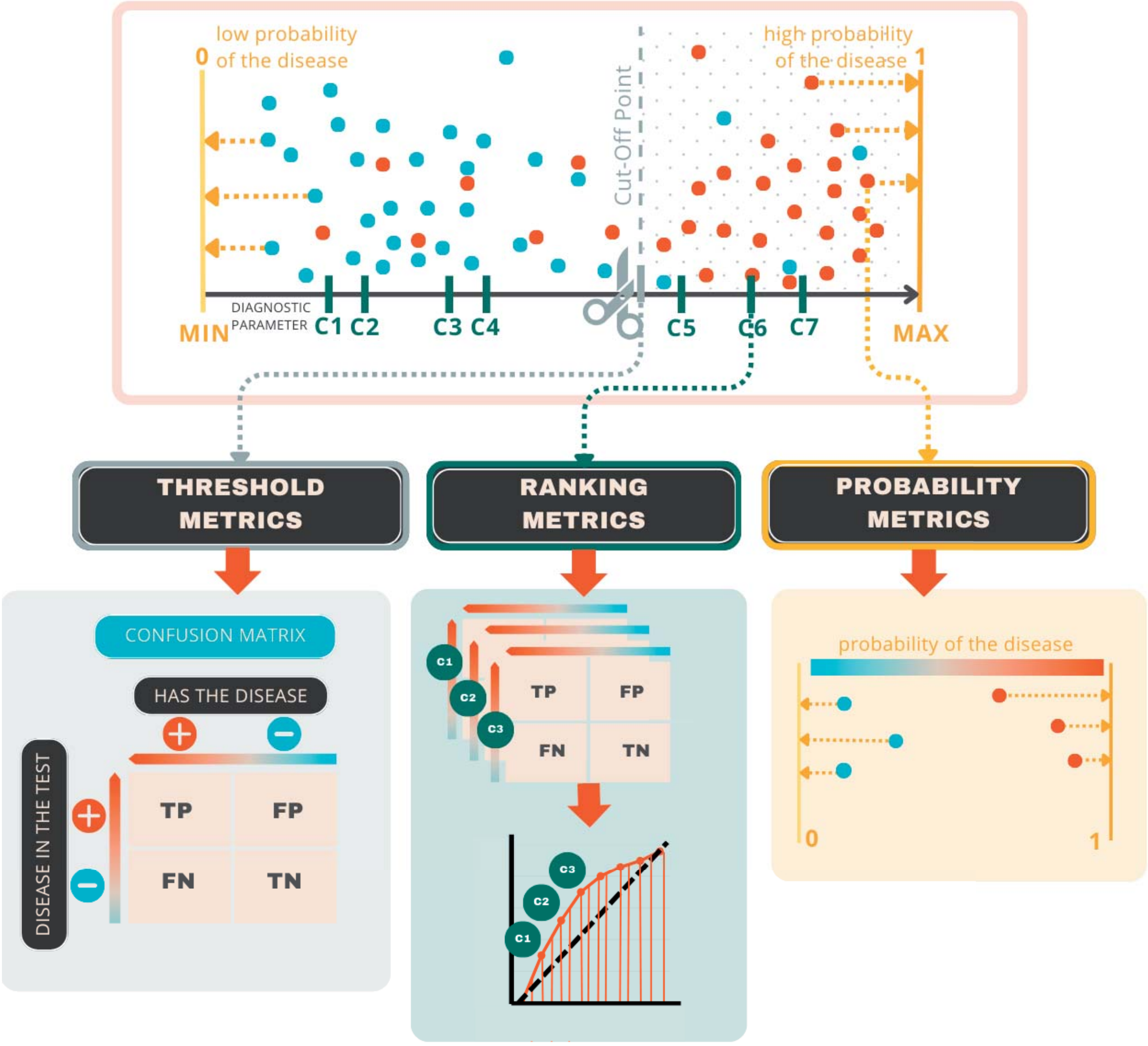


Figure 1. Illustration of the fundamental concepts underlying threshold, ranking, and probabilistic metrics in binary classification. The diagram uses colour coding, with blue indicating patients (positive class) and red indicating healthy individuals (negative class), to represent the core components of classification evaluation visually.

Confusion matrix

A confusion matrix is a table used to define the performance of a classification algorithm. The matrix has two dimensions: predicted classification and actual classification. The predicted classification is the classification the investigator assigns to each patient (or other test subject), whereas the actual classification is the correct classification of each patient. The confusion matrix is shown in **Figure 2**, using the BIRADS scale for breast cancer detection as an example.

Such confusion matrices are used to compute threshold metrics. For Rank Metrics, multiple confusion matrices are created by varying cutoff points; in the example, separate tables are produced for BIRADS = 1, BIRADS = 2, BIRADS = 3, BIRADS = 4, and BIRADS = 5. Probability Metrics do not use such a matrix; instead, they are derived directly from a disease-specific probability value calculated for each patient and from the patient's distance to the actual value.

Metrics for assessing the quality of classification

Table 1 summarises commonly used measures to assess the quality of a classifier, divided into Threshold Metrics, i.e. based on a single confusion matrix, Rank Metrics based on multiple confusion matrices determined for sequentially ordered cut-off points, and Probabilistic Metrics based on classifiers that determine the probability of an event occurring, i.e. a value between 0 and 1.

Threshold Metrics

A simple classifier is binary. For instance, whether a patient exhibits symptoms or has a disease. There are also more complex binary classifiers, such as Naive Bayes, decision trees, and neural networks. Many of these models output continuous scores or class probabilities rather than direct class labels, and a confusion matrix can be obtained only after selecting a decision thresh-

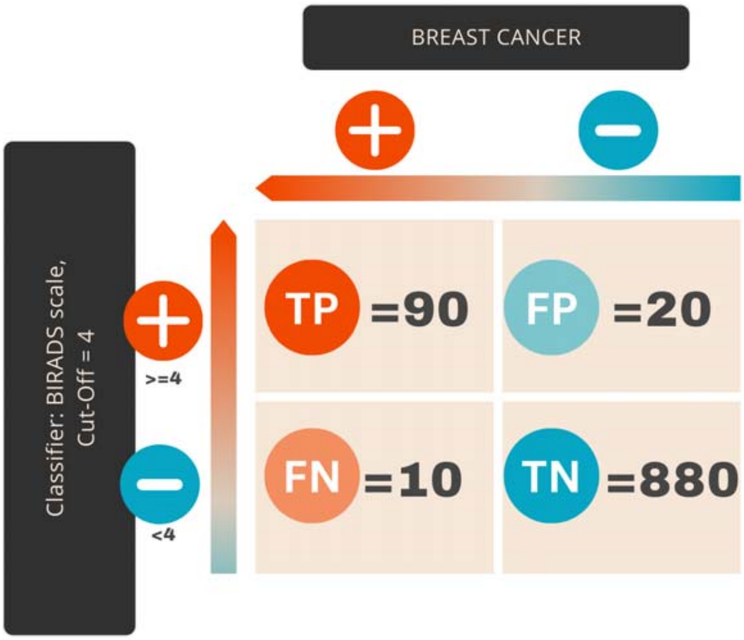


Figure 2. Example of a confusion matrix. In the columns of this matrix, a positive value indicates a positive class (occurrence of an event, in this case, breast cancer), and a negative value indicates a negative class (no event). In the rows of the table, a positive value indicates the detection of an event by the classifier (BIRADS scale >=4 indicates breast cancer), and a negative value indicates the non-detection of this event.

- The four cells of the confusion matrix represent the following: True Positive (TP): 90 patients (who have cancer) and are classified correctly as positive patients (cancer detected according to BIRADS).
- False Positive (FP): 20 patients (healthy), but are classified incorrectly as positive (cancer detected according to BIRADS).
- True Negative (TN): 880 patients (healthy) are classified correctly as negative (no cancer detected according to BIRADS).
- False Negative (FN): 10 patients (who have cancer), but are classified incorrectly as negative (no cancer detected according to BIRADS).

Table 1. Main metrics to evaluate the quality of classification and their definitions.

Term	Definition and ranges of metrics along with interpretation
	Measures based on the Threshold Metrics
Accuracy TP FP FN TN	The proportion of correct predictions among all predictions. Takes a value between 0 (no accuracy) and 1 (complete accuracy). $\frac{TP + TN}{TP + TN + FP + FN}$
Error TP FP FN TN	The proportion of incorrect predictions among all predictions. Takes a value between 0 (no error) and 1 (maximum error). Complementary metrics to Accuracy. $\frac{FP + FN}{TP + TN + FP + FN}$
Sensitivity (Recall or True Positive Rate) TP ○ FN ○	The proportion of true positive predictions (correctly identified sick individuals) out of all actual sick individuals. Takes a value between 0 (no sensitivity) and 1 (maximum sensitivity). $\frac{TP}{TP + FN}$
Specificity ○ FP ○ TN	The proportion of true negative predictions (correctly identified healthy individuals) out of all actual healthy individuals. Takes a value between 0 (no specificity) and 1 (maximum specificity) $\frac{TN}{TN + FP}$
G-mean TP FP FN TN	The geometric mean is sensitivity and specificity combined into a single result that balances both concerns. A G-mean value of 1 indicates a perfect balance between sensitivity and specificity, while a value close to 0 indicates an imbalance or poor performance of the classifier $\sqrt{sensitivity \times specificity} = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}}$
PPV (Precision) TP FP ○ ○	Precision, otherwise known as Positive Predictive Value (PPV) is the proportion of true positives among all positive predictions. Takes a value between 0 (lack of prediction skills for the positive class) and 1 (perfect prediction skills for the positive class) $\frac{TP}{TP + FP}$
NPV ○ ○ FN TN	Negative Predictive Value (NPV) is the proportion of true negatives among all negative predictions. Takes a value between 0 (lack of prediction skills for the negative class) and 1 (perfect prediction skills for the negative class) $\frac{TN}{TN + FN}$
F1-score TP FP FN ○	It focuses on the model's ability to identify positive instances. Precision and Recall combined into one result that tries to balance both concerns. It is calculated as the harmonic average of Precision and Recall. F-score values range from 0 to 1. The higher the F-score value, the better the classifier performs in balancing precision and recall. $2 \times \frac{precision \times recall}{precision + recall} = (\frac{2 \times TP}{2 \times TP + FP + FN})$
F(β) -score TP FP FN ○	It focuses on the model's ability to identify positive instances. A weighted harmonic mean of Precision and Recall. In this formula, β determines the weight assigned to Recall compared to Precision. A higher value of beta gives more weight to Recall, while a lower value of β favors Precision. When beta is equal to 1, the F-beta score is equivalent to the F1-score, which balances Precision and Recall equally $\frac{(1 + \beta^2) \times (precision \times recall)}{\beta^2 \times precision + recall} = \frac{(1 + \beta^2) \times (\frac{TP}{TP + FP} \times \frac{TP}{TP + FN})}{\beta^2 \times \frac{TP}{TP + FP} + \frac{TP}{TP + FN}}$

Table 1. Continuued.

Term	Definition and ranges of metrics along with interpretation
DOR TP FP FN TN	Diagnostic Odds Ratio (DOR) is the ratio of two chances: the chance of a positive classifier result from a diseased person to the chance of a positive classifier result from a healthy person. A higher DOR indicates a better discriminatory power of the classifier, with values greater than 1 suggesting higher odds of a positive classifier result in individuals with the condition compared to those without.
	$\frac{TP \div FN}{FP \div TN}$
MCC [26] TP FP FN TN	Mathews Correlation Coefficient (MCC) metrics the correlation of the true classes with the predicted labels. MCC ranges in the interval [−1,+1], with −1 meaning perfect misclassification and +1 perfect classification. MCC generates a high score in its interval only if the classifier scores a high value in all of the following: sensitivity, specificity, precision, and negative predictive value.
	$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$
Kappa (Cohen's Kappa Coefficient) TP FP FN TN	This metric evaluates the agreement between predicted and actual classes. Also, it considers that some correspondence between predicted and actual classes could occur by chance and eliminates random correspondence. Cohen's Kappa value ranges from -1 to 1, where a value of 1 indicates full agreement and a value of 0 indicates complete randomness. Negative values denote agreement weaker than random is rarely achieved in practice.
	$\frac{observed\ agreement\ - expected\ agreement}{1 - expacted\ agreement}$ or $2 \times \frac{TP \times TN - FN \times FP}{(TP+FP) \times (FP+TN) + (TP+FN) \times FN + TN}$
	Measures based on the Ranking Metrics (Methods)
ROC Curve TP FP FN TN	The Receiver Operating Characteristic Curve (ROC) is a graphical representation of the performance of a binary classifier system as its discrimination cut-off is varied. The higher the ROC curve rises above the diagonal (random line), the better the performance of the classifier. A classifier without skill formulates a line that winds along the diagonal.
	The Y-axis presents sensitivity (true positive rate, TPR), X-axis presents 1- specificity (false positive rate, FPR). Both values are obtained from the successive confusion matrices determined for each of the possible classifier cut-off points.
AUC(ROC) TP FP FN TN	The AUC(ROC) (Area Under the Curve (Receiver Operating Characteristic Curve) is interpreted as the probability that the model will return a higher probability of illness to an arbitrarily selected sick person than an arbitrarily selected healthy person. Where a value of 1 indicates ideal classification (perfect skill classifier), and a value of 0.5 indicates random classification (no skill classifier).
	AUC(ROC) is calculated as the sum of the areas of the trapezoids where the area under the ROC curve is divided.
PR Curve TP FP FN ○	Precision-Recall Curve (PR Curve) is a graphical representation of the performance of a binary classifier system that focuses on TP cases at different decision cut-offs. The closer the Precision-Recall curve is to the point (1,1), the better the performance of the classifier is. The no-skill line changes according to the balance of the classes. This is a horizontal line representing the proportion of positive cases in the data set. In the case of a balanced dataset, it is 0.5; if, for example, the sickness rate is 20%, the line is at 0.2.
	Build a graph with Precision as the y-axis and Recall as the x-axis. Both values are obtained from the successive confusion matrices determined for each of the possible classifier cut-off points.
AUC(PR) TP FP FN ○	The area under the Precision-Recall curve (AUC(PR) is used as a measure of the overall performance of the classifier, where a value of 1 indicates perfect classification (perfect skill classifier) and a value of 0 indicates the worst possible result (no skill classifier).
	AUC(PR) is calculated as the sum of the areas of the trapezoids where the area under the Precision-Recall curve is divided.

Table 1. Continuued.

Term	Definition and ranges of metrics along with interpretation
	Probabilistic Metrics
LogLoss (cross-entropy)	Logarithmic Loss (LogLoss), also known as the logarithmic loss function or cross-entropy loss, is a measure of the magnitude of error used in classification problems. It measures the degree of deviation between the actual values (0 for the no-event class and 1 for the event class) and the probabilities predicted by the classifier. LogLoss values are always non-negative, where a value of 0 indicates a perfect match between actual class values and predicted probabilities, a random model would have a log loss of around 0.693. The higher the LogLoss value, the greater the deviation between the predicted and actual values
	For binary classification is calculated as: $Log\ Loss = = \frac{-1}{N} \times \sum \quad [y \times log(\underline{y}) + (1 - y) \times (1 - y) \times log(1 - \underline{y})]$ where y is the actual value of the class (0 or 1), and $\underline{y}$ is the predicted probability of an event (disease)
Brier Score [Brier GW, Mean squared error)	Brier's score measures the mean square error between actual values (0 for the no-event class and 1 for the event class) and the probabilities predicted by the classifier. Brier's index takes values from 0 to 1. Lower values indicate better classification. An ideal classification would achieve a Brier index of 0, while a completely wrong, unreliable model would achieve a Brier index value of 1 and a random model would have a Brier score of around 0.25.
	$Brier\ Score = \frac{1}{N} \times \sum_{i=1}^{n} \quad (\hat{y}_i - y_i)^2$

old that converts these scores into binary predictions. Once such a threshold is specified, a confusion matrix can be constructed and threshold-based metrics derived from it.

Variables with multiple possible cut-off points are standard. During model development and comparative evaluation, relying on a single confusion matrix at a single chosen threshold can be problematic because threshold-based measures are sensitive to the selected cutoff, and the "optimal" threshold may differ across datasets, even for the same biomarker. In contrast, for a clinically implemented biomarker or diagnostic test, establishing and validating a single, pre-specified threshold that can be applied consistently across laboratories and settings is a strength rather than a limitation. Accuracy and Error may also be misleading in imbalanced datasets, which are frequent in medical studies where patient groups are much smaller than control groups (healthy individuals). Their practical application in current medical research often leads to the "Accuracy Paradox," particularly in imbalanced datasets. For instance, in rare disease screening where pathology prevalence is low (e.g., 1%), a naive classifier predicting all patients as healthy achieves 99% accuracy but fails clinically due to 0% sensitivity. This phenomenon is frequently observed in large-scale health record analyses, where neglecting metrics like Balanced Accuracy or G-mean can obscure a model's inability to detect the minority class of interest [25]. Positive predictive value (PPV) and negative predictive value (NPV) are often more clinically informative in these settings because they quantify the probability of disease given a test result; however, they are strongly dependent on disease prevalence, which limits their transportability between populations.

Threshold metrics, by contrast, offer greater versatility. These metrics can be determined regardless of the type of data on which the marker was measured. It is possible to decide on both quantitative and qualitative data, structured and binary data. The simplicity of calculating and interpreting these metrics is also essential. For instance, utilising a confusion matrix facilitates straightforward computation of various performance measures (as shown in **Table 1**) for BIRADS with the cut-off shown in **Table 2**.

As outlined in **Figure 6**, these metrics are the primary choice only when the core task is binary

Table 2. The calculation results obtained for the confusion matrix presented in Figure 1, together with their interpretation

Threshold Metric	Calculations and Conclusions
Accuracy	(880 + 90) / 1000 = 0.97
	97% of individuals were classified correctly
Error	(10 + 20) / 1000 = 0.03
	3% of individuals were classified incorrectly
Sensitivity (Recall or True Positive Rate)	90 / (10 + 90) = 0.9
	90% of sick individuals were correctly classified
Specificity	880 / (20 + 880) = 0.98
	98% of healthy individuals were correctly classified
G-mean	sqrt(0.9 * 0.98) = 0.94
	Mean sensitivity and specificity is 94%
PPV (Precision)	90 / (20 + 90) = 0.82
	82% of individuals with a positive result (BIRADS>=4) were actually sick.
NPV	880 / (10 + 880) = 0.99
	99% of individuals with a negative result (BIRADS < 4) were actually healthy.
F1-score	2 * ((0.82 * 0.9) / (0.82 + 0.9)) = 0.86
	0.86 indicates a very good balance between precision and recall.
F2-score	F2 score = (1 + 2^2) * ((0.82 * 0.9) / ((2^2 * 0.82) + 0.9)) = 0.88
	With the assumption that recall is twice as important as precision, we still point out that BIRADS classifies fairly accurately.
DOR	(90/10)/(20/880) = 396
	An individual with a positive test result is 396 times more likely to have the disease compared to someone with a negative test result.
MCC	((90*880)-(10*20))/sqrt((90+20)*(90+10)*(880+20)*(880+10)) = 0.842
	0.842 indicates a strong positive correlation and a substantial agreement between BIRADS and reality.
Kappa (Cohen's Kappa Coefficient)	2*(90*880-10*20)/((90+20)*(20+880)+(90+10)*(10+880)) = 0.84
	Agreement for classification between BIRADS and reality after taking into account the agreement.

classification, and the chosen threshold is fixed and clinically validated.

Threshold Optimisation: Selecting the Optimal Cut-off Point

The selection of an optimal cut-off point is crucial for threshold-based classification metrics such as Sensitivity, Specificity, and Accuracy. The choice of threshold directly impacts the classifier's performance and its clinical utility. While a fixed threshold may be appropriate in some cases, optimal threshold selection is often necessary to balance false positives and false negatives effectively. Several methods exist for determining the best threshold, including Youden's Index, cost-benefit analysis, and clinically relevant decision-making frameworks.

Youden's Index: Maximising Sensitivity and Specificity

Youden's Index is one of the most commonly used methods to select an optimal threshold. It is defined as:

$$J = Sensitivity + Specificity - 1$$

The optimal threshold is the point at which Youden's Index is maximised, meaning it provides the best trade-off between true positive rate and actual negative rate. This method is widely used in diagnostic tests where equal importance is placed on detecting disease (high Sensitivity) and avoiding misclassification of healthy individuals (high Specificity).

Cost-Benefit Analysis: Accounting for Clinical Consequences

In many real-world applications, the costs associated with FP and FN classifications are not equal. A cost-benefit analysis helps determine a threshold that minimises the overall impact of classification errors. This approach assigns a weight or cost to each type of error based on its clinical consequences. The optimal threshold is the one that minimises the expected total cost, which is calculated as:

$$Total\ Cost = (CFP \times FP\ Rate) + (CFN \times FN\ Rate)$$

Where *CFP* and *CFN* represent the relative costs of false positives and false negatives, respectively, for example, in cancer screening, a false negative (missed diagnosis) may have a significantly higher price (delayed treatment) than a false positive (leading to additional but unnecessary testing).

Clinically Relevant Decision-Making Frameworks

Beyond mathematical optimisation, clinical decision-making frameworks integrate real-world impact into threshold selection. For instance, in sepsis prediction models, a lower threshold may be preferred to increase early detection rates, even at the cost of a higher false-positive rate. Conversely, in cardiac risk stratification, a stricter threshold may be needed to prevent unnecessary interventions.

One example is the Net Benefit Approach, which considers both the relative utility of true positives and the harm of false positives in medical decision-making. This approach is frequently used in risk-based screening guidelines.

Ranking Metrics

Receiver operating characteristic (ROC) and precision-recall curves are standard methods for evaluating classifier performance. They enable comparison of different diagnostic parameters (or classification models), regardless of the decision threshold. We adjust the threshold by classifying test results as positive (sick) or negative (healthy) to derive these curves (see **Figure 4**). We then construct a confusion matrix to plot the ROC and PR curves for each threshold.

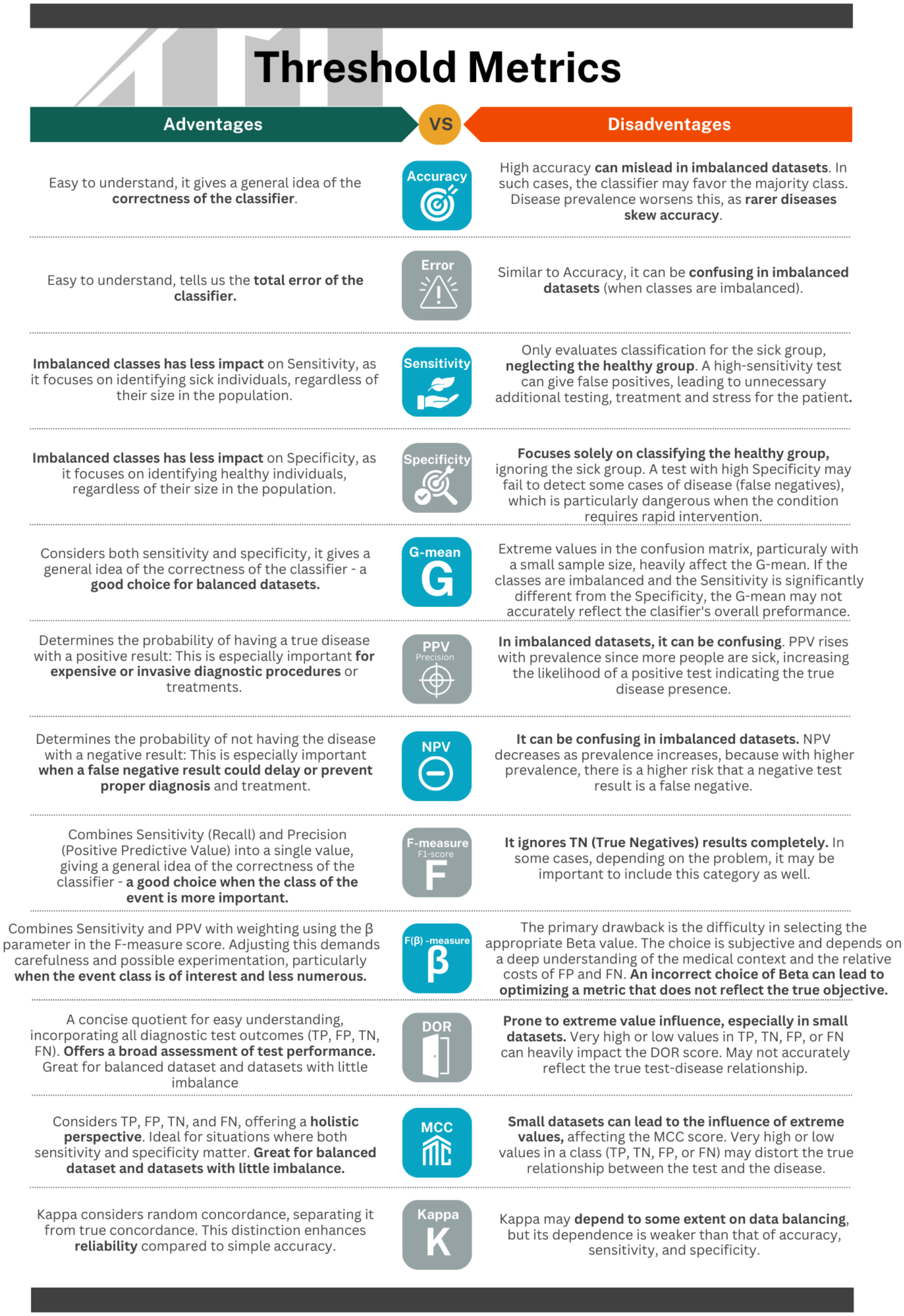
The ROC curve is based on Sensitivity (True Positive Rate, TPR) and a value of 1 1-Specificity (False Positive Rate, FPR = 1-Specificity). For different thresholds, points are plotted on the graph, creating an ROC curve that shows how the test balances between Sensitivity (Y-axis) and 1-Specificity (X-axis) as the classifier's thresholds are successively applied. The PR curve is based on Precision (PPV) and Recall (Sensitivity). We plot points for different thresholds, creating a PR curve that shows how the test balances Precision (Y-axis) and Recall (X-axis) as a function of the classifier's successive tapping thresholds.

Although each of these metrics is based on successive cut-off points (thresholds) and successive confusion matrices, they take into account slightly different aspects of the assessment of classification quality and have distinct advantages and disadvantages (see **Table 3**)

The widespread use of ROC curves is supported by the development of statistical tests to assess the significance of the AUC and to compare ROC curves, such as the DeLong method and the Hanley-McNeil test [26]. Additionally, techniques such as Youden's method facilitate the determination of optimal cutoff points based on ROC curves [26]. It is critical to distinguish two properties: while AUC (ROC) is robust in the sense that its value is mathematically independent of changes in class prevalence (resampling), it can be highly misleading or overly optimistic in imbalanced settings. This "optimism" stems from the fact that AUC (ROC) treats all false alarms (FP) and missed diagnoses (FN) equally, regardless of the actual costs and the clinical dominance of the minority class.

Consider a dataset with a 99:1 ratio of negative to positive examples. A conservative, non-trivial classifier might achieve very high Specificity (low FPR) but low Sensitivity TPR), leading to a misleadingly good AUC(ROC). This optimism arises because the ROC curve does not reflect the low overall prevalence of the positive class. To avoid this misleading optimism, a different metric, such as PR curves, is often preferred.

There are several ways to address this problem. One is to use a different metric, such as Precision-Recall curves. This metric is more sensitive to class distribution and can give a more accurate picture of a model's performance when the minority class is the sick class. See the examples of L-selectin and P-selectin in the detection of psoriasis, and the example of heart rate in the detection of coronary disease (see **Figure 5**).



Threshold Metrics

Advantages	VS	Disadvantages
Easy to understand, it gives a general idea of the **correctness of the classifier**.	Accuracy	High accuracy **can mislead in imbalanced datasets**. In such cases, the classifier may favor the majority class. Disease prevalence worsens this, as **rarer diseases skew accuracy**.
Easy to understand, tells us the **total error of the classifier.**	Error	Similar to Accuracy, it can be **confusing in imbalanced datasets** (when classes are imbalanced).
Imbalanced classes has less impact on Sensitivity, as it focuses on identifying sick individuals, regardless of their size in the population.	Sensitivity	Only evaluates classification for the sick group, **neglecting the healthy group**. A high-sensitivity test can give false positives, leading to unnecessary additional testing, treatment and stress for the patient**.**
Imbalanced classes has less impact on Specificity, as it focuses on identifying healthy individuals, regardless of their size in the population.	Specificity	**Focuses solely on classifying the healthy group,** ignoring the sick group. A test with high Specificity may fail to detect some cases of disease (false negatives), which is particularly dangerous when the condition requires rapid intervention.
Considers both sensitivity and specificity, it gives a general idea of the correctness of the classifier - a **good choice for balanced datasets.**	G-mean G	Extreme values in the confusion matrix, particuraly with a small sample size, heavily affect the G-mean. If the classes are imbalanced and the Sensitivity is significantly different from the Specificity, the G-mean may not accurately reflect the clasifier's overall preformance.
Determines the probability of having a true disease with a positive result: This is especially important **for expensive or invasive diagnostic procedures** or treatments.	PPV Precision	**In imbalanced datasets, it can be confusing**. PPV rises with prevalence since more people are sick, increasing the likelihood of a positive test indicating the true disease presence.
Determines the probability of not having the disease with a negative result: This is especially important **when a false negative result could delay or prevent proper diagnosis** and treatment.	NPV	**It can be confusing in imbalanced datasets.** NPV decreases as prevalence increases, because with higher prevalence, there is a higher risk that a negative test result is a false negative.
Combines Sensitivity (Recall) and Precision (Positive Predictive Value) into a single value, giving a general idea of the correctness of the classifier - **a good choice when the class of the event is more important.**	F-measure F1-score F	**It ignores TN (True Negatives) results completely.** In some cases, depending on the problem, it may be important to include this category as well.
Combines Sensitivity and PPV with weighting using the β parameter in the F-measure score. Adjusting this demands carefulness and possible experimentation, particularly **when the event class is of interest and less numerous.**	F(β) -measure β	The primary drawback is the difficulty in selecting the appropriate Beta value. The choice is subjective and depends on a deep understanding of the medical context and the relative costs of FP and FN. **An incorrect choice of Beta can lead to optimizing a metric that does not reflect the true objective.**
A concise quotient for easy understanding, incorporating all diagnostic test outcomes (TP, FP, TN, FN). **Offers a broad assessment of test performance.** Great for balanced dataset and datasets with little imbalance	DOR	**Prone to extreme value influence, especially in small datasets.** Very high or low values in TP, TN, FP, or FN can heavily impact the DOR score. May not accurately reflect the true test-disease relationship.
Considers TP, FP, TN, and FN, offering a **holistic perspective**. Ideal for situations where both sensitivity and specificity matter. **Great for balanced dataset and datasets with little imbalance.**	MCC	**Small datasets can lead to the influence of extreme values,** affecting the MCC score. Very high or low values in a class (TP, TN, FP, or FN) may distort the true relationship between the test and the disease.
Kappa considers random concordance, separating it from true concordance. This distinction enhances **reliability** compared to simple accuracy.	Kappa K	Kappa may **depend to some extent on data balancing**, but its dependence is weaker than that of accuracy, sensitivity, and specificity.

Figure 3. Advantages and disadvantages of individual Threshold Metrics.

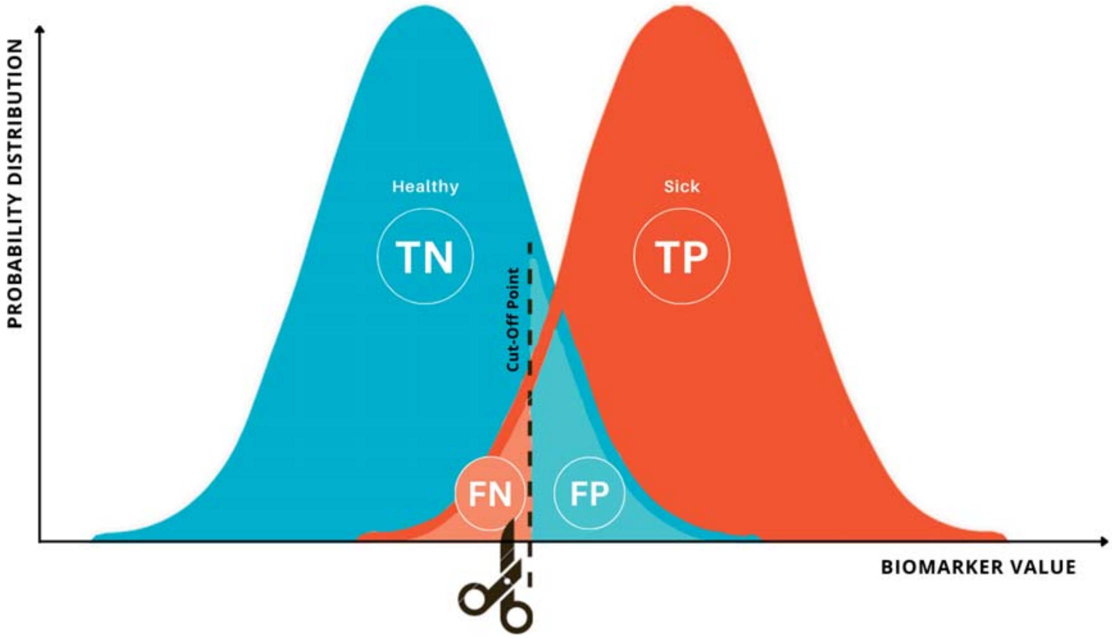
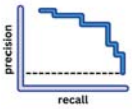


Figure 4. Outcomes of cut-off point (threshold) selection for a continuous biomarker as a potential classifier. The field sizes obtained at the indicated cutoff point yield four quantities: TP, TN, FP, and FN, which are entered into the confusion matrix.

Table 3. Advantages and Disadvantages of Ranking Metrics.

Ranking Metrics	Advantages	Disadvantages
ROC 	**Visual Representation:** ROC curves provide a visual representation of the performance of a classification model across all possible thresholds. **Comprehensive Performance Summary:** AUC (ROC) is a single metric that summarizes the overall performance of a diagnostic test. Represents the probability that a randomly chosen positive case will receive a higher test score than a randomly chosen negative case. **Threshold-Independent Comparison:** Two or more ROC curves can be compared directly even if they are derived from different variables with different units. **Independence from Prevalence:** The ROC coordinates (TPR, FPR) are mathematically independent of the class distribution (prevalence), making the AUC (ROC) value determinable even if the prevalence changes	**Limited Clinical Interpretation:** AUC (ROC), despite its popularity, may not directly translate into meaningful information for clinicians, patients, or healthcare providers. A test with an AUC of 0.9 might be considered "better" than one with an AUC of 0.8, but this difference may not have a significant impact on patient outcomes or treatment decisions. **Focus on All Thresholds:** AUC (ROC) considers the performance of a test across all possible thresholds, including those that may not be clinically relevant or useful in practice. This can lead to an overemphasis on thresholds that are not practical or important for decision-making. **Can be misleading in imbalanced settings:** This is because AUC (ROC) gives equal weight to TPR and FPR, regardless of the actual prevalence. In highly imbalanced datasets, the resulting high AUC may mask poor performance on the minority class, as the FPR remains small due to the dominant TN count. This often leads to a misleadingly high AUC (ROC) when compared to AUC (PR)
PR  	**Visual Representation:** PR curves provide a visual representation of the performance of a classification model at different Recall levels. This allows for a quick and intuitive understanding of the model's ability to identify positive cases (e.g., disease presence) while considering the trade-off between Precision and Recall **Focus on Positive Class:** PR curves specifically focus on evaluating the performance of a model in identifying positive cases (e.g., disease presence), making them particularly useful in scenarios where the detection of these cases is of primary importance. They provide precise information about the model's ability to correctly classify positive instances, even when they are rare or difficult to identify.	**Focus on Positive Class:** PR curves primarily focus on evaluating the performance of a model in identifying positive cases (e.g., disease presence), disregarding the number of true negative results (e.g., disease absence). This can make them less suitable for tasks where both positive and negative classifications are equally important. **Neglect of Healthy Individuals:** PR curves do not directly assess the model's ability to correctly classify healthy individuals (true negatives). **Sensitivity to Data Imbalance:** This sensitivity makes it challenging to compare PR curves from different studies or datasets with varying class imbalances.

Methodological note for illustrative examples

To ensure consistency, calculations for the provided examples (L-selectin, P-selectin, Heart Rate) were performed using individual probabilities obtained from a univariate Logistic Regression model trained on the respective variable (Table 6, Figure 5). Ranking metrics (AUC-ROC and AUC-PR) and probability metrics (Brier Score, Log Loss) were calculated directly from these continuous outputs without applying any arbitrary classification thresholds. The R code used to perform these calculations is available on GitHub (https://github.com/marpatra/Metrics.Selectins-HR)

The ROC curve for L-selectin lies along the line of identity. At the 0.5 level with a balanced dataset (50:50 ratio), the prevalence line is at 0.5 on the PR curve, indicating that L-selectin is a minimal-skill classifier (see **Figure 5**). While the AUC(PR) of 0.523 is technically above the random expectation baseline of 0.50, the value is negligibly close to the no-skill classifier line, justifying its practical classification as non-discriminatory. The ROC and PR curves for P-selectin indicate that low levels of this selectin have discriminatory potential; at high levels, the curves lie along a line indicating no skill.

For the balanced cardiac dataset, we obtained ROC and PR curves, yielding high fields and their plots, and also demonstrating strong classification performance for the heart rate variable. However, for the imbalanced data, the AUC(ROC) value of 0.776 is higher than that for the balanced data (0.745). This finding is due to the substantial increase in the TN, which pushes the FPR close to zero across many thresholds. This results in the ROC curve appearing overly optimistic. In sharp contrast, the AUC(PR) for the same imbalanced data drastically drops to 0.494 (from 0.704 in the balanced setting). This low AUC (PR) indicates the classifier's weakness in the minority class (patients). Specifically, the optimistic AUC(ROC) masks a clinically unacceptable drop in PPV, where many optimistic predictions are actually FP relative to the few TP available. In clinical practice, this would mean a high rate of false alarms, wasting resources and causing unnecessary patient anxiety, a critical factor missed by the ROC curve alone.

To mitigate imbalance during model training, techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or undersampling can be employed to balance class distribution. Alternatively, cost-sensitive learning can be used to assign higher penalties to misclassifying the minority class.

Figure 6 directs the user to these ranking metrics when the primary goal is classification, but when the data are imbalanced or a threshold-agnostic comparison between models is required.

Real-World Applications: Robust Metrics and Resampling Genomic Studies (Metric Selection): As highlighted in recent bioinformatics literature, standard metrics like F1-score can be misleading when the 'negative' class is biologically significant. Chicco and Jurman [21] demonstrated that, in genomic binary classification, the F1-score remained optimistically high even when the model failed to detect negative samples correctly. In contrast, the Matthews Correlation Coefficient (MCC) declined markedly in these scenarios, indicating the model's poor performance. Thus, for omics data, MCC is recommended over F1 as a more truthful performance indicator.

Pandemic Surveillance (Data Resampling):
In scenarios such as a COVID-19 diagnosis, datasets are often skewed toward negative cases. Research on radiomics applied to COVID-19 [2] has shown that models trained on imbalanced internal cohorts typically exhibit high Sensitivity but suffer a sharp decline in Specificity when validated externally due to overfitting. To mitigate this, techniques such as SMOTE (Synthetic Minority Over-sampling Technique) have been successfully employed to balance training sets, preventing the model from becoming biased toward the majority class (healthy individuals) and ensuring reliable detection of infected patients.

Robust Metrics and Resampling

Probabilistic Metrics

As shown in **Figure 6**, these metrics become the priority when the core modelling objective is the prediction of a trustworthy probability value (P), essential for accurate clinical risk assessment

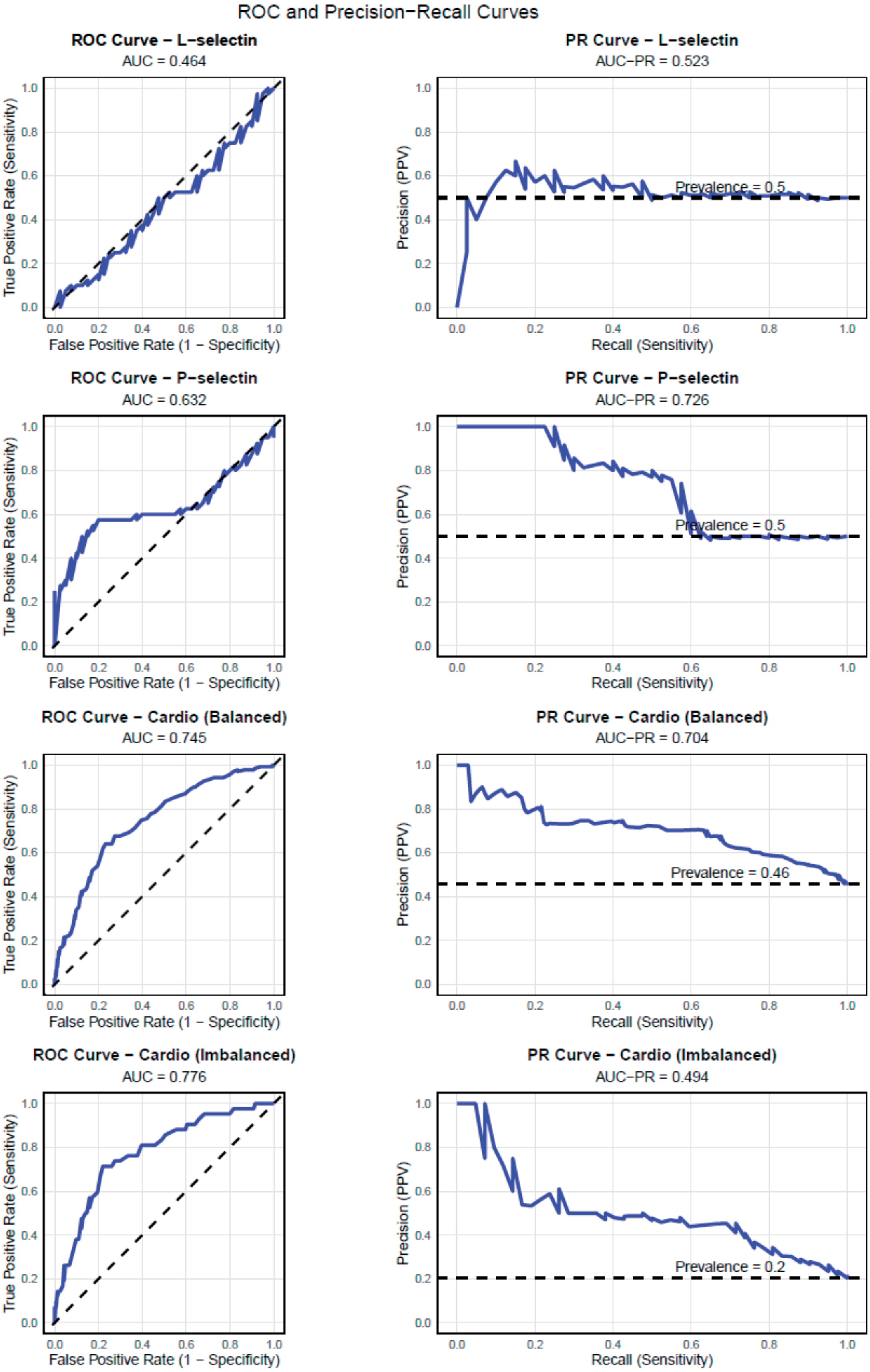


Figure 5. ROC and PR curves for balanced data: no-skill classifier using the example of L-selectin in psoriasis detection, partially-skill classifier using the example of P-selectin in psoriasis detection, skilful classifier using the example of heart rate in coronary disease detection for balanced data; partially skilful classifier using the example of heart rate in coronary disease detection for imbalanced data. Calculations were performed using individual probabilities obtained from a logistic regression model trained on the input variable. The results obtained without using the logistic regression model are identical.

When to Prefer Brier Score and Log Loss

Brier Score and Log Loss are essential metrics for evaluating probabilistic models, particularly when the focus is on the accuracy of predicted probabilities rather than binary classifications. These metrics are preferred over threshold-based metrics (e.g., accuracy, F1-score, sensitivity, specificity, ROC curves, and PR curves) in the following scenarios:

1. Probabilistic Models: When using models such as logistic regression, neural networks, or Bayesian classifiers, which output probabilities rather than binary predictions. These metrics are particularly suited for assessing the quality of probability estimates, which are often more informative than binary decisions in medical applications.
2. Calibration Assessment: When evaluating how well the predicted probabilities align with actual outcomes. For example, a well-calibrated model predicting a 30% risk of an event should observe the event occurring approximately 30% of the time. Calibration is critical in clinical decision-making, where accurate probability estimates are necessary for risk stratification and treatment planning.
3. Sensitivity to Small Errors: When the model's performance depends on accurately predicting probabilities, especially for rare events or imbalanced datasets. Log Loss, in particular, penalises overconfidence in incorrect predictions, making it a valuable tool for training and fine-tuning probabilistic models.

Evaluation of Continuous Probabilities: The Two-Stage Process

The evaluation of a diagnostic test often involves a critical two-stage process that requires different metrics: Stage 1: probability prediction (calibration), and Stage 2: final classification (thresholding).

Stage 1 metrics, such as the Brier Score and Log Loss, assess the quality of the model's raw probability output (P) before any decision threshold is applied. They quantify the model's calibration, ensuring that if the model predicts a probability P, the outcome occurs approximately P per cent of the time. This is critical because, for clinical decision-making, the expected probability P must be trustworthy.

The importance of calibration: an illustrative example

Metrics such as AUC (ROC) are primarily ranking metrics; they assess only the model's ability to correctly order positive cases above negative cases, regardless of the actual probability values. Two models can have identical AUC values but vastly different calibration quality. Consider a small dataset of 10 cases (4 positive (1), 6 negative (0)). Two hypothetical models, Model M1 (well-calibrated) and Model M2 (poorly-calibrated), produce the following probabilities (see **Table 4**):

Despite providing different probability scores, the rank order of cases is identical for both models. Consequently, both Model M1 and Model M2 achieve a perfect AUC(ROC) of 1.0. Based solely on the AUC, we would conclude that both models

Table 4. Comparison of Ranking (AUC) and Probabilistic (Brier Score) Metrics for Two Models with Identical Ranking Ability but Different Calibration (Numerical Example). The R code used to perform these calculations is available on GitHub (https://github.com/marpatra/Metrics.Selectins-HR)

Case	True label	M1 (calibrated prob.)	M2 (uncalibrated prob.)
1	1	0.60	0.90
2	1	0.55	0.85
3	1	0.40	0.70
4	1	0.30	0.60
5	0	0.25	0.55
6	0	0.15	0.40
7	0	0.10	0.30
8	0	0.05	0.20
9	0	0.02	0.10
10	0	0.01	0.05
Summary	AUC-ROC	1.00	1.00
	Brier Score	0.09	0.13

are perfect classifiers. However, when assessing calibration using the Brier Score (BS), the difference becomes clear: BS(M1) = 0.089, BS(M2) = 0.131. Since a lower Brier Score indicates better performance, Model M1 is significantly better calibrated than Model M2. Model M2 systematically underpredicts the risk for positive cases (e.g., predicting 0.60 instead of 0.90 for Case 1). Using Model M2 in a clinical setting would lead practitioners to be consistently overconfident that patients are not sick when a moderate probability is predicted, resulting in poorer clinical decisions despite the model's perfect AUC ranking.

For a more comprehensive visualisation and assessment of calibration, it is standard practice to use Calibration Plots or Reliability Diagrams alongside Brier Score and Log Loss. These graphical tools compare predicted probabilities with observed frequencies across multiple bins, providing an intuitive way to identify systemic biases in the probability output (e.g., over- or underestimation).

Stage 2 involves applying a decision threshold (cutoff point) to the calibrated probability output to produce the final binary classifications (e.g., 'sick' or 'healthy'). At this stage, threshold-dependent metrics such as Sensitivity, Specificity, PPV, NPV, F1-Score and Accuracy are used to assess the final classification performance based on the chosen trade-off.

Practical Applications in Medicine

Prognostic Models in Oncology:
In cancer survival prediction, binary classification (alive/dead) is often insufficient; clinicians require the survival probability to weigh treatment risks. Steyerberg et al. [27] emphasise that a model can have high Accuracy but poor calibration (e.g., consistently predicting 60% risk for patients who actually have a 40% risk). In such cases, the Brier Score is the superior metric because it quantifies the distance between the predicted probability and the actual outcome. A lower Brier score is associated with more reliable risk estimates, which are crucial for deciding whether to administer toxic chemotherapy.

Cardiovascular Risk Scoring:
Similarly, in predicting 10-year cardiovascular event risk (e.g., the Framingham Risk Score), Log Loss is widely used to penalise confident but incorrect predictions. If a model predicts a 99% chance of 'no heart attack' for a patient who subsequently suffers one, Log Loss applies a heavy penalty, forcing the algorithm to be more cautious and realistic in its probability estimates during training."

By incorporating Brier Score and Log Loss into the evaluation process, researchers and clinicians can ensure that their models provide not only accurate classifications but also reliable probability estimates, ultimately improving patient outcomes.

Implementations in Practice

Implementations of these measures are less widely available than other performance metrics, such as Accuracy or Precision. This can make it challenging to use these metrics in some programming environments. It is important to emphasise that Brier Score and Log Loss have strengths and weaknesses, and the choice of appropriate metrics depends on the specifics of the task (see **Table 5**).

Depending on the software used, the values obtained for these measures may vary minimally. In our examples (L-selectin and P-selectin for psoriasis detection, and heart rate for coronary disease detection), we first fitted simple logistic regression models with the respective variable as the predictor, yielding an individual predicted probability of the outcome for each patient. Brier Score was then calculated as the mean squared difference between these case-wise predicted probabilities and the observed outcomes, and Log Loss was computed using the exact individual probabilities. The results for calculations performed in R and in Python are presented in **Table 6**.

Validation

The data used to build and assess a classifier's quality is called learning or training data. In the following steps, each classifier, whether simple (based on a single variable) or complex, like a logistic regression model, neural network, or decision tree, should be validated with independent data, called test or validation data. Repeated testing of the same classifiers or models on

Table 5. Advantages and disadvantages of the chosen Probability Metrics.

Probability Metrics	Advantages	Disadvantages
Brier score	**Interpretability:** Brier Score is easier to interpret than Log Loss. It repre sents the mean squared difference between the predicted probabilities and the actual outcomes (0 or 1). A lower Brier Score indicates better model performance, making it intuitive to understand how well the model performs on average. **Robustness to calibration issues:** Even if a model's predicted probabilities are not perfectly aligned with the actual outcomes, Brier Score can still provide a reasonable assessment of performance.	**No penalty for overconfidence:** does not penalize models for being too confident but incorrect in their predictions. This can lead to preferring models that predict outstanding high/low values, even if they are wrong. In the case of extremely imbalanced data, where one class accounts for less than 1% of observations, **may not reflect the true effectiveness of the model.**
Log Loss	**Suitable for training models:** Minimizing Log Loss during training encourages the model to learn accurate probability estimates. **More sensitive to probability differences:** Log Loss is more sensitive to differences in predicted probabilities compared to Brier Score. This allows it to better distinguish between models that make subtle but significant improvements in probability estimation. **Discourages overconfidence:** Log Loss heavily penalizes models that are overly confident in their wrong predictions. This can be beneficial for tasks where assigning the correct probabilities is crucial.	**Limited interpretability:** Log loss is difficult to interpret in real-world terms. The metric is based on logarithmic values, which makes it difficult to interpret intuitively. **Assumes well-calibrated probabilities:** Log loss works best when a model predicts probabilities with high Accuracy. If a model is inaccurate or poorly calibrated, log loss may not be a reliable measure of performance. This can lead to preferring models that underpredict positive class probabilities, even if they better identify actual outcomes. In the case of extremely imbalanced data, where one class accounts for less than 1% of observations, **may not reflect the true effectiveness of the model.**

Table 6. Brier Score and Log Loss for balanced data for: no-skill classifier using the example of L-selectin in psoriasis detection, partially-skilled classifier using the example of P-selectin in psoriasis detection, skilful classifier using the example of heart rate in coronary disease detection; and for imbalanced data for a partially skilful classifier using the example of heart rate in coronary disease detection. Calculations were performed using individual probabilities obtained from a logistic regression model trained on the input variable and a null model (without variables). Results were presented using the LogLoss function from the MLmetrics package in R, and based on the LogisticRegression, log_loss, and NumPy functions from the scikit-learn and NumPy packages in Python. The data used for these calculations, along with the functions necessary to perform them, are available at https://github.com/marpatra/Metrics.Selectins-HR

	Classifier	Brier Score		Log Loss	
		Python	R	Python	R
L-selectin in psoriasis detection	no-skill	0.23050299261287152	0.230503	0.692726390827548	0.6927264
P-selectin in psoriasis detection	partially skillful	0.23457753518070662	0.2345775	0.656469675228019	0.6564697
heart rate in coronary disease detection	skillful	0.22963327943711911	0.2296331	0.6517807033032978	0.65178
heart rate in coronary disease detection	partially skillful	0.22963327943711911	0.2296331	0.65178	0.65178

***The R** and Python implementations produce similar outputs with slight differences due to numerical precision, optimisation algorithms, library settings, and convergence criteria.

a new dataset will indicate how well the original predictive model and its classifiers perform on new, unseen data. All listed metrics can be calculated on both the training set, to assess the current quality of classification and prediction, and on new validation and test datasets, to generalise this quality to future data on which it may be used.

Model Validation and Overfitting

Model validation is a critical step in ensuring the reliability and generalizability of classification models. Overfitting occurs when a model performs exceptionally well on the training data but fails to generalise to new, unseen data. This typically happens when the model learns noise or specific patterns in the training data that do not apply to the broader population.

Techniques to Prevent Overfitting

k-Fold Cross-Validation: In k-fold cross-validation, the dataset is divided into *k* subsets (folds). The model is trained on *k-1* folds and validated on the remaining fold. This process is repeated *k* times, with each fold used exactly once as the validation set. The results are averaged to provide an estimate of model performance. For instance, k-fold cross-validation is essential for obtaining a stable estimate of AUC(ROC) for the Heart Rate classification model, ensuring that the reported performance is not specific to a single data split. **Leave-One-Out Cross-Validation (LOOCV)**: LOOCV is a special case of k-fold cross-validation in which *k* equals the number of samples in the dataset. Each sample is used once as a validation set, while the remaining samples form the training set. This method is beneficial for small datasets, as it maximises the use of available data. This technique could be employed to rigorously estimate the Sensitivity and Specificity of the BI-RADS scale classifier (e.g., at the >4 threshold) when validating its performance in small, limited patient cohorts.

Bootstrap Methods: Bootstrap resampling is a resampling technique used to assess the variability and internal stability of model performance within the same underlying population (e.g., the same dataset). By repeatedly drawing samples with replacement from the original data and refitting or re-evaluating the model, bootstrap methods provide estimates of the uncertainty and optimism of performance measures (e.g. AUC, Brier Score). However, bootstrap resampling cannot replace evaluation in genuinely different patient populations. Assessment of model performance across settings or populations requires external validation on separate datasets, rather than resampling from a single cohort. Bootstrapping is highly useful for assessing the stability and confidence intervals of the AUC(PR) and Brier Score values reported for the selectin and heart rate models, providing a measure of how much these metrics might vary across different potential patient samples, and thus detecting overfitting.

Implications for classification Metrics: practical examples

Validation techniques are not merely procedural steps; they often reveal critical flaws in metric interpretation that theoretical calculations on training data miss.

Consider a study that employs high-dimensional genomic data to predict cancer subtypes (e.g., the MAQC-II study [28]). A classifier might achieve an Accuracy of 98% and an AUC of 0.99 on the training set due to the model memorising noise (overfitting). However, when subjected to 10-fold cross-validation, the AUC might drop drastically to 0.60. This discrepancy serves as a red flag that the initial high metrics were illusory.

Similarly, in radiomics studies for COVID-19 detection [2], models often show high Sensitivity on training cohorts. However, external validation on data from a different hospital typically indicates a significant drop in Specificity, resulting in a high number of False Positives. This happens because the model may learn scanner-specific artefacts rather than disease pathology. In such cases, relying solely on training F1-scores would be misleading; cross-validation highlights the need for metrics such as the Matthews Correlation Coefficient (MCC), which is more robust to such shifts in confusion-matrix distributions than the F1-score or Accuracy.

Practical Considerations in Medical Applications

In medical applications, where the stakes are high, ensuring that a model generalises well to new data is crucial. Overfitting can lead to overly optimistic performance estimates, potentially resulting in the deployment of unreliable diagnostic tools. Proper validation techniques help mitigate this risk, ensuring that the model's performance is consistent across different datasets and populations. For example, a model trained to predict sepsis must be validated on diverse patient cohorts to ensure its reliability in real-world clinical settings.

Discussion

Summary of Metric Strengths and Weaknesses within the Clinical Context

No single metric is universally optimal. Their selection must reflect the clinical context, priorities, and data characteristics.

Holistic Metrics: Accuracy (ACC) and Error Rate are useful when data are balanced, and misclassification costs are symmetric. Cohen's Kappa corrects these scores for chance agreement [27]. Matthews Correlation Coefficient (MCC) combines holistic assessment with robustness to class imbalance, making it a recommended metric in projects regulated by the U.S. FDA [29,30]. The Diagnostic Odds Ratio (DOR) summarises the overall discriminatory effectiveness of a test. The ROC curve and the area under it (AUC) are used to assess a test's ability to distinguish between groups across various thresholds [31-33]. The PR (Precision-Recall) curve is more informative than the ROC curve for detecting rare events [34]. The selection of an optimal cutoff point can be performed using Youden's index or, better adapted to clinical realities, the tangent (cost) method, which explicitly incorporates the relative costs of FP and FN and prevalence.

Class-Oriented Metrics: As indicated above, Sensitivity, Specificity, PPV, and NPV are fundamental, and their relevance depends on the clinical objective. The F1 score (and its variants) is instrumental in settings with class imbalance, as it combines precision (PPV) and sensitivity (recall).

Calibration Assessment Metrics: Brier Score and Log Loss assess the accuracy of estimated probabilities. The Brier Score is easier to interpret, whereas log loss is more sensitive to minor errors and is commonly used in machine learning.

Guidelines for Metric Selection and Their Clinical Rationale

The prevalence of a disease has a fundamental impact on the interpretation of diagnostic test

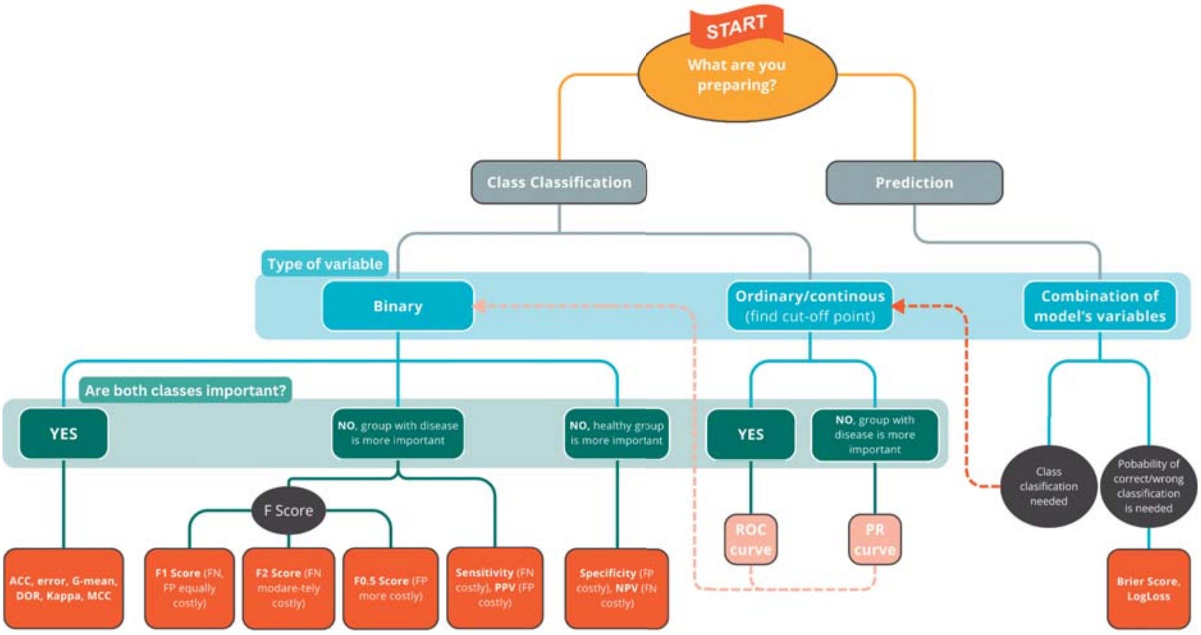

START
What are you preparing?
Class Classification
Prediction
Type of variable
Binary
Ordinary/continous (find cut-off point)
Combination of model's variables
Are both classes important?
YES
NO, group with disease is more important
NO, healthy group is more important
YES
NO, group with disease is more important
Class clasification needed
Pobability of correct/wrong classification is needed
F Score
ROC curve
PR curve
ACC, error, G-mean, DOR, Kappa, MCC
F1 Score (FN, FP equally costly)
F2 Score (FN modare-tely costly)
F0.5 Score (FP more costly)
Sensitivity (FN costly), PPV (FP costly)
Specificity (FP costly), NPV (FN costly)
Brier Score, LogLoss

Figure 6. Graph facilitating the selection of a metric depending on the purpose for which it is determined (assignment to classes, prediction of probability of assignment to classes), type of data, validity or balancing of classes of events (sick) and no events (healthy). While this graph helps establish the primary objective, in practice, it is common to use a combination of indicators to evaluate the model's performance fully.

results and the choice of evaluation metrics. A structured selection framework, presented as a decision tree (see **Figure 6**), comprises three steps: the problem type (classification vs. prediction), the type of target variable, and the relative importance of classes for a given task.

The decision in the final step in medicine is critically determined by the asymmetry of misclassification costs, which varies with the stage of the diagnostic process. This is vividly illustrated in **Figure 7**, which presents the trade-off in cancer screening. At the screening stage, the primary goal is to rule out the disease, making a False Negative (FN) – missing a sick patient – the most critical error, as it delays potentially life-saving treatment. Consequently, screening tests are optimised for high Sensitivity and Negative Predictive Value (NPV). It is crucial to note, however, that a positive screening result typically triggers further confirmatory steps rather than immediate aggressive therapy. At this subsequent confirmatory diagnostic stage, the cost of a False Positive (FP) – subjecting a healthy person to invasive procedures and psychological distress – becomes predominant. Therefore, confirmatory tests must exhibit high Specificity and Positive Predictive Value (PPV) to ensure that treatment is administered only to those who genuinely need it. In practice, minimising one type of error often increases the other, and a common compromise is to use balanced metrics such as the G-mean or metrics from the F-score family.

A Critical Overview and Future Directions

A conscious selection of metrics – often involving a combination of several – is essential for reliable evaluation. The metrics discussed thus far represent established, mathematically rigorous approaches to model assessment. However, the field continues to evolve, driven by the need for more intuitive, actionable, and human-centric evaluation tools.

Classical metrics, together with considerations of disease prevalence and the asymmetric clinical costs of errors at different stages of the patient pathway, enable the selection of measures that ensure a clinically sound and accurate evaluation of diagnostic and predictive models. The future of model evaluation in medicine, however, lies in the synergy between these traditional foundations and the development of new techniques necessary for deploying AI systems (e.g., Shapley Additive exPlanations [35], U-smile [36]. New methods are constantly being introduced, and it was not possible to discuss and present them in a single summary.

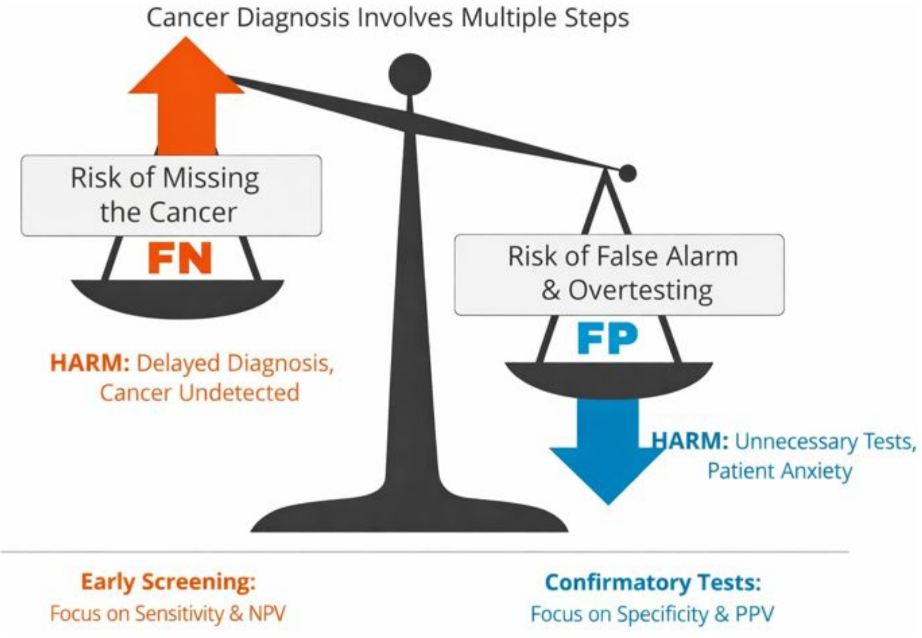

Cancer Diagnosis Involves Multiple Steps
Risk of Missing the Cancer
FN
HARM: Delayed Diagnosis, Cancer Undetected
Risk of False Alarm & Overtesting
FP
HARM: Unnecessary Tests, Patient Anxiety
Early Screening:
Focus on Sensitivity & NPV
Confirmatory Tests:
Focus on Specificity & PPV

Figure 7. The harms caused by false negative and false positive prediction errors in the BI-RADS scale. A False Negative (FN) is typically weighted more heavily than a False Positive (FP), as the primary goal is to identify all potential cases to avoid the risk of disease progression.

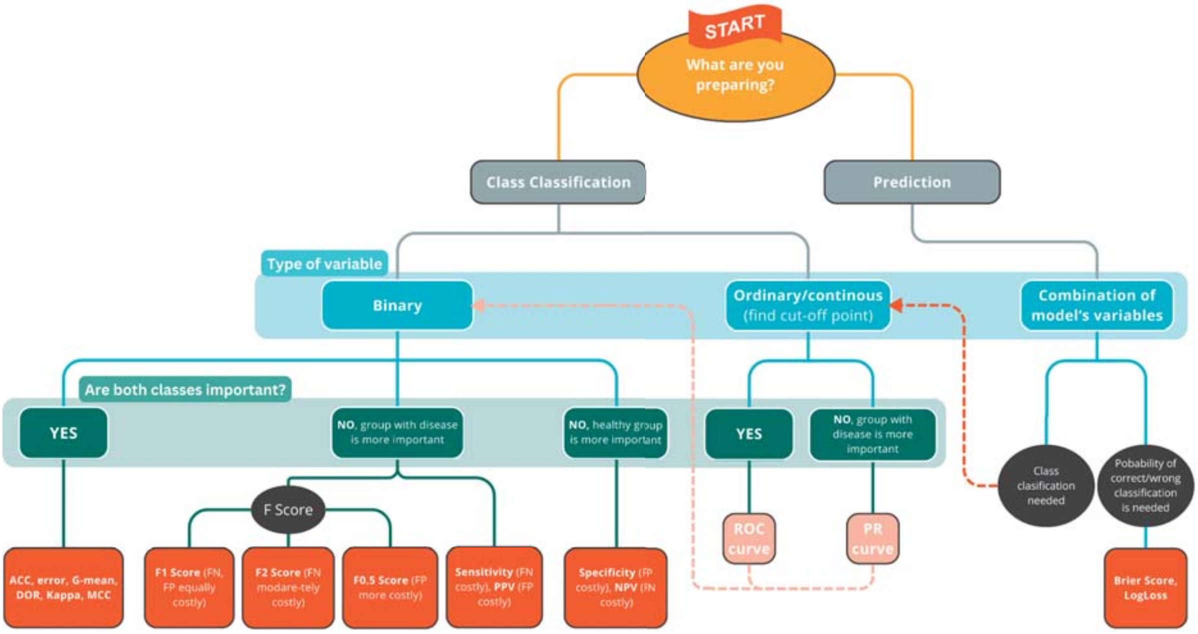

START
What are you preparing?
Class Classification
Prediction
Type of variable
Binary
Ordinary/continous (find cut-off point)
Combination of model's variables
Are both classes important?
YES
NO, group with disease is more important
NO, healthy group is more important
YES
NO, group with disease is more important
F Score
Class clasification needed
Pobability of correct/wrong classification is needed
ROC curve
PR curve
ACC, error, G-mean, DOR, Kappa, MCC
F1 Score (FN, FP equally costly)
F2 Score (FN modare-tely costly)
F0.5 Score (FP more costly)
Sensitivity (FN costly), PPV (FP costly)
Specificity (FP costly), NPV (FN costly)
Brier Score, LogLoss

Figure 8. Algorithm for statistical decision-making.

It is important to note that this tutorial focuses on binary classification. In contrast, multi-class problems—along with their corresponding metrics, interpretability estimation methods, and associated evaluation challenges—constitute an essential and natural direction for future work in this rapidly evolving field.

Acknowledgements

Conflict of interest statement
The authors declare no conflict of interest.

Funding sources
There are no sources of funding to declare.

References

1. Rose G. Sick individuals and sick populations. Int J Epidemiol. 2001;30(3):427–432. https://doi.org/10.1093/ije/30.3.427.
2. Shandhi MMH, Cho PJ, Roghanizad AR, Singh K, Wang W, Enache OM, et al. A method for intelligent allocation of diagnostic testing by leveraging data from commercial wearable devices: a case study on COVID-19. NPJ Digit Med. 2022;5(1):130. https://doi.org/10.1038/s41746-022-00672-z.
3. Xi Y, Ding Y, Cheng Y, Zhao J, Zhou M, Qin S. Evaluation of the medical resource allocation: evidence from China. Healthcare (Basel). 2023;11(6):829. https://doi.org/10.3390/healthcare11060829.
4. Guzik P, Więckowska B. Data distribution analysis – a preliminary approach to quantitative data in biomedical research. J Med Sci. 2023;92(2):e869. https://doi.org/10.20883/medical.e869.
5. George DB, Taylor W, Shaman J, Rivers C, Paul B, O'Toole T, et al. Technology to advance infectious disease forecasting for outbreak management. Nat Commun. 2019;10(1):3932. https://doi.org/10.1038/s41467-019-11901-7.
6. Myers A, Johnston N, Rathore S, Kwon D, Kline J, Jehi L, et al. Electronic health record-based prediction model for acute kidney injury in patients undergoing major gastrointestinal surgery: a pilot study. J Pers Med. 2020;10(1):21. https://doi.org/10.3390/jpm10010021.
7. Flaks-Manov N, Topaz M, Hoshen M, Balicer RD, Shadmi E. Identifying patients at highest-risk: the best timing to apply a readmission predictive model. BMC Med Inform Decis Mak. 2019;19:118. https://doi.org/10.1186/s12911-019-0836-6.
8. Skov Benthien K, Kart Jacobsen R, Hjarnaa L, Mehl Virenfeldt G, Rasmussen K, Toft U. Predicting individual risk of emergency hospital admissions – a retrospective validation study. Risk Manag Healthc Policy. 2021;14:3865–3872. https://doi.org/10.2147/RMHP.S314588.
9. Berchialla P, Lanera C, Sciannameo V, Gregori D, Baldi I. Prediction of treatment outcome in clinical trials under a personalized medicine perspective. Sci Rep. 2022;12(1):4115. https://doi.org/10.1038/s41598-022-07801-4.
10. Selby JV, Fireman BH. Building predictive models for clinical care—where to build and what to predict? JAMA. 2015;313(17):1705–1706. https://doi.org/10.1001/jama.2015.3680.
11. Battineni G, Sagaro GG, Chintalapudi N, Amenta F. Applications of machine learning predictive models in the chronic disease diagnosis. J Pers Med. 2020;10(2):21. https://doi.org/10.3390/jpm10020021.

12. Ghaffar Nia A, Kaplan M, Khelifi A, et al. Evaluation of artificial intelligence techniques in disease diagnosis and prediction. Discov Artif Intell. 2023;3:5. https://doi.org/10.1007/s44163-023-00049-5.
13. Toma I, Wei Y. The application of artificial intelligence methods to public health data. Encyclopedia. 2023;3(2):590–601. https://doi.org/10.3390/encyclopedia3020042.
14. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. 2023;23:689. https://doi.org/10.1186/s12909-023-04698-z.
15. Mansouri A, Mencattini A, Salmeri M, et al. A HUME approach for causal effects estimation in presence of unmeasured confounding. Bioinformatics. 2018;34(24):4274–4283. https://doi.org/10.1093/bioinformatics/bty490.
16. Ferner RE, Aronson JK. Susceptibility to adverse drug reactions. Br J Clin Pharmacol. 2019;85(10):2205–2212. https://doi.org/10.1111/bcp.14017.
17. Twick I, de Vetten JH, ten Berge M, et al. Performance measures for machine learning in a neonatal intensive care unit: a systematic review. J Am Med Inform Assoc. 2022;29(6):1064–1074. https://doi.org/10.1093/jamia/ocac036.
18. Pavlou M, Ambler G, Seaman SR, et al. How to develop a more accurate risk prediction model when there are few events. BMJ. 2015;351:h3868. https://doi.org/10.1136/bmj.h3868.
19. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proc 23rd Int Conf Mach Learn (ICML). New York: ACM; 2006. p.233–240. https://doi.org/10.1145/1143844.1143874.
20. Canbek G, Taskaya Temizel T, Sagiroglu S. PToPI: a comprehensive review, analysis, and knowledge representation of binary classification performance measures/metrics. SN Comput Sci. 2022;4(1):13. https://doi.org/10.1007/s42979-022-01409-1.
21. Chicco D, Jurman G. The Matthews correlation coefficient (MCC) is more informative than Cohen's kappa and Brier score in binary classification assessment. IEEE Access. 2023;16:4. https://doi.org/10.1109/ACCESS.2023.3301604.
22. Ferri C, Hernández-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. Pattern Recognit Lett. 2009;30(1):27–38. https://doi.org/10.1016/j.patrec.2008.08.010.
23. Choi HY, Park HE, Seo H, et al. Fasting plasma glucose and glycated hemoglobin cutoffs for predicting diabetes and prediabetes: the Korean genome and epidemiology study. J Korean Med Sci. 2018;33(50):e93. https://doi.org/10.3346/jkms.2018.33.e93.
24. Wai JH, Turner RM, Koeman J, et al. Outcome analysis for BI-RADS category 3 in the national mammography database. Diagn Interv Imaging. 2017;98(3):179–190. https://doi.org/10.1016/j.diii.2017.01.003.
25. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–1284. https://doi.org/10.1109/TKDE.2008.239.
26. Hassanzad F, Fryback D, Hosmer D, et al. Optimal cut-point analysis: an updated review of methods in medical research. J Appl Stat. 2024;51(4):1222–1242. https://doi.org/10.1080/02664763.2022.2130717.
27. Naulaerts S, et al. The impact of feature selection on the performance of imbalanced binary classification problems. Oncotarget. 2017;8:109343–109353. (DOI could not be verified online; recommend confirming source.)
28. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2.
29. Shi L, Reid LH, Jones WD, et al. The microarray quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol. 2010;28(8):827–838. https://doi.org/10.1038/nbt.1665.
30. Su Z, Łabaj PP, Li S, et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the SEQC/MAQC-III consortium. Nat Biotechnol. 2014;32(9):903–914. https://doi.org/10.1038/nbt.2957.
31. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32–35. https://doi.org/10.1002/1097-0142-(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.
32. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem. 1993;39(4):561–577. https://doi.org/10.1093/clinchem/39.4.561.
33. Obuchowski NA. ROC analysis. AJR Am J Roentgenol. 2005;184(2):364–372. https://doi.org/10.2214/ajr.184.2.01840364.
34. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432.
35. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, von Luxburg U, Bengio S, et al., editors. Advances in Neural Information Processing Systems 30. Red Hook (NY): Curran Associates; 2017. p.4765–4774.
36. Więckowska B, Kubiak KB, Guzik P. Evaluating the three-level approach of the U-smile method for imbalanced binary classification. PLoS One. 2025;20(4):e0321661. https://doi.org/10.1371/journal.pone.0321661.