European Journal of Radiology Artificial Intelligence 3 (2025) 100030

Contents lists available at ScienceDirect

European Journal of Radiology Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/european-journal-of-radiology-artificial-intelligence

Evaluation metrics in medical imaging AI: fundamentals, pitfalls, misapplications, and recommendations

Burak Kocak [a,*,1], Michail E. Klontzas [b,c,d,2], Arnaldo Stanzione [e,3], Aymen Meddeb [f,g,4], Aydın Demircioğlu [h,5], Christian Bluethgen [i,6], Keno K. Bressem [j,k,7], Lorenzo Ugga [l,8], Nathaniel Mercaldo [m,9], Oliver Díaz [n,10], Renato Cuocolo [o,11]

[a] Department of Radiology, Basaksehir Cam and Sakura City Hospital, Istanbul, Turkey
[b] Artificial Intelligence and Translational Imaging (ATI) Laboratory, Department of Radiology, School of Medicine, University of Crete, Heraklion, Crete, Greece
[c] Computational Biomedicine Laboratory, Institute of Computer Science, Foundation for Research and Technology (ICS-FORTH), Heraklion, Crete, Greece
[d] Division of Radiology, Department of Clinical Science Intervention and Technology (CLINTEC), Karolinska Institute, Stockholm, Sweden
[e] Department of Advanced Biomedical Sciences, University of Naples "Federico II" Via S. Pansini 5 – 80131 Naples, Italy
[f] Department of Neuroradiology, Charité – Universitätsmedizin Berlin, Berlin, Germany
[g] Berlin Institute of Health at Charité – Universitätsmedizin Berlin
[h] Institute of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, 45147 Essen, Germany
[i] Institute for Diagnostic and Interventional Radiology, University Hospital Zurich, University of Zurich, Zurich, Switzerland
[j] Department of Diagnostic and Interventional Radiology, Technical University of Munich, School of Medicine and Health, Klinikum rechts der Isar, TUM University Hospital, Ismaninger Str. 22, 81675 Munich, Germany
[k] Department of Cardiovascular Radiology and Nuclear Medicine, Technical University of Munich, School of Medicine and Health, German Heart Center, TUM University Hospital, Lazarethstr. 36, 80636 Munich, Germany
[l] Department of Advanced Medical and Surgical Sciences, University of Campania "Luigi Vanvitelli", P.zza L. Miraglia 2 - 80138 Naples, Italy
[m] Department of Radiology, Massachusetts General Hospital, Boston, MA, USA
[n] Departament de Matemàtiques i Informàtica. Universitat de Barcelona, Spain
[o] Department of Medicine, Surgery, and Dentistry, University of Salerno, Baronissi, Italy

ARTICLE INFO

Key words:
Artificial Intelligence
Diagnostic Imaging
Machine Learning
Computer-Assisted
Algorithms
Reproducibility of Results
Evaluation Studies as Topic
Sensitivity and Specificity

ABSTRACT

Robust assessment of artificial intelligence (AI) models in medical imaging is paramount for reliable clinical integration. This international collaborative review paper provides an overview of key evaluation metrics across diverse tasks, including classification, regression, survival analysis, detection, and segmentation, as well as specialized metrics for calibration, foundation models, large language models, and synthetic images. Challenges of comparing models statistically and translating metric scores to clinical practice are also discussed. For each section, the paper outlines fundamental metrics, identifies common pitfalls and misapplications, and offers recommendations for more robust evaluations. Key recommendations often involve utilizing multiple, complementary metrics tailored to the specific task and dataset properties, transparent reporting of methodology, and critically, considering the clinical utility and real-world implications of model performance. Ultimately, effective evaluation requires a comprehensive, context-aware approach that goes beyond statistical metrics to ensure

* Correspondence to: Department of Radiology, Basaksehir Cam and Sakura City Hospital, Basaksehir, Istanbul 34480, Turkey.
E-mail address: drburakkocak@gmail.com (B. Kocak).
[1] 0000–0002-7307–396X
[2] 0000–0003-2731–933X
[3] 0000–0002-7905–5789
[4] 0000–0001-6537–9419
[5] 0000–0003-0349–5590
[6] 0000–0001-7321–5676
[7] 0000–0001-9249–8624
[8] 0000–0001-7811–4612
[9] 0000–0003-1658–6598
[10] 0000–0001-6789–5177
[11] 0000–0002-1452–1574

https://doi.org/10.1016/j.ejrai.2025.100030
Received 12 June 2025; Received in revised form 1 July 2025; Accepted 4 July 2025
Available online 8 July 2025

model trust and clinical relevance. The authors hope this review will serve as a practical reference for researchers aiming to implement robust and clinically meaningful AI evaluations in medical imaging.

Introduction

Artificial intelligence (AI) has rapidly become a transformative force in medical imaging, enabling diverse applications such as disease classification, lesion detection, image segmentation, prognosis estimation, and synthetic image generation [1–3]. As these AI tools are being increasingly adopted in clinical practice, the need for rigorous, reliable, and context-sensitive evaluation is not merely important but essential for ensuring their safe and effective deployment [4–6].

The current performance evaluation of AI models predominantly relies on quantitative metrics. However, the utility of these metrics is often constrained by their variability across different task types, datasets, and underlying methodological assumptions [7–10]. This challenge is further compounded by the recent advent of foundation models and generative AI, which introduce novel complexities for robust performance measurement [11]. Crucially, inappropriate metric selection or flawed application can lead to misleading interpretations of model performance, potentially impacting downstream clinical decisions and patient care [8,10].

A comprehensive understanding of evaluation metrics, including their strengths, inherent limitations, and appropriate use-cases, is indispensable. This review aims to bridge the current gap by providing a structured overview of key performance metrics relevant to most common AI tasks in medical imaging. For each section, we will introduce fundamentals about metrics, delineate common pitfalls in metric application, and offer actionable recommendations for robust and clinically meaningful model assessments.

Classification metrics

Fundamentals

The robust assessment of classification models is paramount in medical imaging AI, as the choice of metrics can considerably influence the interpretation of model performance and subsequent decision-making. Classification metrics evaluate how well a model assigns inputs to appropriate categories and are typically divided into those derived from a fixed decision threshold (often confusion matrix-based) and those that summarize performance across multiple thresholds (**Table 1**) [8].

The confusion matrix compares model predictions to a reference standard, yielding values such as true positives, false positives, true negatives, and false negatives. From these, commonly used fixed-threshold metrics, such as accuracy, sensitivity, specificity, predictive values (positive and negative), F1-score, and Matthews correlation coefficient (MCC), are derived. Definitions and characteristics are provided in Table 1.

In binary classification tasks, models produce probability scores (either directly or through post hoc calibration) for each class. By varying the decision threshold, one can compute multi-threshold (or threshold-agnostic) metrics, such as the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) [12]. While these metrics aggregate performance across all thresholds, they do not reflect performance at a specific operating point, unlike fixed-threshold metrics such as accuracy or F1-score.

Pitfalls and misapplications

A primary pitfall in evaluating classification models stems from class imbalance within the dataset (**Fig. 1**), which often reflects the natural prevalence of the conditions being studied. When one class substantially outnumbers the other(s), accuracy can be directly skewed, presenting a deceptively optimistic view if the model simply predicts the majority class [13].

Similarly, AUROC can be misleading in imbalanced scenarios. While it offers a threshold-independent measure of overall discrimination, it may remain high even if the model performs poorly on the minority (often critical) class. This is because AUROC includes true negatives, which dominate in such datasets, thus inflating the score [14].

The underlying population prevalence of the condition under study is a critical factor, particularly when assessing a model's real-world clinical utility. Positive and negative predictive values (PPV/NPV) are inherently dependent on this prevalence [15]. Ignoring this dependency when interpreting PPV/NPV or applying them to populations with different prevalences without appropriate consideration constitutes a significant misapplication and can lead to erroneous conclusions about a model's performance in specific clinical contexts.

In multiclass classification, a common pitfall is the uncritical use of averaging strategies (macro, micro, weighted) for metrics such as precision, recall, F1-score, and AUROC. Without specifying these methods, results can misleadingly suggest balanced performance while masking poor outcomes in underrepresented classes [16].

Recommendations

In imbalanced classification tasks (**Fig. 1**), relying solely on accuracy is inadvisable because of its tendency to be misleadingly high. Given AUROC's potential to mask poor minority class performance, as previously discussed, a more insightful evaluation can be achieved using AUPRC, which better reflects a model's ability to identify rare positive instances [14,17].

AUPRC emphasizes the trade-off between precision (the proportion of true positives among all positive predictions, also called PPV) and recall (the proportion of actual positives correctly identified, equivalent to sensitivity). Because it excludes true negatives, it is often more informative than AUROC when the positive class is rare. As shown in Fig. 2, the baseline for AUPRC (random performance) corresponds to the prevalence of the positive class, providing a clearer benchmark for evaluating model performance in rare event detection.

When evaluating performance at a specific decision threshold, which is often crucial for practical applications, the F1-score offers a balanced assessment of the positive class by harmonizing precision and recall. MCC provides another robust, threshold-dependent measure, yielding a high score only if the model performs well across all four confusion matrix elements relative to class sizes [18,19].

PPV and NPV should be considered when clinical decision thresholds and disease prevalence are relevant, such as in screening or triage applications [20].

In multiclass setting, one should specify the averaging method used and align it with the task [16]. Macro-averaging is suited for imbalanced data where all classes are equally important. Micro-averaging captures overall performance but may hide minority class deficiencies. Weighted-averaging reflects real-world class distribution. Reporting per-class metrics and confusion matrices is also essential for transparency in multiclass setting.

Regression metrics

Fundamentals

Regression analysis is used in machine learning to predict a

Table 1
Overview of the common classification metrics.

Category	Metric	Definition and Characteristics	Formula
Confusion matrix-based (fixed decision threshold)	Accuracy	Proportion of correct predictions	(TP + TN) / (TP + TN + FP + FN)
	Sensitivity (Recall)	True positive rate	TP / (TP + FN)
	Specificity	True negative rate	TN / (TN + FP)
	Balanced accuracy	Mean of sensitivity and specificity; useful for imbalanced datasets	(Sensitivity + Specificity) / 2 = (TP / (TP + FN) + TN / (TN + FP)) / 2
	Positive predictive value (PPV; Precision)	Probability of disease given a positive test	TP / (TP + FP)
	Negative predictive value (NPV)	Probability of no disease given a negative test	TN / (TN + FN)
	F1-score	Harmonic mean of precision and recall; balances FP and FN	(2 × TP / (2 × TP + FP + FN))
	Matthews correlation coefficient (MCC)	Balanced measure of binary classifications; considers all four confusion matrix elements (TP, TN, FP, FN)	((TP × TN – FP × FN) / √[(TP + FP) × (TP + FN) × (TN + FP) × (TN + FN)])
Multi-threshold	AUROC	Area under the receiver operating characteristic curve; threshold-independent; measures overall discriminative ability	See formula in[17]
	AUPRC	Area under the precision-recall curve; threshold-independent; emphasizes model's performance on positive class	See formula in[17]

TP, true positive; FP, false positive; TN, true negative; FN, false negative

continuous target variable from a set of predictors. Unlike binary classification, where the target is limited to discrete categories (e.g., benign vs. malignant), continuous regression output values can span a wide range. However, the selection of an appropriate evaluation metric is not straightforward because no single metric universally captures all aspects of model performance, particularly in contexts with significant clinical implications.

Commonly employed metrics include mean absolute error (MAE) [21], root mean squared error (RMSE) [22], derived from mean squared error (MSE), R-squared [23], and mean absolute percentage error (MAPE) [24]. Definitions and characteristics are presented in Table 2.

Pitfalls and misapplications

Understanding the pitfalls of common regression metrics is crucial, especially in clinical contexts, where misinterpretation can have significant consequences.

MAE can oversimplify error assessment by treating all errors uniformly, potentially underestimating the clinical relevance of large deviations in critical diagnostic settings (**Fig. 3**). For example, consider Patient A (actual: 6, prediction: 1) and Patient B (actual: 36, prediction: 31), both of whom have an MAE of 5 months. However, a 5-month error for Patient A, whose actual progression free survival is only 6 months, represents a gross misestimation with potentially severe clinical consequences (e.g., inappropriate treatment de-escalation). In contrast, the same 5-month error for Patient B, with an actual progression-free survival (PFS) of 36 months, might be clinically less impactful in relative terms. MAE treats both these 5-month errors equally, despite their vastly different clinical implications.

MSE calculates the average of squared differences between predicted and actual values, giving more weight to larger errors. RMSE, the square root of MSE, expresses the error in the same unit as the target, making it more interpretable in practical terms. However, RMSE emphasizes large errors and becomes vulnerable to outliers, such as data entry errors or patients with atypical disease progression, which can skew model evaluation [21]. For example (**Fig. 3**), for Patient C (actual: 12, prediction: 3), the model is off by 9 months. The squared error for this single patient is a substantial 81, heavily penalizing this mistake and significantly impacting the overall RMSE. This heightened sensitivity to large clinical errors can be desirable in high-stakes decisions, where such deviations are unacceptable.

MAPE becomes unstable when the true target value is close to zero and is undefined when the true value is exactly zero (as illustrated in the left panel of Fig. 4, where MAPE explodes as actual PFS approaches 0). For instance, a model predicting 6 months when the true PFS is 3 months results in a 100 % prediction error. However, MAPE is undefined when the true PFS is 0 and becomes unstable when the true PFS is very low, which may occur in patients with highly aggressive disease. MAPE can also be biased because the actual value is in the denominator, it tends to penalize under-predictions (prediction < actual) more heavily than over-predictions (prediction > actual) of the same absolute magnitude.

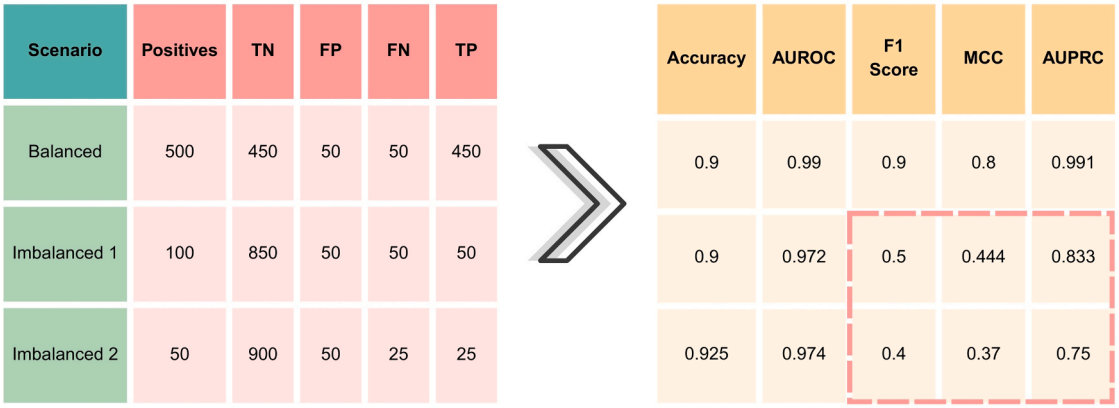


Fig. 1. Impact of class imbalance on classification metrics. While accuracy and area under the receiver operating characteristic curve (AUROC) remain high across scenarios, increasing imbalance leads to sharp declines in F1 Score, Matthews correlation coefficient (MCC), and area under the precision-recall curve (AUPRC), revealing deteriorated performance on the positive class. This highlights the limitations of AUROC and accuracy in imbalanced settings, where metrics focused on the minority class provide a more informative evaluation. Please note that the AUROC and AUPRC values shown are illustrative and not derived from the confusion matrix alone. TN, true negative; FP, false positive; FN, false negative; TP, true positive.

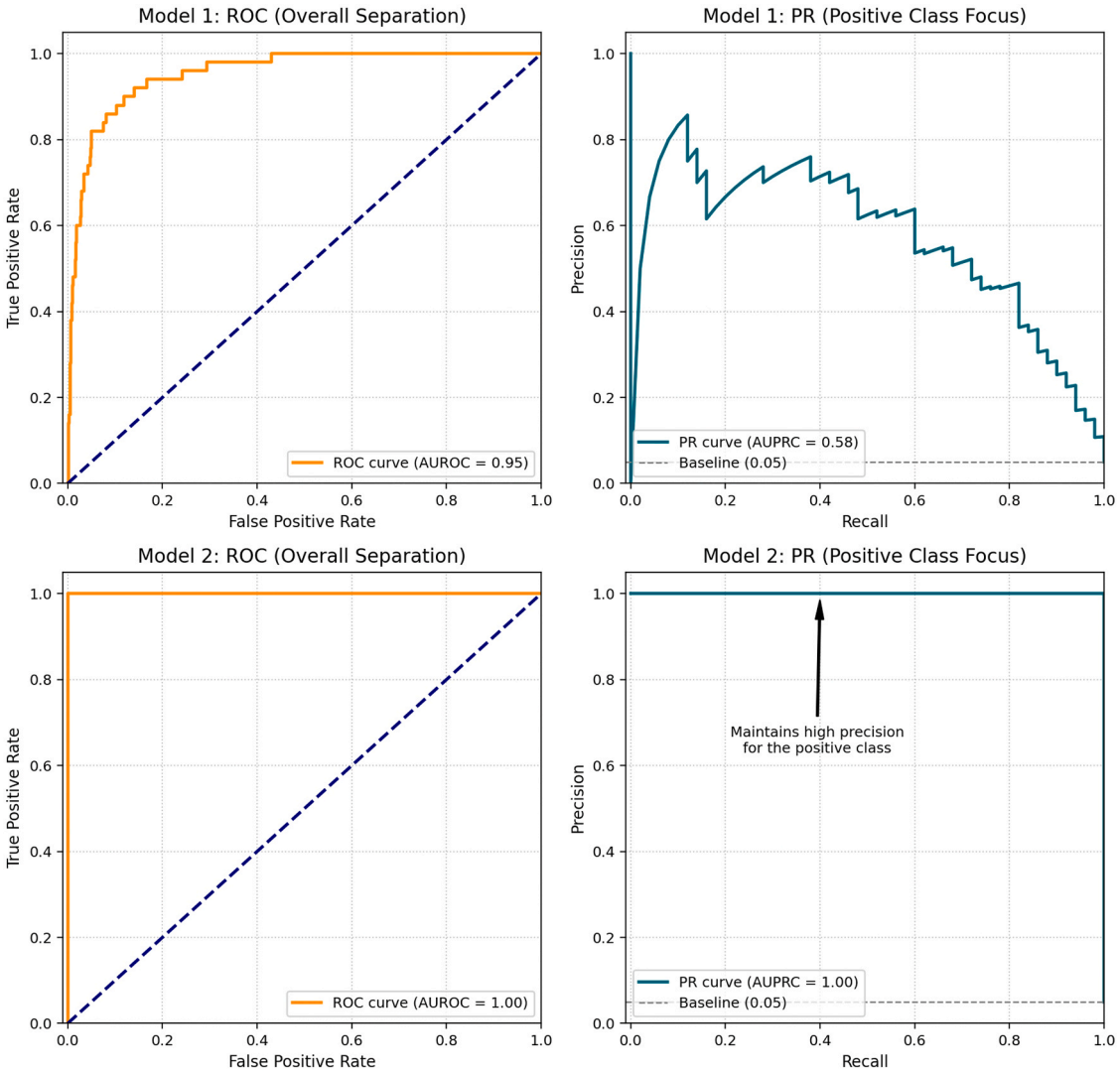


Fig. 2. Comparison of receiver operating characteristic (ROC) and precision-recall (PR) curves for two models under class imbalance. Model 1 shows high AUROC (0.95) indicating strong overall separation, but lower AUPRC (0.58) due to reduced precision on rare positives. Model 2, an ideal (and unrealistic) classifier, achieves perfect AUROC and AUPRC (1.00), consistently identifying positives with no false positives. PR curve baseline reflects the prevalence of the positive class (0.05), while the ROC diagonal indicates random performance.

Table 2
Overview of common regression metrics.

Metric	Definition and Characteristics
Mean absolute error (MAE)	Average of absolute differences between predictions and true values
Root mean squared error (RMSE)	Square root of the mean squared differences
Mean absolute percentage error (MAPE)	Average of absolute percentage errors
R-squared (R^2)	Proportion of variance in the target explained by the model; indicates goodness of fit (proportion of variance explained, 0 −1)

R-squared might appear high even if predictions are clinically inadequate, especially when the dataset contains a narrow range of target values (as depicted in the right panel of Fig. 4). Conversely, extreme outliers can substantially reduce R-squared, even when most predictions are accurate. R-squared can also be inflated by including many irrelevant features, which may lead to overfitting. Importantly, R-squared does not indicate whether the model's predictions are biased (i.e., systematically too high or too low).

Recommendations

Given the distinct sensitivities and limitations of each metric, relying on a single metric can be misleading. A comprehensive evaluation typically involves a combination of metrics interpreted within a specific clinical or research context. The following recommendations address the specific pitfalls discussed.

To address MAE's insensitivity to error magnitude and its potential to overlook critical large errors, it is recommended to combine MAE with clinically weighted error metrics or integrate domain-specific thresholds that reflect real-world implications [25].

To mitigate RMSE's disproportionate sensitivity of RMSE to outliers, which can distort the perceived model performance, robust preprocessing methods, such as outlier detection and removal, or employing robust variants of RMSE, such as Huber loss [26], are advisable.

To counteract MAPE's instability with low or zero true values and its asymmetric penalization, modified metrics like symmetric MAPE (sMAPE) or mean absolute scaled error (MASE), which handle zero and near-zero values more robustly, are recommended [27].

For R-squared, especially given its potential for misinterpretation with narrow data ranges or inflation due to many predictors, using

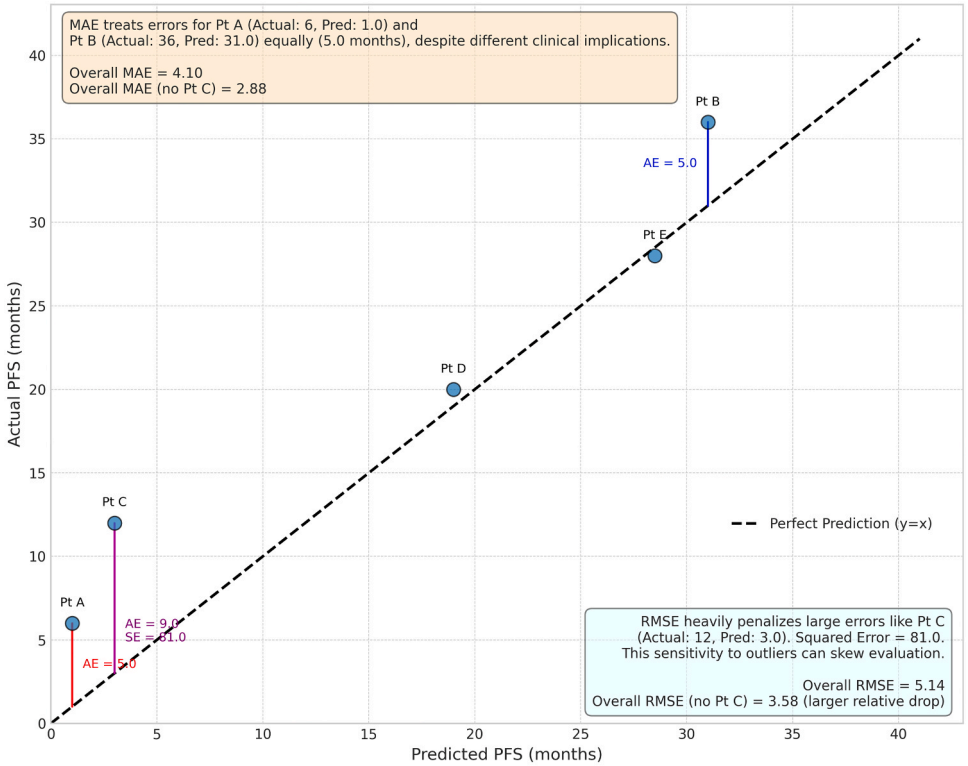


Fig. 3. Pitfalls of mean absolute error (MAE) and root mean squared error (RMSE) in progression-free survival (PFS) prediction as an example use-case. The scatter plot illustrates predicted vs. actual PFS for five patients (Pt). Although MAE treats all absolute errors equally, it overlooks their differing clinical significance. In contrast, RMSE disproportionately penalizes larger errors, as seen in Patient C, increasing the overall error estimates owing to its squared component. AE, absolute error; SE, squared error.

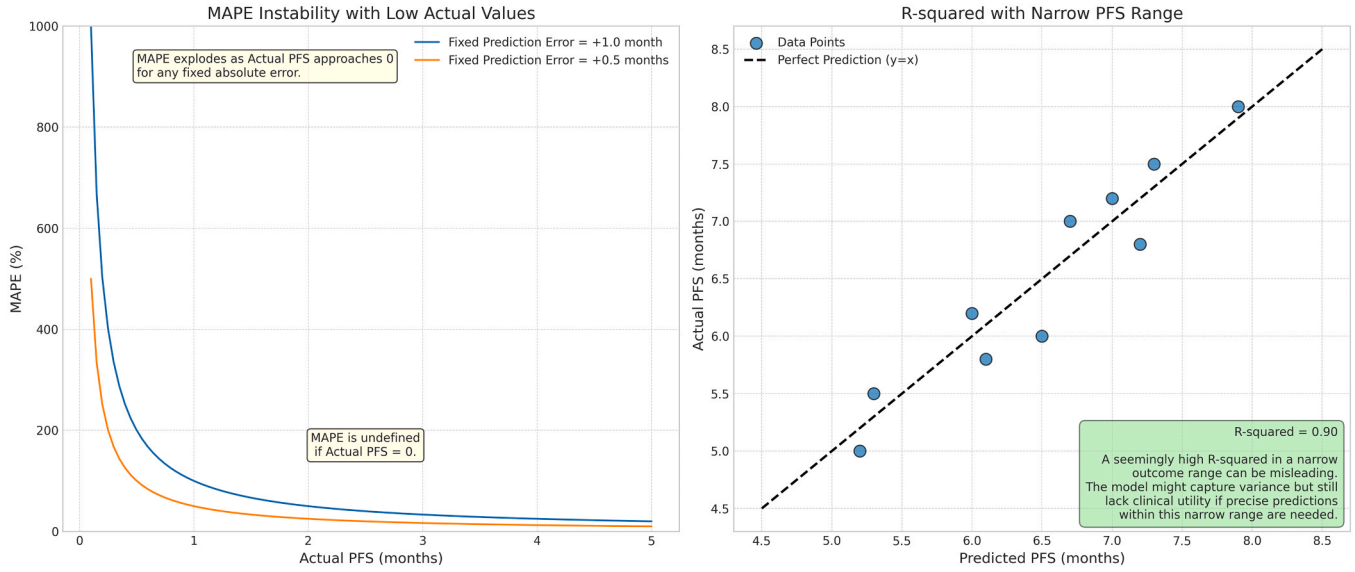


Fig. 4. Limitations of mean absolute percentage error (MAPE) and R-squared in progression-free survival (PFS) prediction. Left: MAPE becomes unstable and undefined as actual PFS values approach zero, exaggerating error estimates for fixed absolute deviations. Right: R-squared can appear high even in clinically unhelpful models when the outcome range is narrow, potentially masking poor predictive precision within a limited clinical context.

adjusted R-squared, which accounts for the number of predictors, alongside additional metrics such as RMSE or MAE, is recommended to obtain a more holistic view of model performance [28]. Visual inspection of predicted versus actual plots is also vital for detecting biases or non-linearities not captured by R-squared alone.

Survival metrics

Fundamentals

Survival analysis refers to statistical modeling of time-to-event data, typically focused on binary outcomes such as death or disease recurrence [29]. In medical imaging, AI holds promise for individualized

Table 3
Key concepts and metrics in survival analysis.

Key Concept/Metric	Definition and Characteristics
Censoring	Event not observed within study time; includes right, left, or interval censoring.
Time-independent output	A single risk score per subject, regardless of time; used with static models; does not capture temporal changes.
Time-dependent output	Risk predictions vary over time (e.g., survival curves, hazard functions).
Harrell's C-index	Discrimination metric estimating concordance between predicted and observed outcomes; valid for time-independent predictions.
Uno's C-index	Variant of C-index using inverse probability weighting; preferred when censoring is substantial or uneven.
Time-dependent AUC	AUC computed at specific time points for dynamic predictions.
Calibration plot	Graphical comparison of predicted vs observed risks.
Hosmer–Lemeshow Test	Statistical test for calibration across risk groups.
Brier score	Mean squared difference between predicted probabilities and actual outcomes; combines calibration and discrimination.
Integrated Brier score	Brier score summarized over time; lower values indicate better performance across time horizon.
DeLong test	Statistical test comparing AUCs from two models; not valid for time-varying AUCs.
Net reclassification improvement (NRI)	Measures correct reclassification of risk categories by a new model.
Integrated discrimination improvement (IDI)	Quantifies gain in sensitivity minus 1-specificity.

AUC, area under the curve

prognostication, but robust evaluation of survival models is critical before clinical use. Evaluation focuses on discrimination (how well the model distinguishes between outcomes) and calibration (how well predicted probabilities reflect actual outcomes). Key concepts and metrics and their definitions are provided in Table 3.

For discrimination, Harrell's C-index applies to time-independent risk scores, while time-dependent area under the curve (AUC) is suitable for dynamic predictions [30].

Calibration is commonly assessed via calibration plots, which compare predicted and observed risks visually. The Brier score provides a unified measure that reflects both calibration and discrimination.

To compare models, tests such as the DeLong test (for AUCs), net reclassification improvement (NRI), and integrated discrimination improvement (IDI) are used to evaluate performance improvements when adding new predictors.

Pitfalls and misapplications

A major pitfall of the C-index lies in its multiple definitions and implementations. Harrell's C-index, though the most commonly adopted, is sensitive to censoring and excludes non-comparable patient pairs, leading to biased estimates in heavily censored datasets (Fig. 5). Different C statistics capture distinct aspects of discrimination and those should not be interchanged or compared without careful consideration [31].

Time-dependent AUC, while better suited for dynamic predictions, also suffers from ambiguities in defining event (e.g., death) and no event groups (e.g., survival) at specific follow-up time. Multiple definitions, such as cumulative/dynamic (C/D), incident/dynamic (I/D), and incident/static (I/S), can yield diverging results for the same data, complicating interpretation and reproducibility (Fig. 6) [32].

Calibration plots, while intuitive, are highly sensitive to sample size and risk group stratification. The Hosmer-Lemeshow test demonstrates limited sensitivity to calibration issues, tending to yield non-significant results in small samples and overly significant results in large cohorts (Fig. 7) [33].

The Brier score, although widely used, depends on prevalence, which can lead to misleading rankings of models across datasets with different event rates. This makes direct comparisons challenging, unless additional decision-analytic measures (e.g., net benefit) are used to evaluate model's clinical value [34].

The DeLong test assumes independence between samples (i.e., Subject 1 should be independent of Subject 2, etc.), as violations can lead to incorrect variance estimates and p-values. It also requires model-level

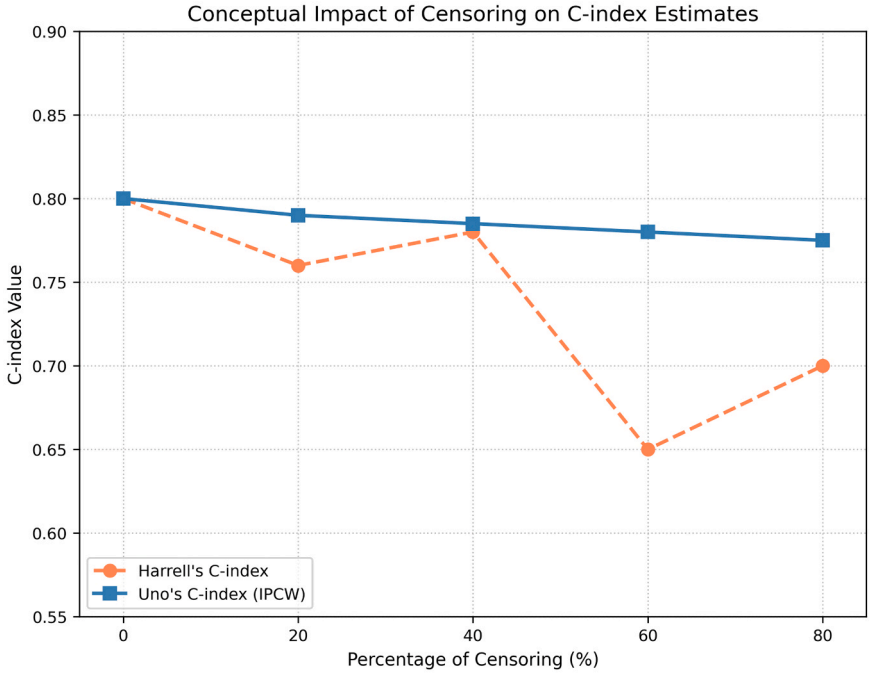


Fig. 5. Conceptual illustration of how increasing censoring affects C-index (concordance Index) estimates. As censoring increases, Harrell's C-index shows a pronounced decline, reflecting its sensitivity to incomplete outcome data. In contrast, Uno's C-index remains relatively stable due to its adjustment for censoring using inverse probability weighting (IPCW).

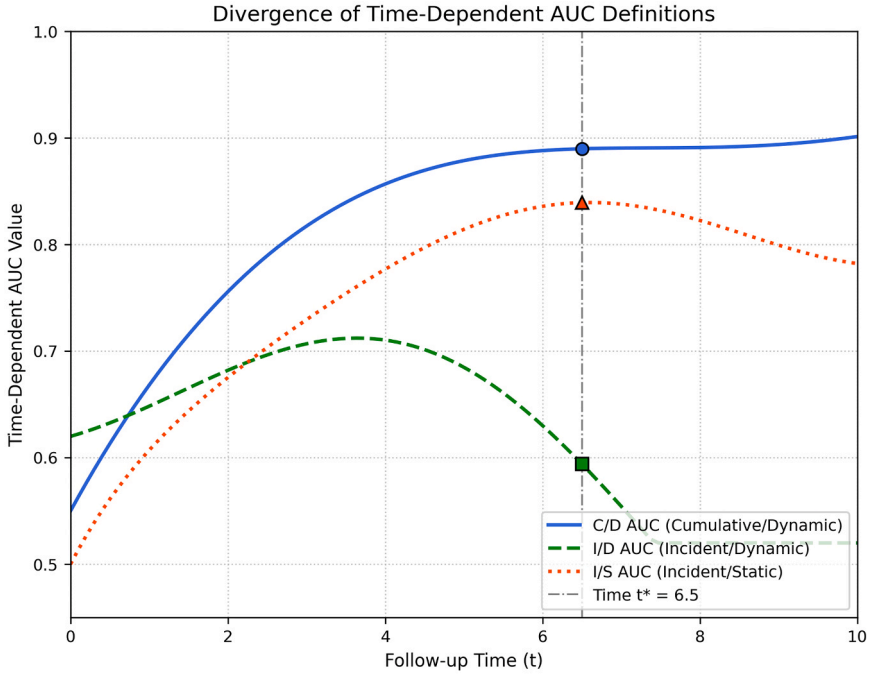


Fig. 6. This conceptual plot illustrates how three definitions of time-dependent area under the curve (AUC), cumulative/dynamic (C/D), incident/dynamic (I/D), and incident/static (I/S), vary across follow-up time. At time t* = 6.5, each AUC type yields a different value, reflecting differing assumptions about risk sets and event handling. These distinctions underscore the importance of selecting and reporting an AUC definition aligned with the prediction target and clinical context.

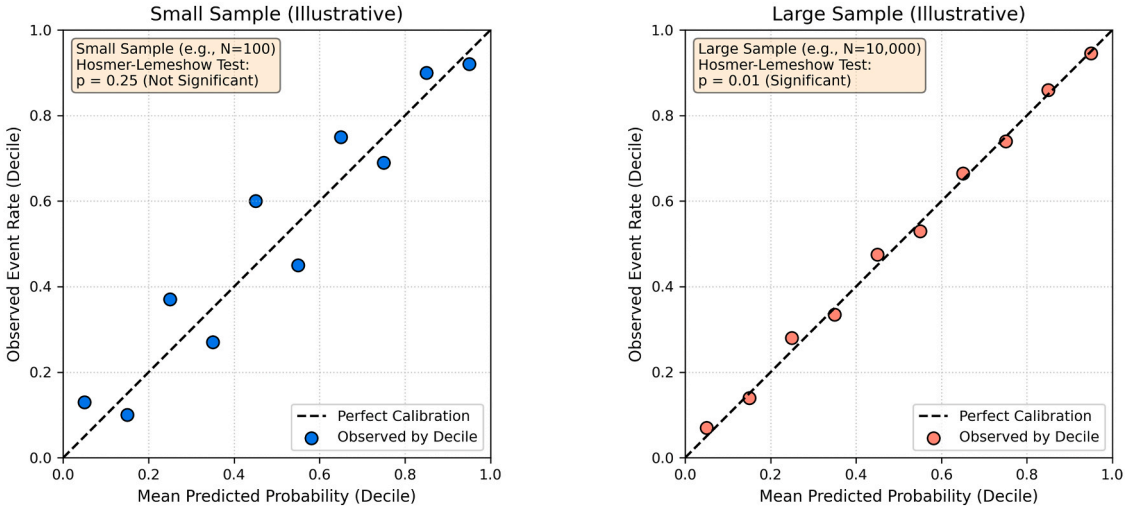


Fig. 7. Hosmer–Lemeshow test sensitivity to sample size. Calibration plots for small (N = 100) and large (N = 10,000) samples. Despite visible miscalibration, the small sample yields a non-significant test (p = 0.25), while the large sample shows near-perfect calibration but a significant result (p = 0.01), highlighting the test's sensitivity to sample size. Note: A low p-value indicates significant miscalibration.

pairing (i.e., Model A's prediction for Subject 1 is paired with Model B's prediction for Subject 1) to ensure a fair comparison by evaluating both models on the exact same set of predictions. Also, this test is not directly applicable to time-dependent AUCs because it is designed for a single, static AUC value rather than an AUC that changes over time. Similarly, the use of NRI and IDI in survival settings requires careful methodological adaptation, particularly when censoring is present [35], because censoring means event times are not observed for all subjects, complicating the classification of improvement or worsening of risk.

Recommendations

Primary evaluation metrics and their specific variants (e.g., which C-index, which time-dependent AUC definition) should be pre-specified prior to analysis. This selection should be justified based on the type of model output, censoring patterns, and the intended clinical application. All chosen metrics should be reported transparently, and their interpretation must be contextualized within their statistical limitations [36].

When using time-independent risk scores, Harrell's C-index remains an appropriate measure, but it requires cautious interpretation in datasets with heavy censoring. In such scenarios, Uno's C-index is recommended due to its enhanced robustness to censoring patterns **(**Fig. 5**)** [30].

For time-dependent AUCs, researchers must explicitly state the chosen variant (C/D, I/D, or I/S), specify the time point(s) of evaluation,

Table 4
Overview of common detection metrics.

Metric	Definition and Characteristics
Precision	TP / (TP + FP); measures correctness of detected positives; high precision = few false positives
Recall (Sensitivity)	TP / (TP + FN); measures completeness; high recall = few missed detections
Average precision (AP)	Interpolated average of precision values across different recall levels; summarizes detection performance at various thresholds; localization evaluated via IoU; different methods for "interpolated average" (e.g., 11-point interpolation, or all-points interpolation which is equivalent to the area under the P-R curve); a way to summarize the precision-recall curve into a single number.
Mean average precision (mAP)	Mean of AP across all object classes or IoU thresholds; reflects both precision and localization quality; can be calculated at a single IoU threshold (e.g., mAP@0.5) or as an average over multiple IoU thresholds (e.g., mAP@[.5:.05:.95])
Intersection over union (IoU)	Area of overlap / area of union between predicted and true boxes; common criterion for determining correct localization; typically > 0.5 indicates TP.
Free-response ROC (FROC)	Sensitivity vs. average FP per image or patient; suitable for multi-lesion tasks; more clinically intuitive for tasks where there can be multiple findings and the negative space is vast.

TP, true positive; FP, false positive; FN, false negative

and provide a clear rationale for these choices. This rationale should be grounded in the clinical setting and the specific prediction target, given that interpretations can vary significantly depending on the definition employed **(**Fig. 6**)** [32].

Calibration assessment should prioritize visual inspection of calibration plots, used in conjunction with the Brier score (or the integrated Brier score for longitudinal assessments). Researchers must exercise caution and avoid over-reliance on the Hosmer-Lemeshow test p-value, given its established sensitivity to sample size and binning strategy **(**Fig. 7**)**. If this test is utilized, its limitations must be explicitly stated. The Brier score's dependence on outcome prevalence should also be acknowledged, and its interpretability enhanced by contextualizing results with domain knowledge [34].

Beyond statistical performance metrics, the clinical utility of the model should be actively assessed. Decision-analytic measures, such as decision curve analysis for estimating net benefit, provide a valuable framework for this purpose. Net benefit offers a more clinically relevant interpretation of a model's value, especially in comparative model evaluations.

When comparing models, the DeLong test is not suitable for time-dependent AUCs, requiring different methods [37]. While NRI and IDI may serve as alternatives, their application requires meticulous handling of censoring and circumspect interpretation, acknowledging their respective limitations [35].

Detection metrics

Fundamentals

Evaluating the performance of AI models for detection tasks in medical imaging, such as identifying lesions or anatomical structures, requires specialized metrics [38–40]. Unlike simpler classification tasks, detection involves not only determining the presence of an object but also its location, often indicated by bounding boxes.

The foundational elements for many detection metrics are based on the confusion matrix concept [9,41]. In object detection, these are typically determined by comparing predicted bounding boxes to ground truth boxes. A prediction is usually considered a true positive if its intersection over union (IoU) with a ground truth box exceeds a predefined threshold (e.g., 0.5) and it is correctly matched to that ground truth. A prediction is a false positive if its IoU with all ground truths is below the threshold, or if it is a duplicate detection of an already matched ground truth. A false negative occurs when a ground truth object is not detected by any prediction with sufficient IoU.

A key characteristic of pure object detection problems is the frequent absence of true negatives, as areas without targets are not explicitly labeled as such [9,41]. Including true negatives in the evaluation can artificially inflate metrics like accuracy (which includes true negatives) and reduce sensitivity to errors in detecting the often much smaller foreground objects. This is why metrics like accuracy are generally not suitable for object detection tasks, and why precision-recall based metrics (like mean average precision (mAP)) are preferred. Common metrics used in detection evaluation are presented in Table 4.

Pitfalls and misapplications

Evaluating detection models in medical imaging faces several challenges. There is significant heterogeneity and a lack of standardized approaches in metric selection and challenge design, making objective comparisons across studies difficult [9,10,42]. The complex properties and limitations of various metrics are often not fully understood by

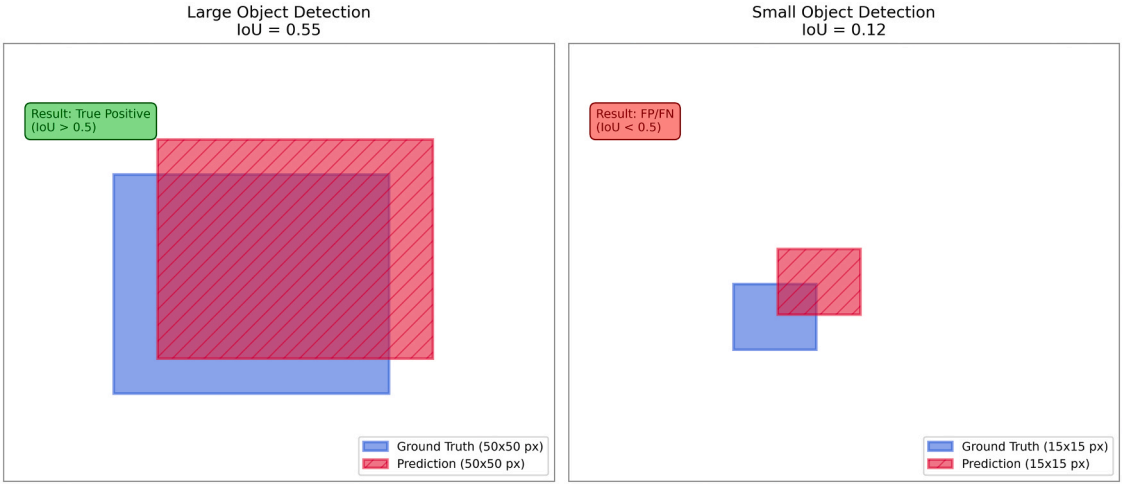


Fig. 8. Illustration of how object size influences intersection over union (IoU)–based evaluation. For large objects, overlapping predictions often exceed the IoU threshold (≥0.5), resulting in a true positive. In contrast, the same degree of misalignment in small objects may reduce the IoU below the threshold, leading to false positive or false negative classification. IoU, Intersection over Union; FP, False Positive; FN, False Negative.

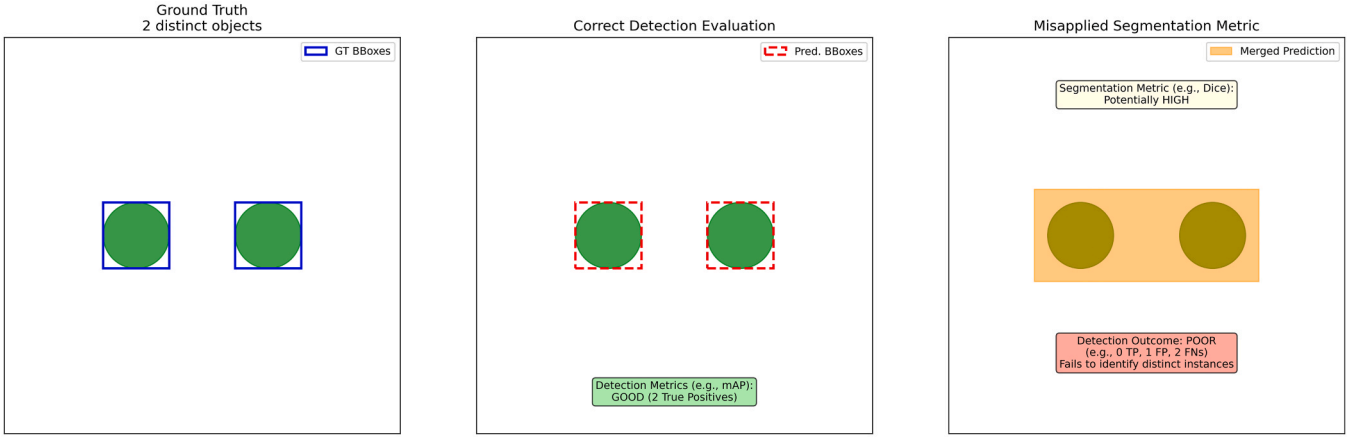


Fig. 9. Example of metric mismatch in object evaluation. While detection metrics (e.g., mAP) correctly identify two distinct objects, segmentation metrics (e.g., Dice) may misleadingly report high performance despite merging instances into one region. This highlights the risk of using segmentation metrics for instance detection tasks. In the misapplied case (right), the merged prediction fails to uniquely match either ground truth object, resulting in 0 true positives (no prediction sufficiently overlaps a single ground truth), 1 false positive (merged region does not match any object), and 2 false negatives (both objects remain undetected). mAP, mean average precision; Dice, Dice similarity coefficient; TP, true positive; FP, false positive; FN, false negative.

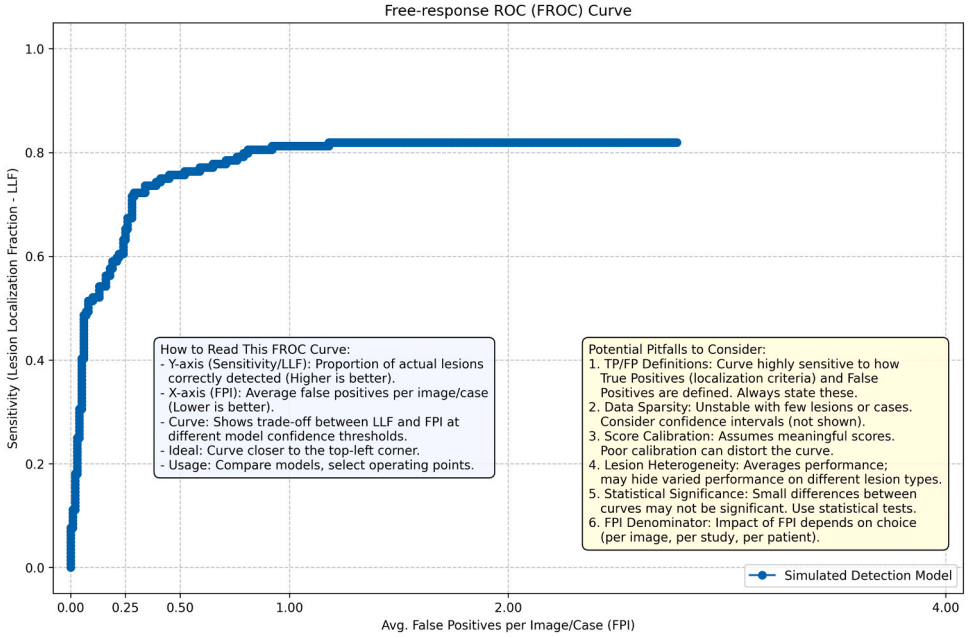


Fig. 10. Free-response receiver operating characteristic (FROC) curve illustrating the performance of a simulated multi-lesion detection model. The curve shows the trade-off between lesion localization sensitivity (lesion localization fraction, LLF) and average false positives per image or case (FPI). While FROC offers valuable insight into detection performance, its interpretation can be confounded by several factors (yellow text box in the figure) such as inconsistent true/false positive definitions, small object sensitivity (e.g., IoU thresholding), and hierarchical data structures. Simply averaging metrics across images without accounting for lesion counts or localization criteria can lead to misleading conclusions.

researchers and clinicians. Furthermore, no single metric can comprehensively capture all desirable aspects of a model's performance in terms of detection. Datasets themselves can present challenges, such as small target structures or high variability in object sizes, which can disproportionately affect certain metrics [9,39]. For example, Fig. 8 illustrates how a constant pixel offset in a bounding box prediction results in a much lower IoU for a small object compared to a large one, potentially leading to the small object detection being misclassified as a false positive despite reasonable localization accuracy.

Several pitfalls can compromise the validity of detection evaluation. One significant pitfall is the inadequate choice of problem category, such as applying metrics designed for segmentation (which focuses on exact boundaries) to object detection (which prioritizes localization of distinct instances) (Fig. 9) [8,9]. As shown in Fig. 9, a segmentation metric like Dice might yield a high score for a prediction that merges two distinct objects, whereas a detection evaluation would correctly identify this as poor performance (e.g., one true positive, one false negative, and potentially a false positive for the merged area, depending on matching rules) because it fails to identify distinct instances.

Poor metric selection can occur by disregarding the specific biomedical domain interest (e.g., not accounting for the clinical impact of different errors), the properties of target structures (like size sensitivity of some metrics), or dataset properties (such as the issue of empty references or predictions causing undefined metric values) [9].

Issues with metric application include inadequate implementation, insufficient reporting of variability, and, crucially for detection,

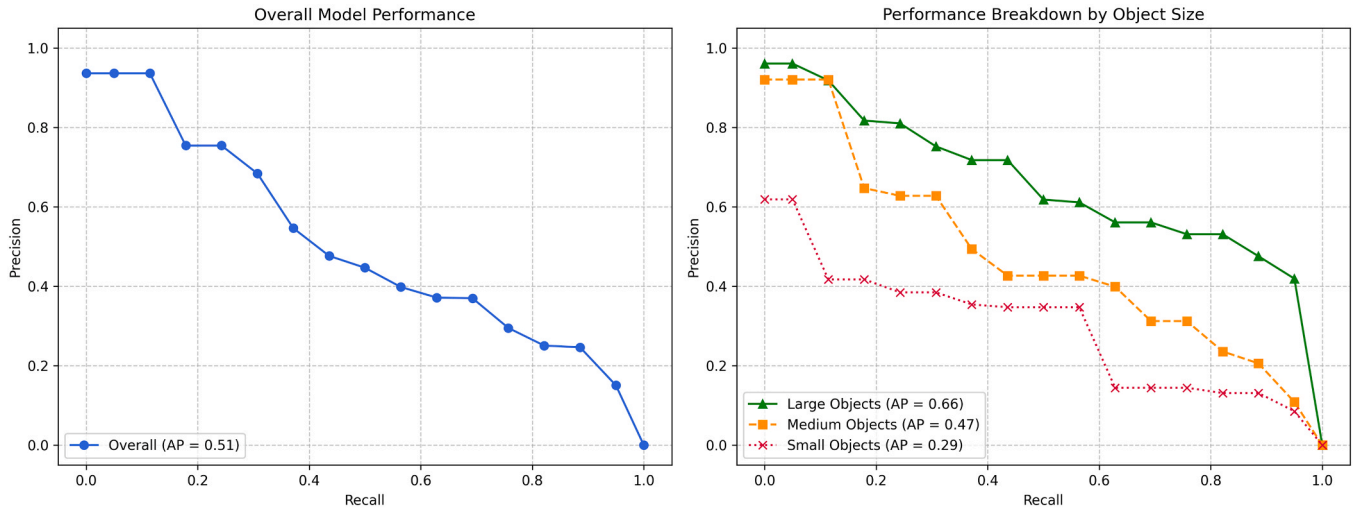


Fig. 11. Importance of stratified evaluation for understanding model limitations. Precision-recall (PR) curves illustrating overall model performance and a breakdown by object size. While overall performance provides a general overview, subclass analysis reveals substantial performance disparities. AP, average precision.

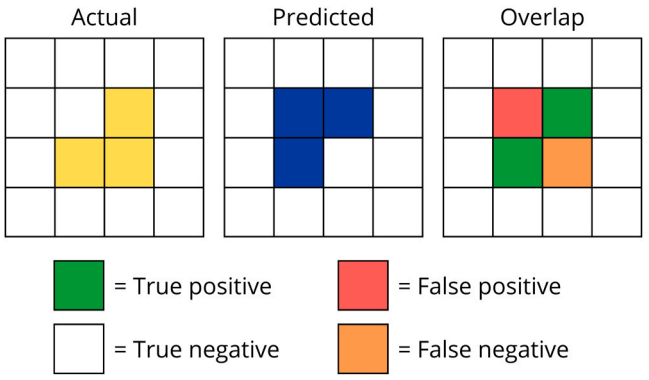


Fig. 12. Example of a semantic segmentation task on a 4 × 4 pixel matrix, with corresponding pixel-level definition of true and false positive and true and false negative.

inadequate aggregation of scores [8,9]. Simply averaging metrics across images without considering the number of objects or hierarchical data structures can be misleading.

Localization criteria themselves, like IoU, can also have pitfalls, particularly with small structures [9,39]. For example, the Free-Response ROC (FROC) curve, a common metric for evaluating multi-lesion detection tasks, illustrates this complexity. As shown in Fig. 10, while FROC provides valuable insights into the trade-off between sensitivity and false positives per image/case, its interpretation requires careful consideration of several potential pitfalls (please see them in the figure) related to its definition and application.

Recommendations

It is crucial to use multiple, complementary metrics rather than relying on a single one, selecting them based on their suitability for the specific problem, dataset, and clinical goals [38,43].

Leveraging structured frameworks like Metrics Reloaded, which uses a 'problem fingerprint' to guide metric selection, can help avoid common pitfalls and recommend metrics that align with domain interest [8]. Metrics should be applied properly, including aggregating results appropriately, often per class or instance where applicable, and accounting for data hierarchies [8,9].

Reporting performance variability (e.g., confidence intervals) and using informative visualizations are also essential for robust evaluation [8,10]. Calculating metrics on blinded data and utilizing standardized open-source tools can enhance reproducibility and comparability. Evaluating performance on specific subclasses or challenging cases, like small objects, can provide deeper insights into model behavior **(**Fig. 11**)** [39].

Segmentation metrics

Fundamentals

Image segmentation is a common task in AI applications within medical imaging. Even when it is not the main endpoint of a research study or software tool, automated segmentation of one or more structures within the image represents a common pre-requisite for further analyses. Therefore, the need for clear quantitative metrics to assess the reliability of automated segmentation models is paramount to ensure the robustness of the entire software stack which depends on this output [44] and ultimately the trustworthiness of downstream analyses that can directly impact clinical decisions (e.g., treatment planning based on tumor volume or radiomics modeling).

From a practical perspective, the easiest way to understand semantic (i.e., pixel/voxel level) image segmentation accuracy metrics is to envision it as a direct transposition of classification metrics from a patient or exam level to the pixel one. Imagine each pixel as an individual data point. The actual or ground truth mask indicates the true class of each pixel (e.g., background or lesion), often labeled with "0" and "1" values on a binary mask. At time of inference, the automated segmentation tool's output will in turn classify each pixel as belonging to one class or the other. After this process, matching labels will be true positives (correctly identified lesion pixels) and true negatives (correctly identified background pixels), while mismatches will produce false positive (background pixels incorrectly identified as lesion) and false negative (lesion pixels missed by the model) **(**Fig. 12**)**. Notably, true negatives are typically excluded from the common overlap-based metric formulas discussed below. In medical imaging, the background (true negative region) often constitutes the vast majority of pixels. Including true negatives would make these metrics artificially high and less sensitive to errors in segmenting the usually much smaller foreground regions of interest.

For example, the Dice similarity coefficient (DSC), commonly employed in medical imaging segmentation, corresponds to the F1 score [45]. Conceptually, DSC relates twice the intersection (true positive, correctly identified pixels) to the sum of the total pixels in the ground truth mask (true positive + false negative) and the predicted mask (true

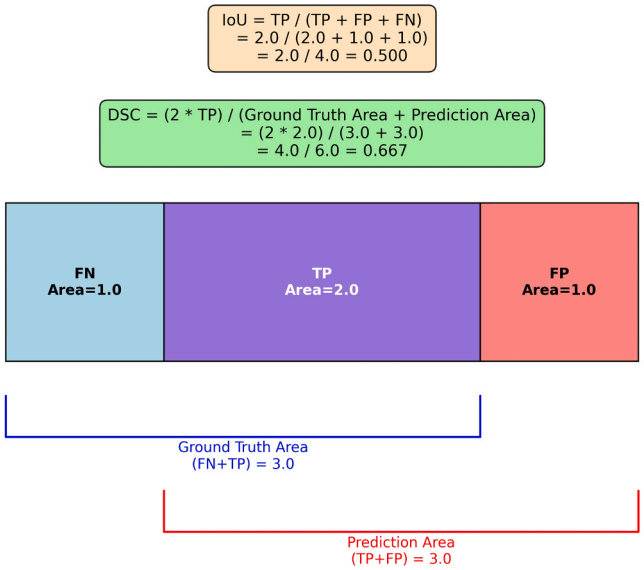


Fig. 13. Conceptual diagram illustrating the components for intersection over union (IoU) and Dice similarity coefficient (DSC) calculations.

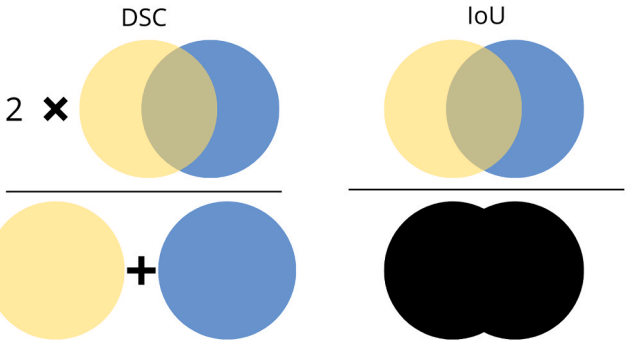


Fig. 14. Visual comparison of the Dice similarity coefficient (DSC) and intersection over union (IoU) formulas. DSC (left) effectively relates twice the intersection (overlap area) to the sum of the individual ground truth (yellow circle) and prediction (blue circle) areas. IoU (right) relates the intersection to their combined union (black shape).

positive + false positive).

An alternative accuracy metric is represented by the intersection over union (IoU), also known as Jaccard Index. IoU measures similarity by dividing the size of the intersection (true positive) by the size of the union of the ground truth and predicted areas (true positive + false positive + false negative). Fig. 13 illustrates their calculation with respective formulas.

Fig. 14 visually contrasts the calculations of DSC and IoU. The metrics are directly related by the formula: DSC = 2 * IoU / (IoU + 1). This means DSC will generally yield higher values than IoU for any imperfect segmentation, with both reaching 1 for a perfect match.

It should be noted that other discrimination accuracy metrics, such as precision and recall, are still applicable to segmentation tasks, and may be included when reporting performance of such tools.

Pitfalls and misapplications

Users should be aware that segmentation metrics, such as DSC and IoU, do not represent perfect estimates of performance to be used blindly. Rather, as is often the case, these are measurement tools to be used within the context of pre-existing domain knowledge and in light of the question that we wish to answer.

DSC tends to value overlap between predicted and ground truth label more than absolute precision, while IoU is a more stringent metric from this perspective. More specifically, IoU penalizes any misclassified pixel within the union area somewhat uniformly. Consequently, for small objects, even minor pixel errors (which represent a larger proportion of the small object) can lead to a significant drop in IoU (Fig. 15). DSC, by normalizing against the sum of areas (effectively averaging their sizes), can be less sensitive to the same absolute error in larger objects compared to smaller ones (Fig. 15).

Recommendations

For segmenting small structures where high boundary precision is critical (e.g., small tumors, subtle lesions), IoU provides a stricter evaluation. For larger structures where overall volumetric agreement is more important than pinpoint boundary accuracy, DSC might be more reflective of perceived quality or less punishing of minor boundary discrepancies.

A second consideration is the overlap-centric nature of DSC and IoU. These metrics primarily quantify regional agreement and may not adequately penalize boundary inaccuracies or differentiate error distributions if overall overlap is similar. Thus, visual inspection of segmentations is essential. For applications demanding high boundary fidelity (e.g., surgical planning, radiomics analysis), supplementing with boundary-specific metrics like Hausdorff distance or average surface distance is also recommended.

Segmentation metric interpretation is context-dependent. A good score varies with anatomical complexity, image quality, and inter-observer variability, precluding universal thresholds. For instance, a DSC of 0.7 might be excellent for a highly complex and variable

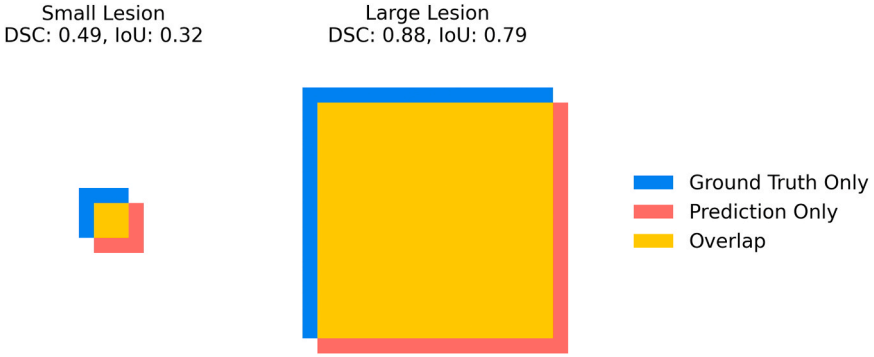
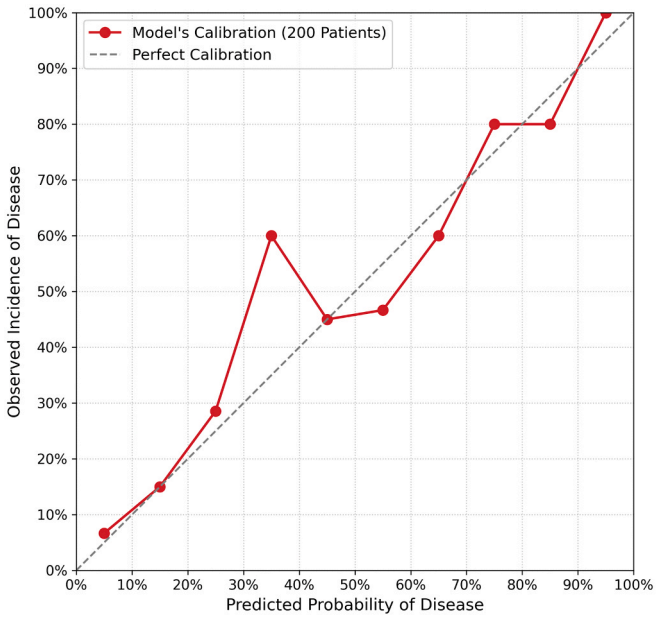


Fig. 15. Effect of lesion size on segmentation metric sensitivity under fixed misalignment. Illustration compares small and large lesion segmentations with the same absolute spatial offset (e.g., 3 pixels). Despite equal misalignment, intersection over union (IoU) drops more sharply for the small lesion, while Dice similarity coefficient (DSC) shows a more moderate change. This indicates IoU's heightened sensitivity to minor errors in small structures.



Pred. Prob. Bin	Patients in Bins	Actual Cases	Observed Inc. (%)	Calibration
0-10%	30	2	6.7	Good
10-20%	40	6	15.0	Good
20-30%	35	10	28.6	Good
30-40%	25	15	60.0	Poor
40-50%	20	9	45.0	Good
50-60%	15	7	46.7	Poor
60-70%	10	6	60.0	Good
70-80%	10	8	80.0	Poor
80-90%	10	8	80.0	Good
90-100%	5	5	100.0	Good

Fig. 16. Calibration plot and table for a disease prediction model (n = 200). Calibration assesses how closely predicted probabilities align with actual outcomes, here represented by the observed incidence of disease within each predicted probability bin. The red curve shows model calibration across bins, while the diagonal line represents perfect calibration. The adjacent table summarizes the predicted probability bins, number of patients per bin, actual disease cases, observed incidence (%), and a qualitative assessment of calibration.

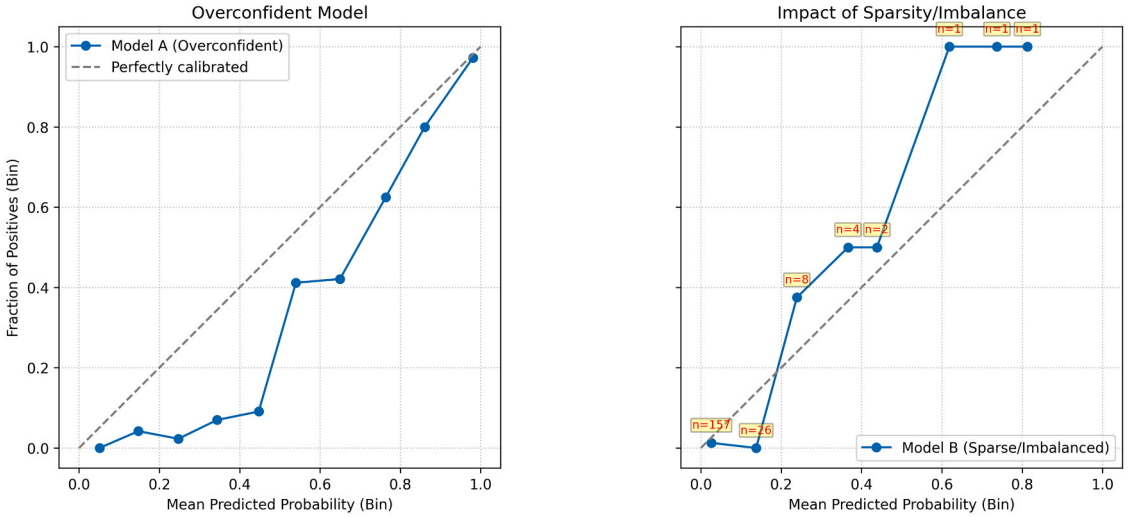


Fig. 17. Reliability diagrams and their pitfalls. Left panel shows an overconfident model (Model A), where the predicted probabilities are systematically higher than the observed frequencies of positives across bins. This results in a curve that falls below the diagonal, indicating poor calibration despite seemingly high confidence, which, if ignored or misinterpreted, can lead to pitfalls in downstream decision-making. Right panel demonstrates the impact of class imbalance and bin sparsity (Model B). Many bins contain very few samples (e.g., n = 1 or n = 2), especially at higher predicted probabilities. This sparsity can cause misleading interpretations of calibration performance due to high variability.

structure but mediocre for a simple, well-defined one. Therefore, scores must be interpreted within the specific application context, establishing task-specific baselines and, ideally, comparing against human inter-rater reliability (of segmentation result or subsequent analysis, e.g., reproducibility of extracted features from segmentation) for a robust assessment.

The choice of primary metric, and the interpretation of its value, should align with the clinical impact of potential segmentation errors. Small differences might be critical for some applications but negligible for others. From a practical perspective, DSC is very common in relation to segmentation while IoU can be found in detection tasks. Being aware of these conventions can aid in comparing with existing work.

Calibration metrics

Fundamentals

Calibration metrics play a critical role in evaluating the trustworthiness of probabilistic predictions in medical imaging AI. Calibration becomes even more crucial when models are deployed in clinical settings where uncertainty quantification is vital, for instance, in triaging ambiguous cases for expert review or combining AI outputs with clinician assessments [46].

While conventional performance metrics such as accuracy or AUROC provide insight into a model's discriminative power, they do not reflect

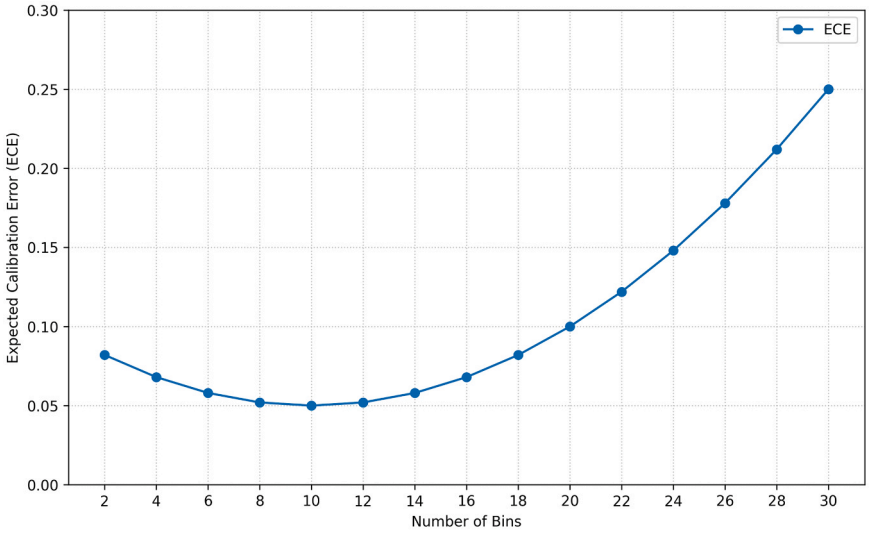


Fig. 18. Expected calibration error (ECE)'s sensitivity to binning strategy. Illustrative plot shows ECE varies with the number of bins used. While initially relatively stable at lower bin counts, ECE becomes increasingly sensitive at higher bin counts, underscoring its dependence on binning strategy.

how well the predicted probabilities align with actual outcomes, which is an essential aspect in high-stakes medical decisions [46]. Calibration assesses whether, for example, predictions labeled with a 70 % probability of disease actually correspond to a 70 % incidence of disease in reality. Reliability diagrams provide a granular visual tool, plotting predicted probabilities against observed frequencies to highlight regions of under- or overconfidence **(**Fig. 16**)**.

A key metric in this domain is the expected calibration error (ECE), which summarizes the average discrepancy between predicted probabilities and observed outcomes across multiple bins [47].

Another important measure is the Brier score, which quantifies the mean squared error between predicted probabilities and actual binary outcomes [48]. It is a strictly proper scoring rule, meaning it encourages honest probability estimates by uniquely rewarding the model for outputting its true estimates about the likelihood of an event.

Various techniques can be applied to improve model calibration, often post hoc (after initial model training). Common methods include parametric approaches like Platt scaling [49], non-parametric methods such as isotonic regression, and Bayesian techniques [50]. These are increasingly being utilized, particularly when initial model outputs exhibit poor calibration.

Pitfalls and misapplications

Interpreting calibration plots, such as reliability diagrams, can be challenging [51]. Visual assessment of deviations from the ideal diagonal line can be subjective, varying between observers [52]. Furthermore, the choice of binning strategy for the plot itself (number and width of bins) can alter its appearance, potentially influencing interpretation [53]. One common pitfall is overconfidence, where predicted probabilities are systematically higher than observed frequencies, resulting in a curve falling below the diagonal, as illustrated by Model A in Fig. 17. This can give a false impression of high confidence despite poor calibration. Overconfidence may be overlooked, especially when its effects are subtle or confounded by binning artifacts. Their utility may also degrade in the presence of class imbalance, very common in medical datasets [54]. Fig. 17 (right panel, Model B) demonstrates this impact where, many bins contain very few samples. This sparsity, often exacerbated by class imbalance, can cause misleading interpretations of calibration performance due to high variability within those sparse bins, rendering the perceived fair calibration unreliable [53].

ECE is intuitive and widely used, but it can obscure local miscalibrations due to its dependence on binning schemes [55]. Because ECE averages errors across all bins, significant miscalibration within a narrow range of probabilities (i.e., in specific bins) might be numerically masked or 'cancelled out' by good calibration in other bins, especially if the overall binning strategy is not granular enough. Fig. 18 graphically illustrates this sensitivity. This issue highlights the risk of drawing misleading conclusions about a model's calibration if the binning strategy is not carefully chosen, justified, and reported.

The Brier score, while comprehensive, blends calibration and discrimination [51], limiting interpretability of the model's calibration quality unless its decomposed components are analyzed. A lower Brier score does not necessarily mean better calibration, as it reflects both discrimination and calibration simultaneously.

Other challenges in calibration for medical imaging AI arise from small sample sizes, distributional shifts, and overconfident neural networks, especially those trained with cross-entropy loss [56–58].

Recommendations

Reporting both ECE and Brier Score (ideally its decomposed components: calibration, refinement/resolution, and uncertainty) alongside visual tools like reliability diagrams is recommended.

The binning strategy (number of bins and how they are defined) used for ECE and reliability diagrams should always be specified to ensure transparency and reproducibility.

As Brier score combines calibration with discrimination, extensions such as the decomposed Brier score can be considered to separate these components, though this is less frequently reported in medical imaging literature.

Given the propensity for neural networks, particularly those trained with cross-entropy loss, to be overconfident, applying post hoc recalibration methods is often needed before model deployment in clinical practice.

Calibration should be rigorously evaluated not only on internal test sets but also on diverse, representative external datasets. This helps assess robustness against potential distributional shifts that models are likely to encounter in real-world clinical scenarios.

Foundation model metrics for medical imaging-related tasks

Fundamentals

Foundation models (FMs) are deep learning models trained on massive datasets that can flexibly handle various downstream tasks (e.

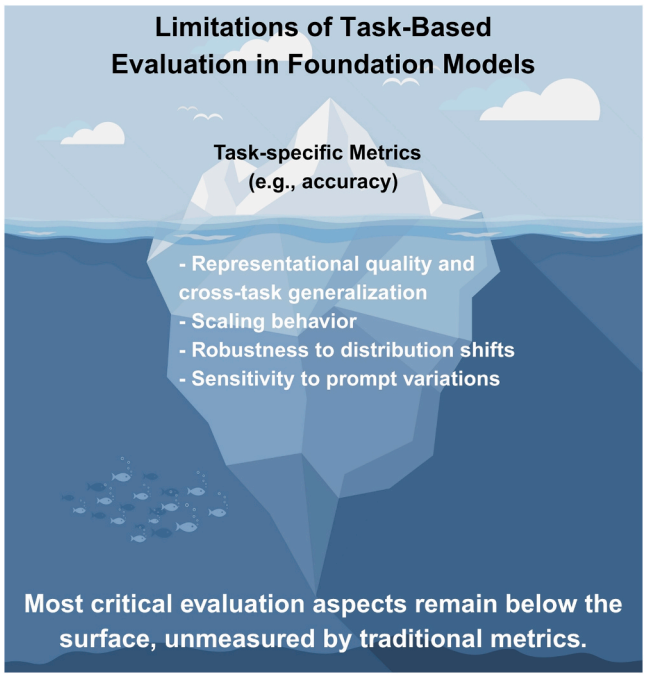


Fig. 19. Iceberg illusion in foundation model evaluation. Traditional task-based metrics may capture only a small visible portion of model performance. Critical aspects remain unmeasured and require more comprehensive evaluation frameworks.

g., classification, segmentation) with minimal task-specific training (i.e., continued training on labeled data) [11,59]. Their evaluation introduces additional complexities compared to traditional AI models, owing to their scale, emergent properties, and multi-task adaptability. This section focuses on image-based and multimodal FMs; text-only FMs (LLMs) are discussed in the next section.

FMs are typically pretrained using weakly or self-supervised learning, enabling them to learn generalizable representations without human-provided labels. These rich representations support performance across diverse downstream tasks. Their large scale can lead to "emergent" abilities, such as instruction following [60] or zero-/few-shot generalization [61]. General-purpose FMs can be fine-tuned for medical domains (e.g., Segment Anything Model [62,63]) and may be uni-modal [62,64–66] or multimodal, combining images and associated reports [67–71].

FMs are increasingly being explored for diverse medical imaging applications, such as disease classification from radiographs [69–71], segmentation of organs or lesions in MRI or CT [68,72], and generational tasks like medical report generation or free-text visual question answering [71]. Each task requires appropriate performance metrics (e.g., accuracy for classification, IoU for localization). However, FM-specific attributes, scale, multimodality, zero/few-shot transferability, and adaptability, necessitate broader evaluations beyond task-specific metrics, including tests for generalization, representation quality, and data efficiency.

Emerging benchmarks increasingly cover multiple diverse task types for the evaluation of a single FM (e.g., segmentation, classification, outcome prediction). Comprehensive frameworks include the assessment of bias, fairness, toxicity, and robustness [73,74]. Such multi-dimensional evaluation frameworks are especially important for clinical AI deployment where safety, bias, and interpretability are crucial.

The high computational cost of FM training makes studying scaling laws essential [75]. For medical applications, which are often compute- or data-constrained, adapting FMs is interesting due to their potential for greater label efficiency (less labeled training data is needed for specific tasks) and computational efficiency (fewer training steps are needed for adaptation). These can be evaluated by tracking performance (e.g., accuracy) across decreasing labeled datasets or training steps. Plotting performance against data, steps, or parameter count then helps visualize scaling behavior. Additionally, ablation studies can isolate the contribution of specific components (e.g., attention blocks or multi-modal inputs).

FM adaptability stems from the quality of their learned representations. Zero-shot evaluation tests performance on unseen tasks. For instance, an FM may identify a rare tumor not present in the training data based on general visual features that favor malignancy. Few-shot evaluation assesses performance using a small labeled set, via (i) similarity-based matching, (ii) prompting (without fine-tuning), or (iii) minimal fine-tuning. In promptable FMs, the choice, type (e.g., text, points, bounding boxes), and design of prompts can substantially affect results [62,64,76].

Linear probing evaluates the usefulness of FM features for a specific task (e.g., classification) by training a simple linear classifier that uses the FM's features as input and labeled data as targets [77].

Retrieval-based evaluation assesses how well FM representations capture concepts by using them to find semantically similar items, e.g., identifying CT studies with the same findings, using metrics such as cosine similarity or CLIPScore [78]. In multimodal settings, retrieval operates across modalities, for instance to assess image-text alignment, a critical capability for image-report linking in radiology.

Pitfalls and misapplications

FM evaluation in medical imaging poses specific challenges. Key pitfalls include over-reliance on task-specific metrics (Fig. 19), which miss generalization and alignment capabilities and sensitivity of few-/zero-shot performance to prompt design. Users may also assume generalizability across tasks or modalities without proper validation.

Table 5 summarizes the evaluation strategies and highlights metric-specific limitations in detail.

Recommendations

To address the challenges, users should complement standard metrics with alignment and representation evaluations, assess label and compute efficiency, and validate generalization across datasets. For open-ended, generative tasks like image-grounded report generation, expert review often remains essential.

Large language model metrics for text-based tasks

Fundamentals

Evaluating LLMs for tasks such as translation, summarization, or open-ended generation requires metrics that capture both surface-level accuracy and deeper semantic coherence. The evaluation metrics encompasses both traditional metrics, primarily focused on surface-level agreement (e.g., BLEU, ROUGE) or fluency (e.g., Perplexity), and newer, embedding-based metrics designed to capture deeper semantic coherence (e.g., BERTScore, BARTScore) [79–86]. When combined, these metrics may allow for more nuanced evaluation, especially in tasks involving paraphrasing, abstraction, translation, or creative generation of, for example, radiology reports or clinical summaries.

Table 6 details these key evaluation approaches, outlining their definitions and core characteristics.

Pitfalls and misapplications

While widely adopted, traditional n-gram metrics like BLEU and ROUGE exhibit limitations in semantic evaluation and often fail to capture meaning when lexical variation is high [81,87]. These metrics

Table 5
Evaluation methods, metrics, and challenges for foundation models.

Category	Method / Evaluation Type	Metric	Considerations & Challenges
Transferability	k-shot evaluation	Task-specific as appropriate (e.g., accuracy, Dice score)	Performance highly sensitive to prompt/example selection and task setup; measures generalization with no (zero-shot) or minimal adaptation (few-shot). Few-shot evaluation also assesses label efficiency and allows plotting a performance metric by number of few-shot examples curve.
	Linear probing	Task-specific	Evaluates the quality of base model features; requires labeled data and training a simple linear layer on top of a (frozen) base FM.
Semantic quality & cross-modal alignment	Retrieval	Recall@k	Measures how many of the *total* known relevant items in the dataset were found within the top k results. Sensitive to *k*. Requires a labeled database (to know the total of relevant items) which can be costly to create. Often presented together with Precision@ *k*.
	Retrieval	Precision@k	Measures the proportion of relevant items within the top *k* retrieved results. E.g., if the query is a chest CT with pneumonia. Precision@ 5 = 0.8 means 4 out of the top 5 retrieved CTs show similar cases. Sensitive to *k*.
	Retrieval / Ranking	Average precision (AP)	Evaluates relevance and ranking order for a *single* query; balances precision and recall.
	Retrieval / Ranking	Mean average precision (mAP)	Average of AP scores over *multiple* queries; aggregate metric for retrieval system performance.
	Retrieval / Ranking	Mean reciprocal rank (MRR)	Evaluates rank of the first relevant item; best suited when only the top result matters (e.g., QA).
	Clustering	Visual analysis, normalized mutual information (NMI)	Evaluate feature quality by assessing separation of learned representations, usually in a lower-dimensional space. Common methods include k-nearest neighbor (KNN), t-stochastic neighbor embeddings (t-SNE) and UMAP. Allows visual analysis of how well different classes are separately represented by FM features (e.g., clusters presenting distinct anatomical structures). Sensitive to clustering algorithm & hyperparameters.
Scaling behavior	Computational efficiency	Performance (task-specific) by FLOPs (floating point operations)	Theoretical compute cost for inference or training pass; allows hardware-independent comparison of computational load.
	Parameter efficiency	Performance by model size (number of parameters)	Evaluates trade-off between model size/complexity and performance. Train multiple model variants of different sizes (e.g., varying depth, width) on the same dataset. High computational cost.
	Data efficiency	Performance by training dataset size (number of examples or fraction of full dataset)	Measures gains from more training data. Identifies saturation points. Requires training identical models on varying dataset fractions. Challenges include high cost, need for representative sampling, and ambiguous dataset size definitions (e.g., samples vs. tokens). Can also be used to assess label efficiency.

can penalize valid paraphrases or alternate phrasings and are sensitive to surface form, making them less reliable in contexts involving creative generation, paraphrasing, or abstraction. This may also limit their utility in real-world clinical applications that require domain-specific terminology, such as radiology report generation.

Perplexity, although valuable for evaluating fluency and likelihood, does not assess whether the output is factually correct or contextually appropriate. Its sensitivity to the underlying language model's distributional assumptions also limits cross-model comparisons [88].

Even with these more advanced embedding-based metrics like BERTScore and BARTScore, challenges remain. These metrics depend on pretrained models that may carry biases or domain misalignments, and they can be influenced by prompt wording or task framing [89,90]. Moreover, while they correlate better with human judgments, they still may not fully capture task-specific criteria like factual accuracy, coherence across multiple sentences, or alignment with clinical context in specialized domains like medical natural language processing, particularly in specialized applications like radiology report generation.

Recommendations

Despite the challenges noted, embedding-based neural metrics like BERTScore and BARTScore generally demonstrate superior correlation with human judgments compared to traditional approaches, particularly in scenarios involving paraphrasing and semantic equivalence. The contextual understanding inherent in these models allows them to recognize semantic similarities even when lexical overlap is minimal,

Table 6
Common metrics for evaluating LLMs in text-based tasks.

Metric	Definition	Characteristics
BLEU	Measures n-gram precision between generated and reference text, with a brevity penalty to penalize short outputs [79,80].	Language-independent and efficient; limited in capturing semantics and sensitive to word order[81,87].
ROUGE	Evaluates n-gram recall and overlap between system output and reference (especially in summarization)[81,82].	Emphasizes recall; ROUGE-N is commonly used. Provides both precision and recall, enabling a more comprehensive quality assessment[81].
Perplexity	Assesses how well a model predicts text, defined as the exponent of the cross-entropy loss[83,84].	Lower values suggest better fluency. Common in autoregressive language models; depends on model architecture and pretraining setup[88]. Only applicable when the model is probabilistic and mostly used during training or on held-out data.
BERTScore	Computes cosine similarity between contextualized token embeddings of candidate and reference using a pretrained BERT model[85].	Captures deeper semantic similarity; more robust to paraphrasing. Requires pretrained transformer models [89,90].
BARTScore	Uses BART to compute the log-probability of generating one text given another (candidate -reference)[86].	Provides bidirectional evaluation (precision and recall). Combines generation modeling with similarity assessment.

addressing a significant limitation of n-gram-based metrics. Moreover, both BERTScore and BARTScore exhibit robust performance across various text generation tasks, including machine translation, summarization, and dialogue generation, demonstrating their versatility as evaluation metrics for diverse natural language processing applications [85,86,90]. Their flexibility across various NLP tasks makes them suitable for evaluating medical LLMs, though domain adaptation remains critical.

To ensure robust evaluation of LLM outputs, it is advisable to report multiple metrics, ideally including one or more contextual embedding-based scores alongside traditional n-gram or perplexity-based measures (see Table 7 for a comparative overview to aid selection). Human evaluation remains critical, especially for open-ended or high-stakes applications, and task-specific rubrics or qualitative error analysis can supplement quantitative scores for a more holistic assessment of model quality.

Evaluating AI-generated synthetic medical images

Fundamentals

The increasing use of AI-generated synthetic medical images in radiology applications, particularly through generative adversarial networks (GANs), variational autoencoders and diffusion models, necessitates robust evaluation metrics to assess their realism, fidelity, and clinical utility [91].

Synthetic images are applied in tasks such as data augmentation, image-to-image translation, or unsupervised training, and their evaluation must go beyond pixel-level similarity to include diagnostic

Table 7
Comparative overview of common LLM evaluation metrics.

Metric	Type	Semantic Sensitivity	Fluency Sensitivity	Robustness to Paraphrasing	Requires Reference	Relative Computational Cost
BLEU	Surface-based	Low	Indirect (via n-grams)	Low	Yes	Low
ROUGE	Surface-based	Low	Indirect (via recall)	Low	Yes	Low
Perplexity	Likelihood-based	None	High	High	No	Low
BERTScore	Embedding-based	High	Indirect (via similarity)	High	Yes	High
BARTScore	Generation-based	High	High	High	Yes	High

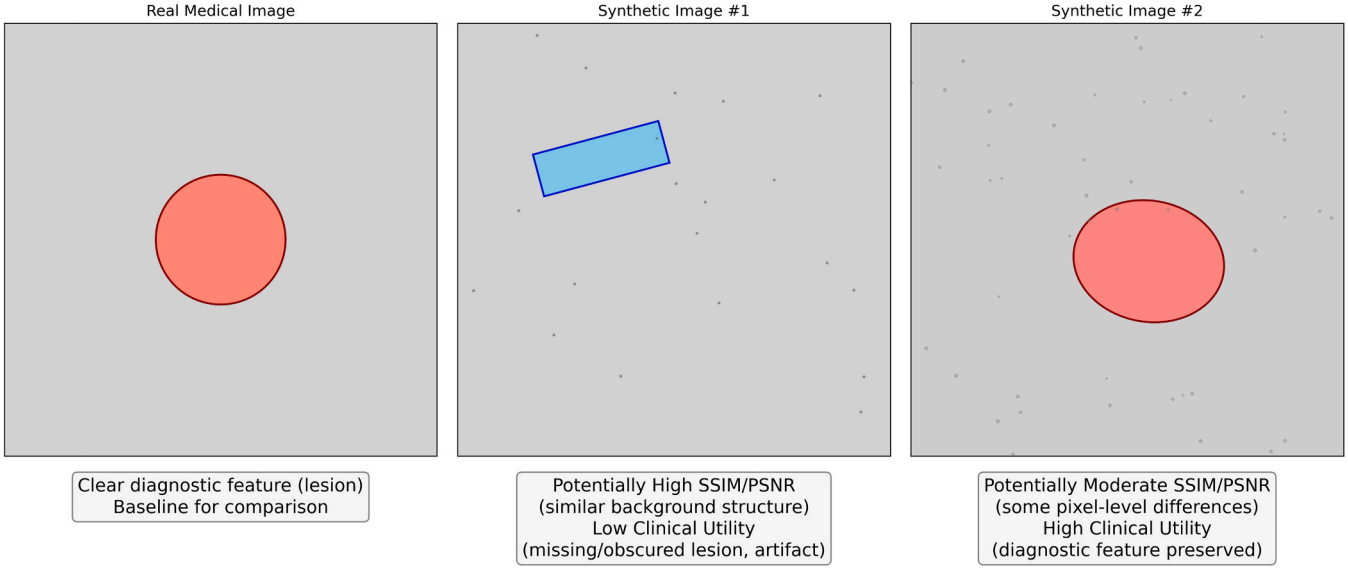


Fig. 20. Misalignment between conventional image quality metrics and clinical utility. Standard metrics may not reflect clinical relevance. Synthetic Image #1 shows high structural similarity (SSIM) and peak signal-to-noise ratio (PSNR) despite missing the diagnostic feature, indicating low clinical utility. In contrast, Synthetic Image #2 preserves the lesion and thus maintains clinical relevance, even with lower SSIM/PSNR values. This highlights the importance of domain-specific evaluation criteria in medical imaging.

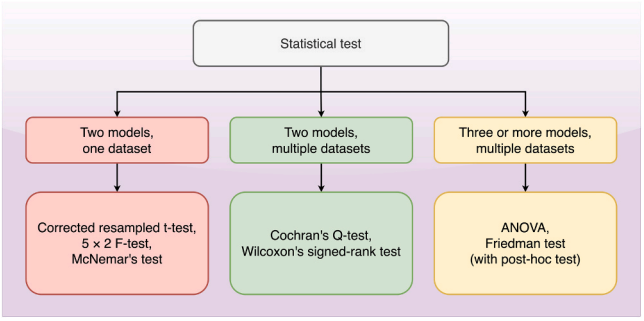


Fig. 21. Overview of often-used statistical tests. The flowchart guides test selection based on factors like the number of models being compared and the structure of the data (e.g., single vs. multiple datasets). The analysis of variance (ANOVA) and Friedman tests are omnibus tests, assessing whether there is any difference among the models. If a significant difference is detected, post hoc tests such as the Nemenyi test need to be applied to identify which models differ significantly. For tests involving human raters, inter-rater reliability measures such as Cohen's or Fleiss's Kappa can be employed. If applying other tests like the (uncorrected) *t*-test or the DeLong test, one needs to ensure that the application is correct, since these tests may have certain assumptions (e.g., independent test statistics, or non-nested models). More information on the tests can be found in [103,110].

credibility and clinical plausibility.

Initially, evaluation often relied on a variety of image quality metrics employed to assess the similarity between synthetic and real images, particularly in image-to-image translation tasks. These include traditional metrics such as structural similarity index measure (SSIM) [92], MSE, mean absolute error (MAE), peak signal-to-noise ratio (PSNR), histogram matching scores, and learned perceptual image patch similarity (LPIPS) [93].

To better address the specific requirements of evaluating deep generative models, dedicated model-based metrics have been developed. The inception score (IS) [94] and the Fréchet inception distance (FID) [95] are among the most commonly used for assessing the realism and diversity of generated images. These metrics evaluate statistical similarity between synthetic and real image distributions, typically using a pretrained feature extractor.

More recently, domain-adapted metrics such as the Fréchet radiomics distance (FRD) have been proposed [96], aiming to bridge the gap by comparing the distribution of radiomic features, quantitative imaging biomarkers, between real and synthetic images, thus focusing on clinical and biological relevance.

Pitfalls and misapplications

However, a significant pitfall is that metrics like IS and FID, being derived from models trained on natural images, may not fully capture domain-specific features relevant to radiology. Moreover, these metrics can introduce biases that limit their utility in clinical contexts [97–99]. Likewise, traditional metrics such as SSIM and PSNR, while useful for quantifying pixel-level similarity, often fail to reflect clinical realism, anatomical correctness, or perceptual coherence, especially in pathologically important regions.

While these metrics provide useful low-level comparisons, they often fail to align with human judgment of perceptual similarity, particularly in the complex and nuanced domain of medical imaging. Evaluating synthetic images in radiology requires metrics that reflect not only image quality but also clinical interpretability (Fig. 20). Without incorporating domain-specific criteria, a synthetic image that scores well in general-purpose metrics might still be diagnostically misleading or even harmful in downstream applications.

Recommendations

To capture the full range of clinical and perceptual fidelity, human observer studies remain indispensable. These involve expert radiologists assessing the realism, diagnostic quality, and anatomical accuracy of synthetic images, often via visual Turing tests or diagnostic tasks [100]. Additionally, downstream task performance, such as the effect of synthetic data on training classification or segmentation models, provides indirect but meaningful validation of utility.

In conclusion, no single metric suffices. A comprehensive multimetric framework combining traditional, learned, and domain-specific evaluations, alongside irreplaceable human expertise, is essential for

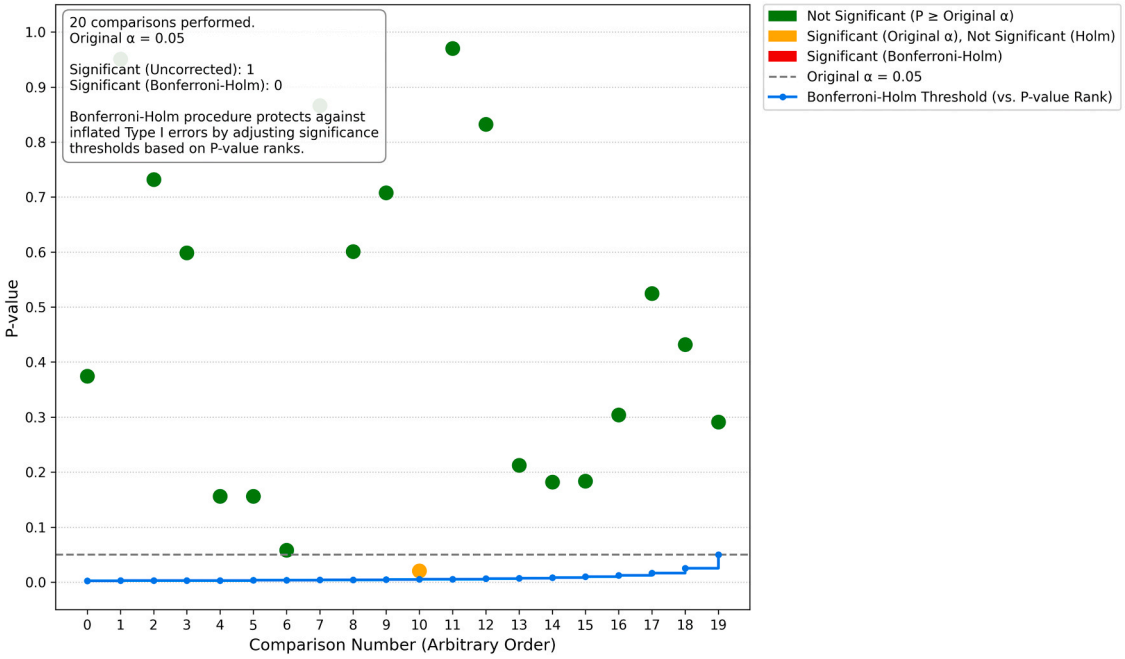


Fig. 22. Importance of multiplicity testing. Only one test appears significant under the uncorrected threshold (orange), but none remain significant after correction, illustrating the method's control over Type I error inflation.

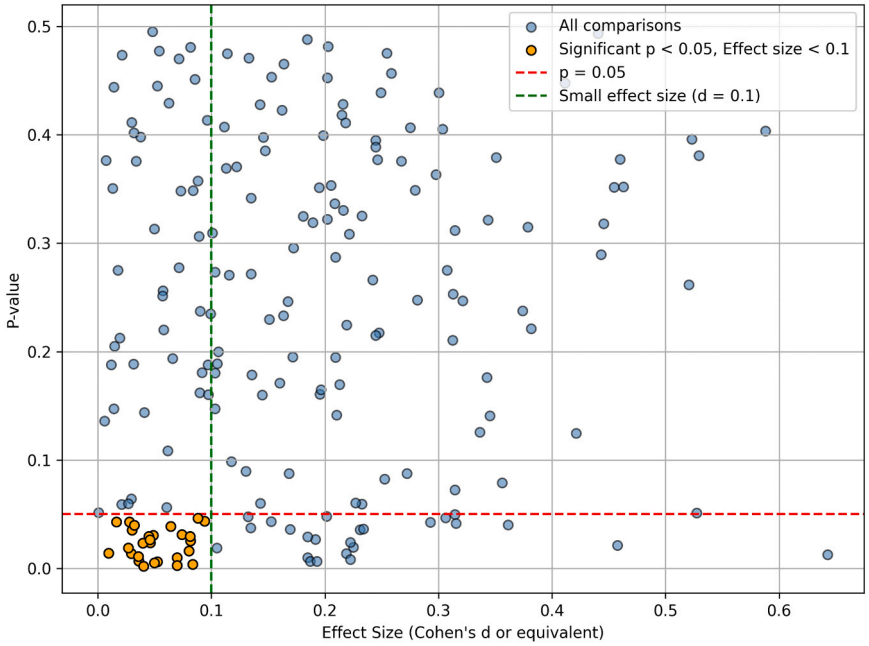


Fig. 23. Scatterplot of effect size versus p-value across model comparisons. Orange points represent statistically significant results ($p < 0.05$) with small effect sizes ($d < 0.1$), meaning that statistical significance does not always imply practical relevance.

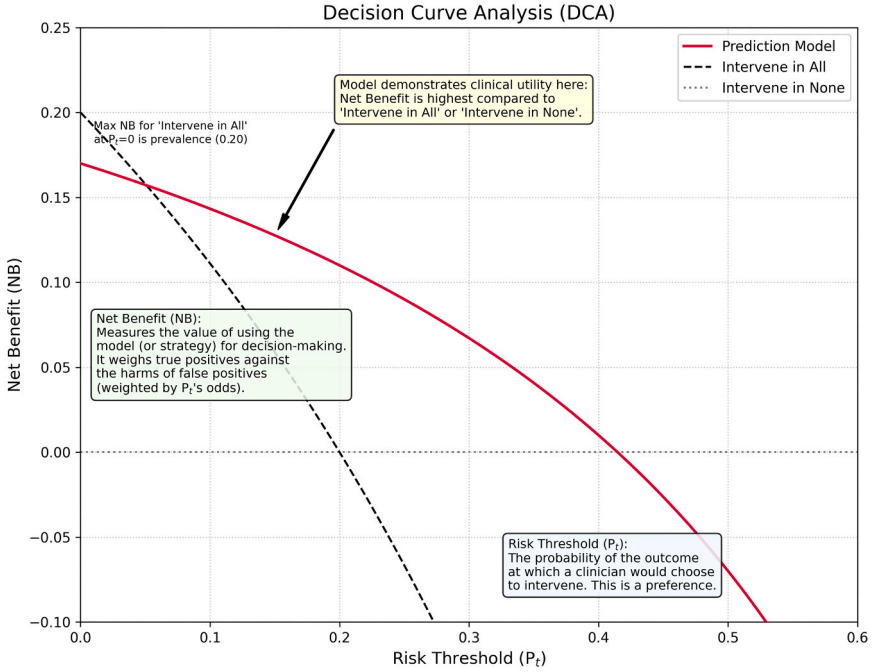


Fig. 24. Basic interpretation of decision curve analysis (DCA). Net benefit (NB) is plotted against varying risk thresholds (Pt), illustrating the clinical utility of a prediction model. The model is considered beneficial where its curve lies above both "Intervene in All" and "Intervene in None" strategies. NB quantifies the trade-off between true positives and false positives, adjusted by Pt's odds.

the rigorous and clinically meaningful assessment of synthetic medical images.

Comparing metrics across models

Fundamentals

A model's performance metric, when viewed in isolation, often lacks meaningful interpretation. Its significance emerges only through comparison with alternative models, baseline models (such as a constant or random model), or established reference standards [101].

Statistical testing is the standard method for formalizing these comparisons (Fig. 21). In this process, chosen metrics are calculated and compared for the models on independent, representative test data to determine if observed differences are significant or likely due to chance. When dedicated independent test data is scarce, or to make more robust estimates, resampling methods such as cross-validation or bootstrapping are commonly used to maximize the use of the data [102]. These

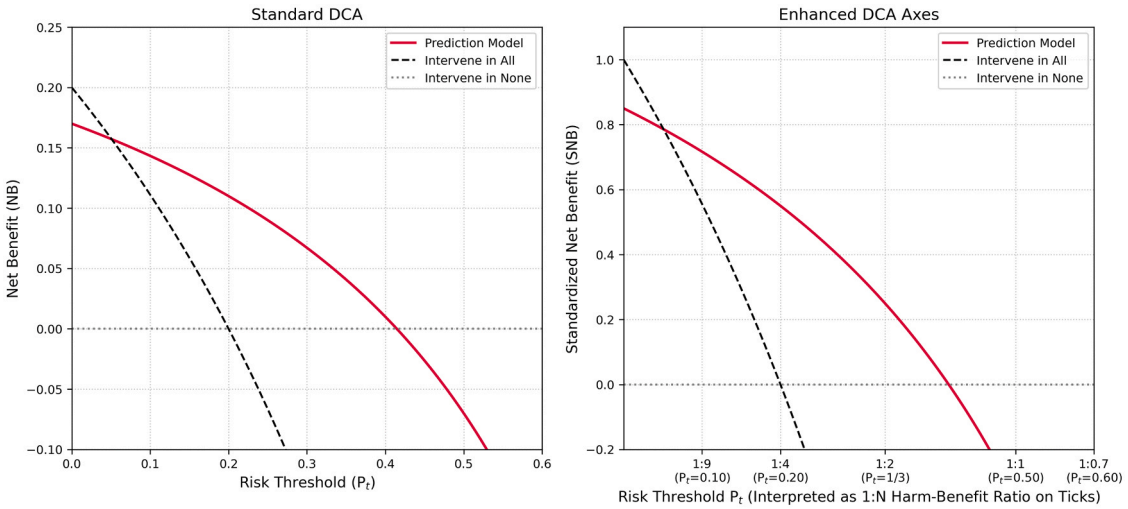


Fig. 25. Comparison of standard and enhanced decision curve analysis (DCA) axes. The left panel shows net benefit (NB) across risk thresholds (Pt), while the right panel presents standardized net benefit (SNB) with Pt reinterpreted as harm-benefit ratios (1:N). This enhances clinical interpretability; for instance, Pt = 0.10 corresponds to a 1:9 trade-off between missing an outcome and unnecessary intervention.

systematically split the data into subsets, ensuring that some parts are used exclusively for training while others are used for testing. The choice of resampling method often depends on the data size and the available computational resources.

Pitfalls and misapplications

One key challenge is identifying the appropriate statistical test [103] (Fig. 21). Parametric tests, such as the *t*-test, rely on assumptions about data distributions that, if unmet, can potentially lead to misleading or invalid conclusions. Non-parametric alternatives are more robust to such assumptions but generally have lower statistical power, meaning they require more data or larger effect sizes to detect a real difference.

Dedicated test data are often not available in sufficient quality and are instead created by splitting the existing data. While this allows for independent evaluation, it reduces the sample size available for training. If too many samples are reserved, the model may underperform; if too few, the test set may be unrepresentative, reducing generalizability or statistical power, that is the ability to detect a true difference (Type II error) [104].

When resampling is applied, the resulting metrics are often not independent, since the training is often on overlapping data. In this case, standard tests that assume independence not apply directly and either corrections, like the corrected resampled *t*-test, or alternative strategies such as permutation tests or nested cross-validation, must be applied. Also, resampling may distort class balance; hence, stratified sampling, which preserves class distributions across splits, should be used when applicable, especially for classification tasks with imbalanced datasets.

Attention must also be given to multiple testing, which increases the risk of observing differences when in reality there are none (Type I error). Corrections like the Bonferroni-Holm method should be applied in this case **(**Fig. 22**)** [103].

Observed significance can be misleading for several reasons. First, not all aspects of model quality can be captured adequately by metrics [105], for example, there is currently no widely accepted metric for the explainability of a model. Second, for some metrics, a higher score does not necessarily indicate better performance, or they were designed to be applied only in specific contexts [106]. In such situations, additional evaluation using human raters may be beneficial. Third, the choice of the metric itself can be subjective. Furthermore, other aspects might play a role that are not directly part of the model. For example, a newly designed model can perform better than established models, but the latter could be simpler, more known and trusted, which would affect its later use [105]. In addition, when comparing to reference standards, significance may arise from confounding factors such as differences in training data or hyperparameter tuning strategies.

Moreover, multiple metrics are often of interest, which may not only be on different scales but can also favor different models. For instance, one model might achieve higher precision, while another performs better in terms of recall, leading to conflicting conclusions about which model is superior.

Recommendations

To avoid ‘p-hacking’, the set of metrics and the primary metric(s) for decision-making should be defined in advance, and models should not be repeatedly improved and re-tested on the same evaluation data until statistical significance is achieved.

In cases, where multiple metrics yield conflicting results, an analysis employing weighted aggregation of the metrics can help resolve the contradictions and support a more balanced model selection [107].

While statistical testing is widely used, it is not without criticism [108]. A statistically significant result may not translate into practical relevance [101]. To address this, confidence intervals should be reported to capture the uncertainty surrounding metric estimates, and effect sizes should be calculated to better understand the magnitude of the observed differences **(**Fig. 23**)** [109].

Where applicable, results should also be compared on standardized benchmark datasets to enhance transparency and reproducibility. Ultimately, performance comparisons should be designed with careful attention to test conditions, the nature of the metrics used, the assumptions of statistical methods, and the clinical or research context in which the models are applied.

Translating metric-based evaluations to real-world clinical practice

Fundamentals

In an era of abundant medical data, translating this information into actionable knowledge remains challenging. Clinical utility was developed to aid in bridging the gap between researchers who design these novel prediction tools and the decision makers, including clinicians, who would like to apply them in practice [111]. Commonly used statistical methods for quantifying clinical utility, such as decision curve analyses (DCA), are simplified versions of cost-benefit analyses, as they

Table 8
Summary of key pitfalls and recommendations.

Metric Category	Key Pitfalls	Key Recommendations
Classification	– Accuracy can be misleading in imbalanced datasets. – AUROC may overestimate performance on minority class. – Predictive values are prevalence-dependent and often misinterpreted. – In multiclass settings, inappropriate selection or reporting of averaging methods (e.g., macro, micro, weighted) can obscure poor performance on rare or clinically critical classes.	– Avoid relying solely on accuracy, especially in imbalanced data. – Use AUPRC over AUROC when the positive class is rare. – Report F1-score and MCC at chosen thresholds. – Interpret PPV/NPV with population prevalence in mind. – In multiclass settings, clearly specify the chosen averaging method (macro, micro, weighted) and ensure it aligns with the task objectives and class distribution.
Regression	– MAE treats large and small errors equally. – RMSE is overly sensitive to outliers. – MAPE is unstable near zero and penalizes under-prediction more. – R^2 can appear good despite poor clinical performance.	– Use multiple metrics (MAE, RMSE, R^2, MAPE) together. – Replace MAPE with sMAPE or MASE in low-value contexts. – Use adjusted R^2 and visualize predictions vs. actual. – Consider clinical relevance of error.
Detection	– No consensus on metric use limits cross-study comparisons. – IoU is highly sensitive to object size. – Segmentation metrics can be misused for detection. – Improper averaging misrepresents performance.	– Use multiple, problem-tailored detection metrics. – Avoid segmentation metrics for instance detection. – Account for object size sensitivity in IoU. – Report variability and stratify performance.
Survival	– C-index is affected by censoring. – Time-dependent AUC definitions vary. – Hosmer-Lemeshow is sensitive to binning. – Brier score conflates discrimination and calibration.	– Specify the variant of C-index or AUC used. – Use Uno's C-index in high-censoring datasets. – Avoid overreliance on Hosmer-Lemeshow. – Use net benefit to assess clinical utility.
Segmentation	– DSC may under-penalize boundary errors. – IoU is overly sensitive for small structures. – Overlap metrics ignore boundary accuracy. – Scores lack context sensitivity.	– Use IoU for small and DSC for large structures. – Supplement with boundary-specific metrics. – Interpret within anatomical/clinical context. – Compare to inter-observer variability.
Calibration	– ECE may hide local miscalibrations. – Reliability diagrams can be misleading in imbalance. – Brier score blends calibration and discrimination. – Overconfident models often go unchecked.	– Report both ECE and Brier score with binning details. – Include reliability diagrams. – Apply recalibration methods pre-deployment. – Assess calibration internally and externally.
Foundation Models	– Overreliance on task metrics hides generalization gaps. – Prompt sensitivity distorts few-/zero-shot evaluations. – Setup transparency often lacking. – Generalizability assumed without validation.	– Report setup details and prompt sensitivity. – Include transferability, alignment, and scaling behaviors. – Use retrieval and probing metrics. – Validate cross-domain generalization explicitly.
LLMs	– BLEU/ROUGE fail at capturing semantics. – Perplexity does not reflect correctness. – Embedding metrics can be biased. – Factual consistency hard to measure.	– Combine traditional and contextual metrics. – Add human evaluations for critical tasks. – Monitor for domain mismatch and bias.
Synthetic Images	– SSIM/PSNR do not reflect diagnostic fidelity. – IS and FID use non-medical pretrained models. – Metrics miss clinical plausibility. – No single metric is sufficient.	– Combine traditional and learned metrics. – Include expert assessments for realism and value. – Use downstream task performance. – Do not rely solely on SSIM/PSNR.
Model Comparison	– Wrong tests invalidate comparisons. – Small sets reduce power. – Multiple tests inflate false positives. – Conflicting metrics confuse interpretation.	– Predefine metrics and decision criteria. – Use statistical tests with correction. – Report effect sizes and confidence intervals. – Use weighted aggregation when metrics conflict.
Clinical Translation	– Risk thresholds misunderstood as model parameters. – Net benefit does not reflect individual utility. – Decision curve analysis can be misinterpreted without context. – Oversimplified usage undermines value.	– Use decision curve analysis – Prefer standardized over absolute net benefit. – Reframe thresholds as harm-benefit ratios. – Clearly state assumptions and decision thresholds.

AUROC, Area Under the Receiver Operating Characteristic Curve; AUPRC, Area Under the Precision-Recall Curve; MCC, Matthews Correlation Coefficient; PPV, Positive Predictive Value; NPV, Negative Predictive Value; MAE, Mean Absolute Error; RMSE, Root Mean Squared Error; MAPE, Mean Absolute Percentage Error; sMAPE, Symmetric Mean Absolute Percentage Error; MASE, Mean Absolute Scaled Error; IoU, Intersection over Union; DSC, Dice Similarity Coefficient; ECE, Expected Calibration Error; LLMs, Large Language Models; SSIM, Structural Similarity Index Measure; PSNR, Peak Signal-to-Noise Ratio; IS, Inception Score; FID, Frechet Inception Distance; DCA, Decision Curve Analysis.

indirectly account for the clinical consequences of medical decisions [112–114]. These simplifications have led to frequent misinterpretations of the methods [115,116].

DCA is a graphical approach to characterize the clinical utility of a prediction tool under multiple intervention strategies **(**Fig. 24**)**. Three strategies are typically considered: one guided by the prediction tool and two reflecting the extremes where the intervention is applied to either all patients or none [114].

The x-axis of this figure denotes a clinician's risk threshold [115]. This represents the probability of an outcome at which a clinician (or patient/policy) would choose to intervene. It is quantified using the risk threshold and reflects how the clinician weighs the outcomes associated with intervening or not [112,114,115].

The y-axis denotes the net benefit of administering the intervention when utilizing a treatment strategy. It is defined:

sensitivity x prevalence – (1-specificity) x (1-prevalence) x the odds of the risk threshold [115].

Essentially, net benefit measures the value of using the model or strategy, by weighting true positives against the harms of false positives, where the weighting is determined by the chosen risk threshold's odds.

If the difference in net benefits between a prediction tool and a comparator strategy is positive, then this is evidence of its clinical utility for a predefined risk threshold.

Pitfalls and misapplications

Vickers et al. note that confusion around DCA often stems from misinterpreting its core metrics [115].

A significant pitfall relates to the risk threshold. It is not a value derivable from the model's performance data itself. Instead, it reflects subjective factors including the perceived severity of the outcome, the effectiveness and side effects of the treatment, patient preferences, and resource availability. Misunderstanding it as an objective model output rather than a subjective preference can lead to flawed interpretations of clinical utility.

Similarly, net benefit has its interpretational challenges. It is a population-level metric, often conceptualized in terms of net true positives gained per patient. This means it does not directly translate to the benefit for an individual patient but rather reflects the overall advantage of a strategy across a population. These limitations, if not properly understood, can result in DCA being designed or interpreted superficially in many studies.

Recommendations

To enhance the interpretability and practical application of DCA, several modifications and alternative conceptualizations have been recommended.

As a more intuitive alternative to the risk threshold, the harm-benefit ratio (the odds of the risk threshold) can be used (Fig. 25); for instance, a 10 % threshold corresponds to a 1:9 ratio, meaning missing an outcome is considered nine times worse than an unnecessary intervention [115].

Standardized net benefit, defined as the net benefit divided by prevalence, may be preferred since the maximum net benefit equals the prevalence. A standardized net benefit of 75 % indicates that the prediction tool achieves 75 % of the maximum possible utility [116]. Together, these axis modifications enhance DCA interpretability (Fig. 25).

Additional details regarding DCA assumptions, other common misinterpretations, and further nuances of its application have been well documented [112,114,115,117].

Final thoughts

Robust model evaluation in medical imaging AI requires more than reporting frequently used standard metrics. As summarized in the Table 8, each metric type presents distinct pitfalls that may lead to misinterpretation if unaddressed. To mitigate these, the authors offer task-specific recommendations that support more reliable and meaningful assessments (see Table 8 for summary of recommendations as well). Overall, evaluators should align metric selection with the clinical question, dataset characteristics, and intended use. Using multi-metric strategies, clarifying assumptions, and avoiding common misapplications are essential for trustworthy model evaluation. We hope this review serves as a practical reference for researchers striving to implement rigorous and clinically relevant AI evaluations in medical imaging.

Funding

None.

Ethical Statement

N/a.

CRediT authorship contribution statement

Keno K. Bressem: Writing – review & editing, Writing – original draft. **Christian Bluethgen:** Writing – review & editing, Writing – original draft. **Aydın Demircioğlu:** Writing – review & editing, Writing – original draft. **Aymen Meddeb:** Writing – review & editing, Writing – original draft. **Arnaldo Stanzione:** Writing – review & editing, Writing – original draft. **Michail E. Klontzas:** Writing – review & editing, Writing – original draft. **Burak Kocak:** Writing – review & editing, Writing – original draft, Conceptualization. **Renato Cuocolo:** Writing – review & editing, Writing – original draft. **Oliver Díaz:** Writing – review & editing, Writing – original draft. **Nathaniel Mercaldo:** Writing – review & editing, Writing – original draft. **Lorenzo Ugga:** Writing – review & editing, Writing – original draft.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Burak Kocak, Michail Klontzas, and Renato Cuocolo serve as Editors for European Journal of Radiology Artificial Intelligence (EJR AI). None of them were involved in the review or selection process for this article.

Data availability

No data was used for the research described in the article.

References

[1] K. Pierre, A.G. Haneberg, S. Kwak, K.R. Peters, B. Hochhegger, T. Sananmuang, P. Tunlayadechanont, P.J. Tighe, A. Mancuso, R. Forghani, Applications of artificial intelligence in the radiology roundtrip: process streamlining, workflow optimization, and beyond, Semin. Roentgenol. 58 (2023) 158–169, https://doi.org/10.1053/j.ro.2023.02.003.

[2] L. Pinto-Coelho, How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications, Bioeng. (Basel) 10 (2023) 1435, https://doi.org/10.3390/bioengineering10121435.

[3] M. Khalifa, M. Albadawy, AI in diagnostic imaging: revolutionising accuracy and efficiency, Comput. Methods Prog. Biomed. Update 5 (2024) 100146, https://doi.org/10.1016/j.cmpbup.2024.100146.

[4] K. Drukker, W. Chen, J. Gichoya, N. Gruszauskas, J. Kalpathy-Cramer, S. Koyejo, K. Myers, R.C. Sá, B. Sahiner, H. Whitney, Z. Zhang, M. Giger, Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment, J. Med Imaging (Bellingham) 10 (2023) 061104, https://doi.org/10.1117/1.JMI.10.6.061104.

[5] R.K. Samala, K. Drukker, A. Shukla-Dave, H.-P. Chan, B. Sahiner, N. Petrick, H. Greenspan, U. Mahmood, R.M. Summers, G. Tourassi, T.M. Deserno, D. Regge, J.J. Näppi, H. Yoshida, Z. Huo, Q. Chen, D. Vergara, K.H. Cha, R. Mazurchuk, K. T. Grizzard, H. Huisman, L. Morra, K. Suzuki, S.G. Armato, I.I.I., L. Hadjiiski, AI and machine learning in medical imaging: key points from development to translation, BJR|Artif. Intell. 1 (2024) ubae006, https://doi.org/10.1093/bjrai/ubae006.

[6] S. Keni, Evaluating artificial intelligence for medical imaging: a primer for clinicians, Br. J. Hosp. Med 85 (2024) 1–13, https://doi.org/10.12968/hmed.2024.0312.

[7] O. Rainio, J. Teuho, R. Klén, Evaluation metrics and statistical tests for machine learning, Sci. Rep. 14 (2024) 6086, https://doi.org/10.1038/s41598-024-56706-x.

[8] L. Maier-Hein, A. Reinke, P. Godau, M.D. Tizabi, F. Buettner, E. Christodoulou, B. Glocker, F. Isensee, J. Kleesiek, M. Kozubek, M. Reyes, M.A. Riegler, M. Wiesenfarth, A.E. Kavur, C.H. Sudre, M. Baumgartner, M. Eisenmann, D. Heckmann-Nötzel, T. Rädsch, L. Acion, M. Antonelli, T. Arbel, S. Bakas, A. Benis, M.B. Blaschko, M.J. Cardoso, V. Cheplygina, B.A. Cimini, G.S. Collins, K. Farahani, L. Ferrer, A. Galdran, B. van Ginneken, R. Haase, D.A. Hashimoto, M. M. Hoffman, M. Huisman, P. Jannin, C.E. Kahn, D. Kainmueller, B. Kainz, A. Karargyris, A. Karthikesalingam, F. Kofler, A. Kopp-Schneider, A. Kreshuk, T. Kurc, B.A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A.L. Martel, P. Mattson, E. Meijering, B. Menze, K.G.M. Moons, H. Müller, B. Nichyporuk, F. Nickel, J. Petersen, N. Rajpoot, N. Rieke, J. Saez-Rodriguez, C.I. Sánchez, S. Shetty, M. van Smeden, R.M. Summers, A.A. Taha, A. Tiulpin, S.A. Tsaftaris, B. Van Calster, G. Varoquaux, P.F. Jäger, Metrics reloaded: recommendations for image analysis validation, Nat. Methods 21 (2024) 195–212, https://doi.org/10.1038/s41592-023-02151-z.

[9] A. Reinke, M.D. Tizabi, C.H. Sudre, M. Eisenmann, T. Rädsch, M. Baumgartner, L. Acion, M. Antonelli, T. Arbel, S. Bakas, P. Bankhead, A. Benis, M. Blaschko, F. Buettner, M.J. Cardoso, J. Chen, V. Cheplygina, E. Christodoulou, B. Cimini, G.S. Collins, S. Engelhardt, K. Farahani, L. Ferrer, A. Galdran, B. van Ginneken, B. Glocker, P. Godau, R. Haase, F. Hamprecht, D.A. Hashimoto, D. Heckmann-Nötzel, P. Hirsch, M.M. Hoffman, M. Huisman, F. Isensee, P. Jannin, C.E. Kahn, D. Kainmueller, B. Kainz, A. Karargyris, A. Karthikesalingam, A.E. Kavur, H. Kenngott, J. Kleesiek, A. Kleppe, S. Kohler, F. Kofler, A. Kopp-Schneider, T. Kooi,

M. Kozubek, A. Kreshuk, T. Kurc, B.A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A.L. Martel, P. Mattson, E. Meijering, B. Menze, D. Moher, K.G.M. Moons, H. Müller, B. Nichyporuk, F. Nickel, M.A. Noyan, J. Petersen, G. Polat, S.M. Rafelski, N. Rajpoot, M. Reyes, N. Rieke, M. Riegler, H. Rivaz, J. Saez-Rodriguez, C.I. Sánchez, J. Schroeter, A. Saha, M.A. Selver, L. Sharan, S. Shetty, M. van Smeden, B. Stieltjes, R.M. Summers, A.A. Taha, A. Tiulpin, S.A. Tsaftaris, B.V. Calster, G. Varoquaux, M. Wiesenfarth, Z.R. Yaniv, P. Jäger, L. Maier-Hein, Common Limitations of Image Processing Metrics: A Picture Story, (2023). ⟨https://doi.org/10.48550/arXiv.2104.05642⟩.

[10] A. Reinke, M.D. Tizabi, M. Baumgartner, M. Eisenmann, D. Heckmann-Nötzel, A. E. Kavur, T. Rädsch, C.H. Sudre, L. Acion, M. Antonelli, T. Arbel, S. Bakas, A. Benis, F. Buettner, M.J. Cardoso, V. Cheplygina, J. Chen, E. Christodoulou, B. A. Cimini, K. Farahani, L. Ferrer, A. Galdran, B. van Ginneken, B. Glocker, P. Godau, D.A. Hashimoto, M.M. Hoffman, M. Huisman, F. Isensee, P. Jannin, C. E. Kahn, D. Kainmueller, B. Kainz, A. Karargyris, J. Kleesiek, F. Kofler, T. Kooi, A. Kopp-Schneider, M. Kozubek, A. Kreshuk, T. Kurc, B.A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A.L. Martel, E. Meijering, B. Menze, K.G.M. Moons, H. Müller, B. Nichyporuk, F. Nickel, J. Petersen, S.M. Rafelski, N. Rajpoot, M. Reyes, M.A. Riegler, N. Rieke, J. Saez-Rodriguez, C.I. Sánchez, S. Shetty, R. M. Summers, A.A. Taha, A. Tiulpin, S.A. Tsaftaris, B. Van Calster, G. Varoquaux, Z.R. Yaniv, P.F. Jäger, L. Maier-Hein, Understanding metric-related pitfalls in image analysis validation, Nat. Methods 21 (2024) 182–194, https://doi.org/10.1038/s41592-023-02150-0.

[11] M. Paschali, Z. Chen, L. Blankemeier, M. Varma, A. Youssef, C. Bluethgen, C. Langlotz, S. Gatidis, A. Chaudhari, Foundation models in radiology: what, how, why, and why not, Radiology 314 (2025) e240597, https://doi.org/10.1148/radiol.240597.

[12] J.T. Hancock, T.M. Khoshgoftaar, J.M. Johnson, Evaluating classifier performance with highly imbalanced big data, J. Big Data 10 (2023) 42, https://doi.org/10.1186/s40537-023-00724-5.

[13] Learning from Imbalanced Data, (n.d.). ⟨https://ieeexplore.ieee.org/document/5128907⟩ (accessed May 23, 2025).

[14] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, PLoS One 10 (2015) e0118432, https://doi.org/10.1371/journal.pone.0118432.

[15] D.G. Altman, J.M. Bland, Diagnostic tests 2: predictive values, BMJ 309 (1994) 102, https://doi.org/10.1136/bmj.309.6947.102.

[16] M. Grandini E. Bagli G. Visani Metrics multi-class classification: an overview 2020 doi: 10.48550/arXiv.2008.05756.

[17] M.B.A. McDermott H. Zhang L.H. Hansen G. Angelotti J. Gallifant A closer look. auroc auprc cl. imbalance 2025 doi: 10.48550/arXiv.2401.06091.

[18] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genom. 21 (2020) 6, https://doi.org/10.1186/s12864-019-6413-7.

[19] A. Luque, A. Carrasco, A. Martín, A. de las Heras, The impact of class imbalance in classification performance metrics based on the binary confusion matrix, Pattern Recognit. 91 (2019) 216–231, https://doi.org/10.1016/j.patcog.2019.02.023.

[20] L.D. Maxim, R. Niebo, M.J. Utell, Screening tests: a review with examples, Inhal. Toxicol. 26 (2014) 811–828, https://doi.org/10.3109/08958378.2014.955932.

[21] C.J. Willmott, K. Matsuura, Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, Clim. Res. 30 (2005) 79–82, https://doi.org/10.3354/cr030079.

[22] R.J. Hyndman, A.B. Koehler, Another look at measures of forecast accuracy, Int. J. Forecast. 22 (2006) 679–688, https://doi.org/10.1016/j.ijforecast.2006.03.001.

[23] S.A. Glantz, B.K. Slinker, T.B. Neilands, In: primer of applied regression and analysis of variance, 3e, McGraw-Hill Education,, New York, NY, 2017.

[24] A.D. Myttenaere, B. Golden, B.L. Grand, F. Rossi, Mean absolute percentage error for regression models, Neurocomputing 192 (2016) 38–48, https://doi.org/10.1016/j.neucom.2015.12.114.

[25] T. Chai, R.R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)? – arguments against avoiding RMSE in the literature, Geosci. Model Dev. 7 (2014) 1247–1250, https://doi.org/10.5194/gmd-7-1247-2014.

[26] P.J. Huber, Robust estimation of a location parameter, Ann. Math. Stat. 35 (1964) 73–101, https://doi.org/10.1214/aoms/1177703732.

[27] S. Kim, H. Kim, A new metric of absolute percentage error for intermittent demand forecasts, Int. J. Forecast. 32 (2016) 669–679, https://doi.org/10.1016/j.ijforecast.2015.12.003.

[28] D. Chicco, M.J. Warrens, G. Jurman, The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation, PeerJ Comput. Sci. 7 (2021) e623, https://doi.org/10.7717/peerj-cs.623.

[29] T.G. Clark, M.J. Bradburn, S.B. Love, D.G. Altman, Survival analysis part I: basic concepts and first analyses, Br. J. Cancer 89 (2003) 232–238, https://doi.org/10.1038/sj.bjc.6601118.

[30] H. Uno, T. Cai, M.J. Pencina, R.B. D'Agostino, L.J. Wei, On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data, Stat. Med 30 (2011) 1105–1117, https://doi.org/10.1002/sim.4154.

[31] M.J. Pencina, R.B. D'Agostino, L. Song, Quantifying discrimination of Framingham risk functions with different survival C statistics, Stat. Med 31 (2012) 1543–1553, https://doi.org/10.1002/sim.4508.

[32] P.J. Heagerty, Y. Zheng, Survival model predictive accuracy and ROC curves, Biometrics 61 (2005) 92–105, https://doi.org/10.1111/j.0006-341X.2005.030814.x.

[33] K. Han, K. Song, B.W. Choi, How to develop, validate, and compare clinical prediction models involving radiological parameters: study design and statistical methods, Korean J. Radio. 17 (2016) 339–350, https://doi.org/10.3348/kjr.2016.17.3.339.

[34] M. Assel, D.D. Sjoberg, A.J. Vickers, The brier score does not evaluate the clinical utility of diagnostic tests or prediction models, Diagn. Progn. Res 1 (2017) 19, https://doi.org/10.1186/s41512-017-0020-3.

[35] M.J. Pencina, R.B. D'Agostino, R.B. D'Agostino, R.S. Vasan, Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond, Stat. Med 27 (2008) 157–172, https://doi.org/10.1002/sim.2929.

[36] S.Y. Park, J.E. Park, H. Kim, S.H. Park, Review of statistical methods for evaluating the performance of survival or other time-to-event prediction models (from conventional to deep learning approaches), Korean J. Radio. 22 (2021) 1697–1707, https://doi.org/10.3348/kjr.2021.0223.

[37] P. Blanche, J.-F. Dartigues, H. Jacqmin-Gadda, Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks, Stat. Med 32 (2013) 5381–5397, https://doi.org/10.1002/sim.5958.

[38] S.A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M.A. Riegler, P. Halvorsen, S. Parasa, On evaluation metrics for medical applications of artificial intelligence, Sci. Rep. 12 (2022) 5979, https://doi.org/10.1038/s41598-022-09954-8.

[39] J.-H. Oh, H.-G. Kim, K.M. Lee, Developing and evaluating deep learning algorithms for object detection: key points for achieving superior model performance, Korean J. Radiol. 24 (2023) 698–714, https://doi.org/10.3348/kjr.2022.0765.

[40] R. Yang, Y. Yu, Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis, Front. Oncol. 11 (2021), https://doi.org/10.3389/fonc.2021.638182.

[41] R. Padilla, S.L. Netto, E.A.B. Da Silva, A survey on performance metrics for object-detection algorithms, in: 2020 International Conference on Systems, Signals and Image Processing (IWSSIP), IEEE, Niterói, Brazil, 2020, pp. 237–242, https://doi.org/10.1109/IWSSIP48289.2020.9145130.

[42] L. Maier-Hein, M. Eisenmann, A. Reinke, S. Onogur, M. Stankovic, P. Scholz, T. Arbel, H. Bogunovic, A.P. Bradley, A. Carass, C. Feldmann, A.F. Frangi, P. M. Full, B. van Ginneken, A. Hanbury, K. Honauer, M. Kozubek, B.A. Landman, K. März, O. Maier, K. Maier-Hein, B.H. Menze, H. Müller, P.F. Neher, W. Niessen, N. Rajpoot, G.C. Sharp, K. Sirinukunwattana, S. Speidel, C. Stock, D. Stoyanov, A. A. Taha, F. van der Sommen, C.-W. Wang, M.-A. Weber, G. Zheng, P. Jannin, A. Kopp-Schneider, Why rankings of biomedical image analysis competitions should be interpreted with care, Nat. Commun. 9 (2018) 5217, https://doi.org/10.1038/s41467-018-07619-7.

[43] D. Zimmerer, K. Maier-Hein, Beyond Heatmaps: A Comparative Analysis of Metrics for Anomaly Localization in Medical Images, in: 2024. ⟨https://openreview.net/forum?id=bwaTJzL6fn&referrer=%5Bthe%20profile%20of%20David%20Zimmerer%5D(%2Fprofile%3Fid%3D~David_Zimmerer1)⟩) (accessed May 6, 2025).

[44] A. Stanzione, R. Cuocolo, L. Ugga, F. Verde, V. Romeo, A. Brunetti, S. Maurea, Oncologic imaging and radiomics: a walkthrough review of methodological challenges, Cancers (Basel) 14 (2022) 4871, https://doi.org/10.3390/cancers14194871.

[45] K.H. Zou, S.K. Warfield, A. Bharatha, C.M.C. Tempany, M.R. Kaus, S.J. Haker, W. M. Wells, F.A. Jolesz, R. Kikinis, Statistical validation of image segmentation quality based on a spatial overlap index, Acad. Radio. 11 (2004) 178–189, https://doi.org/10.1016/s1076-6332(03)00671-8.

[46] R.O. Lane A compr. rev. classif. probab. calibration metr. 2025 doi: 10.48550/ARXIV.2504.18278.

[47] N. Posocco A. Bonnefoy Estim. expect. calibration errors 2021 doi: 10.48550/ARXIV.2109.03480.

[48] K. Rufibach, Use of Brier score to assess binary predictions, J. Clin. Epidemiol. 63 (2010) 938–939, https://doi.org/10.1016/j.jclinepi.2009.11.009.

[49] J. Platt, others, probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Adv. Large Margin Classif. 10 (1999) 61–74.

[50] A.-J. Rousseau, T. Becker, S. Appeltans, M. Blaschko, D. Valkenborg, Post hoc calibration of medical segmentation models, Discov. Appl. Sci. 7 (2025) 180, https://doi.org/10.1007/s42452-025-06587-0.

[51] Y. Huang, W. Li, F. Macheret, R.A. Gabriel, L. Ohno-Machado, A tutorial on calibration measurements and calibration models for clinical prediction models, J. Am. Med. Inform. Assoc. 27 (2020) 621–633, https://doi.org/10.1093/jamia/ocz228.

[52] F. Harrell, Is medicine mesmerized by machine learning? Stat. Think. (2018). ⟨https://www.fharrell.com/post/medml/⟩ (accessed May 25, 2025).

[53] T. Dimitriadis, T. Gneiting, A.I. Jordan, Stable reliability diagrams for probabilistic classifiers, Proc. Natl. Acad. Sci. USA 118 (2021) e2016191118, https://doi.org/10.1073/pnas.2016191118.

[54] J. Schwarz, D. Heider, GUESS: projecting machine learning scores to well-calibrated probability estimates for clinical decision-making, Bioinformatics 35 (2019) 2458–2465, https://doi.org/10.1093/bioinformatics/bty984.

[55] M. Kelly, P. Smyth, Variable-Based Calibration for Machine Learning Classifiers, (2023). ⟨https://doi.org/10.48550/arXiv.2209.15154⟩.

[56] D.-B. Wang, L. Feng, M.-L. Zhang, Rethinking calibration of deep neural networks: do not be afraid of overconfidence, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, J.W. Vaughan (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc, 2021, pp. 11809–11820, in: ⟨https://proceedings.neurips.cc/paper_files/paper/2021/file/61f3a6dbc9120ea78ef75544826c814e-Paper.pdf⟩.

[57] S.G. Finlayson, A. Subbaswamy, K. Singh, J. Bowers, A. Kupke, J. Zittrain, I. S. Kohane, S. Saria, The clinician and dataset shift in artificial intelligence, N. Engl. J. Med 385 (2021) 283–286, https://doi.org/10.1056/NEJMc2104626.

[58] R.D. Riley, K.I.E. Snell, L. Archer, J. Ensor, T.P.A. Debray, B. van Calster, M. van Smeden, G.S. Collins, Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study, BMJ 384 (2024) e074821, https://doi.org/10.1136/bmj-2023-074821.

[59] R. Bommasani D.A. Hudson E. Adeli R. Altman S. Arora S. von Arx M.S. Bernstein J. Bohg A. Bosselut E. Brunskill E. Brynjolfsson S. Buch D. Card R. Castellon N. Chatterji A. Chen K. Creel J.Q. Davis D. Demszky C. Donahue M. Doumbouya E. Durmus S. Ermon J. Etchemendy K. Ethayarajh L. Fei-Fei C. Finn T. Gale L. Gillespie K. Goel N. Goodman S. Grossman N. Guha T. Hashimoto P. Henderson J. Hewitt D.E. Ho J. Hong K. Hsu J. Huang T. Icard S. Jain D. Jurafsky P. Kalluri S. Karamcheti G. Keeling F. Khani O. Khattab P.W. Koh M. Krass R. Krishna R. Kuditipudi A. Kumar F. Ladhak M. Lee T. Lee J. Leskovec I. Levent X.L. Li X. Li T. Ma A. Malik C.D. Manning S. Mirchandani E. Mitchell Z. Munyikwa S. Nair A. Narayan D. Narayanan B. Newman A. Nie J.C. Niebles H. Nilforoshan J. Nyarko G. Ogut L. Orr I. Papadimitriou J.S. Park C. Piech E. Portelance C. Potts A. Raghunathan R. Reich H. Ren F. Rong Y. Roohani C. Ruiz J. Ryan C. Ré D. Sadigh S. Sagawa K. Santhanam A. Shih K. Srinivasan A. Tamkin R. Taori A.W. Thomas F. Tramèr R.E. Wang W. Wang B. Wu J. Wu Y. Wu S.M. Xie M. Yasunaga J. You M. Zaharia M. Zhang T. Zhang X. Zhang Y. Zhang L. Zheng K. Zhou P. Liang Oppor. Risks Found. Models 2021 doi: 10.48550/ARXIV.2108.07258.

[60] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C.L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback. Proceedings of the 36th International Conference on Neural Information Processing Systems, Curran Associates Inc,, Red Hook, NY, USA, 2022, pp. 27730–27744.

[61] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners. Proceedings of the 34th International Conference on Neural Information Processing Systems, Curran Associates Inc,, Red Hook, NY, USA, 2020, pp. 1877–1901.

[62] A. Kirillov E. Mintun N. Ravi H. Mao C. Rolland L. Gustafson T. Xiao S. Whitehead A.C. Berg W.-Y. Lo P. Dollár R. Girshick Segm. Anything 2023 doi: 10.48550/arXiv.2304.02643.

[63] J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang, Segment anything in medical images, Nat. Commun. 15 (2024) 654, https://doi.org/10.1038/s41467-024-44824-z.

[64] M. Oquab T. Darcet T. Moutakanni H. Vo M. Szafraniec V. Khalidov P. Fernandez D. Haziza F. Massa A. El-Nouby M. Assran N. Ballas W. Galuba R. Howes P.-Y. Huang S.-W. Li I. Misra M. Rabbat V. Sharma G. Synnaeve H. Xu H. Jegou J. Mairal P. Labatut A. Joulin P. Bojanowski DINOv2 learn. robust. vis. features superv. 2024 doi: 10.48550/arXiv.2304.07193.

[65] S. Pai I. Hadzic D. Bontempi K. Bressem B.H. Kann A. Fedorov R.H. Mak H.J.W.L. Aerts Vis. found. models comput. tomogr. 2025 doi: 10.48550/ARXIV.2501.09001.

[66] F. Pérez-García, H. Sharma, S. Bond-Taylor, K. Bouzid, V. Salvatelli, M. Ilse, S. Bannur, D.C. Castro, A. Schwaighofer, M.P. Lungren, M.T. Wetscherek, N. Codella, S.L. Hyland, J. Alvarez-Valle, O. Oktay, Exploring scalable medical image encoders beyond text supervision, Nat. Mach. Intell. 7 (2025) 119–130, https://doi.org/10.1038/s42256-024-00965-w.

[67] I.E. Hamamci S. Er C. Wang F. Almas A.G. Simsek S.N. Esirgun I. Doga O.F. Durugol W. Dai M. Xu M.F. Dasdelen B. Wittmann T. Amiranashvili E. Simsar M. Simsar E.B. Erdemir A. Alanbay A. Sekuboyina B. Lafci C. Bluethgen K. Batmanghelich M.K. Ozdemir B. Menze Dev. Gen. Found. Models a Multimodal Dataset 3D Comput. Tomogr. 2024 doi: 10.48550/ARXIV.2403.17834.

[68] L. Blankemeier J.P. Cohen A. Kumar D.V. Veen S.J.S. Gardezi M. Paschali Z. Chen J.-B. Delbrouck E. Reis C. Truyts C. Bluethgen M.E.K. Jensen S. Ostmeier M. Varma J.M.J. Valanarasu Z. Fang Z. Huo Z. Nabulsi D. Ardila W.-H. Weng E.A. Junior N. Ahuja J. Fries N.H. Shah A. Johnston R.D. Boutin A. Wentland C.P. Langlotz J. Hom S. Gatidis A.S. Chaudhari Merlin a vis. lang. found. model 3d comput. tomogr. 2024 doi: 10.48550/arXiv.2406.06512.

[69] E. Tiu, E. Talius, P. Patel, C.P. Langlotz, A.Y. Ng, P. Rajpurkar, Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning, Nat. Biomed. Eng. 6 (2022) 1399–1406, https://doi.org/10.1038/s41551-022-00936-9.

[70] Z. Chen, M. Varma, J. Xu, M. Paschali, D. Van Veen, A. Johnston, A. Youssef, L. Blankemeier, C. Bluethgen, S. Altmayer, J.M.J. Valanarasu, M.S.E. Muneer, E.P. Reis, J.P. Cohen, C. Olsen, T.M. Abraham, E.B. Tsai, C.F. Beaulieu, J. Jitsev, S. Gatidis, J.-B. Delbrouck, A.S. Chaudhari, C.P. Langlotz, A Vision-Language Foundation Model to Enhance Efficiency of Chest X-ray Interpretation, (2024). ⟨https://doi.org/10.48550/ARXIV.2401.12208⟩.

[71] N. Deperrois H. Matsuo S. Ruipérez-Campillo M. Vandenhirtz S. Laguna A. Ryser K. Fujimoto M. Nishio T.M. Sutter J.E. Vogt J. Kluckert T. Frauenfelder C. Blüthgen F. Nooralahzadeh M. Krauthammer RadVLM a multitask. conversat. vis. lang. model radiol. 2025 doi: 10.48550/ARXIV.2502.03333.

[72] L. Zhao, X. Chen, E.Z. Chen, Y. Liu, T. Chen, S. Sun, Retrieval-augmented Few-shot, Med. Image Segm. Found. Models (2024), https://doi.org/10.48550/ARXIV.2408.08813.

[73] T. Lee, H. Tu, C.H. Wong, W. Zheng, Y. Zhou, Y. Mai, J. Roberts, M. Yasunaga, H. Yao, C. Xie, others, Vhelm: a holistic evaluation of vision language models, Adv. Neural Inf. Process. Syst. 37 (2024) 140632–140666.

[74] T. Lee, M. Yasunaga, C. Meng, Y. Mai, J.S. Park, A. Gupta, Y. Zhang, D. Narayanan, H. Teufel, M. Bellagente, others, Holistic evaluation of text-to-image models, Adv. Neural Inf. Process. Syst. 36 (2023) 69981–70011.

[75] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L.A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, O. Vinyals, J.W. Rae, L. Sifre, Training compute-optimal large language models. Proceedings of the 36th International Conference on Neural Information Processing Systems, Curran Associates Inc,, Red Hook, NY, USA, 2022, pp. 30016–30030.

[76] J. Zhu A. Hamdi Y. Qi Y. Jin J. Wu Med. SAM 2 Segm. Med. Images Video via Segm. Anything Model 2 2024 doi: 10.48550/ARXIV.2408.00874.

[77] G. Alain Y. Bengio Underst. Intermed. layers Using Linear Classif. probes 2018 doi: 10.48550/arXiv.1610.01644.

[78] J. Hessel A. Holtzman M. Forbes R.L. Bras Y. Choi CLIPScore A Ref. Free Eval. Metr. Image Captioning 2021 doi: 10.48550/ARXIV.2104.08718.

[79] Z. Guo, R. Jin, C. Liu, Y. Huang, D. Shi, Supryadi, L. Yu, Y. Liu, J. Li, B. Xiong, D. Xiong, Eval. Large Lang. Model. A Compr. Surv. (2023), https://doi.org/10.48550/arXiv.2310.19736.

[80] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[81] L. Chin-Yew, Rouge: A package for automatic evaluation of summaries. in: Proceedings of the Workshop on Text Summarization Branches Out, 2004.

[82] A.R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, D. Radev, Summeval: Re-evaluating summarization evaluation, Trans. Assoc. Comput. Linguist. 9 (2021) 391–409.

[83] S.F. Chen D. Beeferman R. Rosenfeld. Eval. Metr. Lang. Models 1998.

[84] D. Jurafsky, J. Martin, Machine translation, speech and language processing, Prentice-Hall,, Englewood Cliffs, 2000.

[85] T. Zhang V. Kishore F. Wu K.Q. Weinberger Y. Artzi BERTScore Eval. Text. Gener. BERT 2020 doi: 10.48550/arXiv.1904.09675.

[86] W. Yuan, G. Neubig, P. Liu, Bartscore: evaluating generated text as text generation, Adv. Neural Inf. Process. Syst. 34 (2021) 27263–27277.

[87] J. Novikova, O. Dušek, A.C. Curry, V. Rieser, Why we need new evaluation metrics for NLG, Proc. 2017 Conf. Empir. Methods Nat. Lang. Process. (2017) 2231–2240, https://doi.org/10.18653/v1/D17-1237.

[88] Y. Wang J. Deng A. Sun X. Meng Perplexity PLM Is. Unreliable Eval. Text. Qual. 2023 doi: 10.48550/arXiv.2210.05892.

[89] A. Pradhan, K. Todi, Understanding large language model based metrics for text summarization, Proc. 4th Workshop Eval. Comp. NLP Syst. (2023) 149–155.

[90] A.B. Sai, A.K. Mohankumar, M.M. Khapra, A survey of evaluation metrics used for NLG systems, ACM Comput. Surv. (CSUR) 55 (2022) 1–39.

[91] R. Osuala, K. Kushibar, L. Garrucho, A. Linardos, Z. Szafranowska, S. Klein, B. Glocker, O. Diaz, K. Lekadir, Data synthesis and adversarial networks: a review and meta-analysis in cancer imaging, Med. Image Anal. 84 (2023) 102704, https://doi.org/10.1016/j.media.2022.102704.

[92] Zhou Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process 13 (2004) 600–612, https://doi.org/10.1109/TIP.2003.819861.

[93] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Salt Lake City, UT, 2018, pp. 586–595, https://doi.org/10.1109/CVPR.2018.00068.

[94] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, X. Chen, Improved techniques for training GANs, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc, 2016, in: ⟨https://proceedings.neurips.cc/paper_files/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf⟩.

[95] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local nash equilibrium. in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc,, Red Hook, NY, USA, 2017, pp. 6629–6640.

[96] R. Osuala, D.M. Lang, P. Verma, S. Joshi, A. Tsirikoglou, G. Skorupko, K. Kushibar, L. Garrucho, W.H.L. Pinaya, O. Diaz, J.A. Schnabel, K. Lekadir, Towards learning contrast kinetics with multi-condition latent diffusion models, in: M.G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, J. A. Schnabel (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, Springer Nature Switzerland, Cham, 2024, pp. 713–723.

[97] M.J. Chong, D. Forsyth, Effectively unbiased FID and inception score and where to find them. in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[98] T. DeVries, A. Romero, L. Pineda, G.W. Taylor, M. Drozdzal, Eval. Cond. GANs (2019), https://doi.org/10.48550/arXiv.1907.08175.

[99] A. Borji, Pros and cons of GAN evaluation measures: new developments, Comput. Vis. Image Underst. 215 (2022) 103329, https://doi.org/10.1016/j.cviu.2021.103329.

[100] L. Garrucho, K. Kushibar, R. Osuala, O. Diaz, A. Catanese, J. del Riego, M. Bobowicz, F. Strand, L. Igual, K. Lekadir, High-resolution synthesis of high-density breast mammograms: application to improved fairness in deep learning based mass detection, Front Oncol. 12 (2023) 1044496, https://doi.org/10.3389/fonc.2022.1044496.

[101] N. Wolfrath, J. Wolfrath, H. Hu, A. Banerjee, A.N. Kothari, Stronger Baseline Models – A Key Requirement for Aligning Machine Learning Research with Clinical Utility, (2024). ⟨https://doi.org/10.48550/ARXIV.2409.12116⟩.
[102] T.J. Bradshaw, Z. Huemann, J. Hu, A. Rahmim, A guide to cross-validation for artificial intelligence in medical imaging, Radiol. Artif. Intell. 5 (2023) e220232, https://doi.org/10.1148/ryai.220232.
[103] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[104] M. Sivakumar, S. Parthasarathy, T. Padmapriya, Trade-off between training and testing ratio in machine learning for medical image processing, PeerJ Comput. Sci. 10 (2024) e2245, https://doi.org/10.7717/peerj-cs.2245.
[105] C.J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, D. King, Key challenges for delivering clinical impact with artificial intelligence, BMC Med 17 (2019) 195, https://doi.org/10.1186/s12916-019-1426-2.
[106] E. Reiter, A structured review of the validity of BLEU, Comput. Linguist. 44 (2018) 393–401, https://doi.org/10.1162/coli_a_00322.
[107] R. Longjohn G. Gopalan E. Casleton Stat. Uncertain. Quantif. Aggreg. Perform. Metr. Mach. Learn. Benchmarks 2025 doi: 10.48550/ARXIV.2501.04234.
[108] C. Drummond, N. Japkowicz, Warning: statistical benchmarking is addictive. kicking the habit in machine learning, J. Exp. Theor. Artif. Intell. 22 (2010) 67–80, https://doi.org/10.1080/09528130903010295.
[109] G. Cumming. Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis, 1st ed..,, Routledge,, 2013, https://doi.org/10.4324/9780203807002.
[110] N. Japkowicz, Z. Boukouvalas, Machine learning evaluation: towards reliable and responsible AI, Cambridge University Press,, 2024.
[111] P.M. Bossuyt, J.B. Reitsma, K. Linnet, K.G. Moons, Beyond diagnostic accuracy: the clinical utility of diagnostic tests, Clin. Chem. 58 (2012) 1636–1643, https://doi.org/10.1373/clinchem.2012.182576.
[112] A.J. Vickers, B. Van Calster, E.W. Steyerberg, Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests, BMJ (2016) i6, https://doi.org/10.1136/bmj.i6.
[113] V.X. Liu, D.W. Bates, J. Wiens, N.H. Shah, The number needed to benefit: estimating the value of predictive analytics in healthcare, J. Am. Med. Inform. Assoc. 26 (2019) 1655–1659, https://doi.org/10.1093/jamia/ocz088.
[114] A.J. Vickers, E.B. Elkin, Decision curve analysis: a novel method for evaluating prediction models, Med Decis. Mak. 26 (2006) 565–574, https://doi.org/10.1177/0272989X06295361.
[115] A.J. Vickers, B. Van Calster, E.W. Steyerberg, A simple, step-by-step guide to interpreting decision curve analysis, Diagn. Progn. Res 3 (2019) 18, https://doi.org/10.1186/s41512-019-0064-7.
[116] K.F. Kerr, M.D. Brown, K. Zhu, H. Janes, Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use, JCO 34 (2016) 2534–2540, https://doi.org/10.1200/JCO.2015.65.5654.
[117] A. Vickers. Seven. common errors decis. curve anal. stat. think. 2023.(accessed May 19, 2025).⟨https://www.fharrell.com/post/edca/⟩.