Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluation Metrics for Image Classification, Segmentation and Localization Mathematical Foundations, Comparative Analysis and Research Recommendations

Behnam Kiani Kalejahi

b.kiani@centralasian.uz

Central Asian University

Mohammad Javad Rajabi[2]

INTI International University

Systematic Review

Keywords: image classification, semantic segmentation, instance segmentation, object detection, landmark localization, performance metrics, calibration, uncertainty, statistical validation, medical imaging

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11366673/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

CRITICAL REVIEW AND REPRODUCIBLE METHODS RESOURCE

Evaluation Metrics for Image Classification, Segmentation and Localization

Mathematical Foundations, Comparative Analysis and Research Recommendations

Behnam Kiani Kalejahi[1,2], Mohammad Javad Rajabi[2]

[1]Department of Computer Science, School of Engineering, Central Asian University, Tashkent, Uzbekistan
[2]Faculty of Data Science and Information Technology, INTI International University, Putra Nilai, Malaysia
Corresponding author: Behnam Kiani Kalejahi (b.kiani@centralasian.uz), ORCID 0000-0002-7118-0382

Abstract

Evaluation metrics dictate the incentives for an image analysis system, detect remaining failures, and verify any proposed improvement in scientific or clinical claims. Despite the mathematical similarity of classification, segmentation, and localization, the units, matching criteria, and practice implications vary across these tasks. The current review analyzes the distinctions through the lens of a selected sample of 150 bibliographically validated references, including 10 papers from the provided author profile. It incorporates foundational statistics, state-of-the-art benchmarking practices, medical imaging use cases, and recent developments in uncertainties, selective prediction, and deployment. The review organizes 72 metrics and corresponding evaluation quantities by the question they answer: correctness of decision, score discrimination, reliability of probabilities, spatial correspondence, recovery of objects, localization of geometry, or practical utility. Comparisons describe equations, assumptions, aggregation approaches, advantages, and failures. Ten deterministic worked examples demonstrate prevalence effects, constant discrimination and worsening probability scores, reverse rankings because of aggregation, hidden border failures by overlap, detection matching, physical landmark errors, and the cost of abstaining. Twenty-four numeric instances from 4 separate publications supply additional citations and examples without pooling the inconsistent results. The primary takeaway is that a sound evaluation requires a small and preselected set of metrics, complementary to each other, applicable to the defined population, unit, reference standard, and decision. For classification, operating point performance is always required, along with discrimination and probability metrics. Segmentation demands appropriate measures of overlap, borders, and object topology. Localization needs explicitly stated criteria of correspondence and tolerance and false positive control. Independent partitioning, uncertainties, subgroups, and workflow outputs determine the interpretability of the metrics. This manuscript is an organized critical review and methodology resource, but not a completed systematic review according to PRISMA criteria, meta-analysis, or benchmark of a new model. The attached list of metrics, reference export, extraction notes, computational code, and reporting checklist help design and validate reproducible studies and author reviews.

Keywords: image classification; semantic segmentation; instance segmentation; object detection; landmark localization; performance metrics; calibration; uncertainty; statistical validation; medical imaging.

1 Introduction

Image analysis has progressed from hand-designed features to learned architectures that provide labels, masks, bounding boxes, key points, and probability distributions. Still, there is always a possibility that a model will be tested against an improper target. High image-level accuracy does not imply a good performance on the rare diagnoses. High foreground overlap may go along with a disconnected vessel or distant false-positive lesion. Accurate coordinates of successfully localized landmarks may conceal a large proportion of total failure. These problems result from the relationship between the task and its measurement, not any architecture choice. [1, 2]

An evaluation metric is a number that summarizes a certain comparison. The meaning is added to the summary after the observation; the reference and aggregation procedure are described. For example, recall may be measured per positive patients, per foreground voxels, lesions or landmarks – the denominators are different. The scientific claim is incomplete if it mentions recall without a unit. A review of classification, segmentation, and localization tasks requires a unified evaluation language while keeping the differences. [3, 4, 5]

Current systems make task separation even more relevant. Self-configurable segmentation pipelines combine the architecture with preprocessing, sampling, and postprocessing. Promptable models provide masks conditioned on the user information. Instance models recognize the position, identity, and pixel membership of the object. Evaluation should cover the entire prediction process that includes the input information. [6, 7, 8, 9]

This review aims to achieve three interrelated goals. First, to explain the meaning of commonly used metrics mathematically and practically. Second, compare different options under realistic conditions of class imbalance, annotation uncertainty and partial prediction. Third, to transfer those comparisons into guidelines for study design, statistics and reporting. Medical imaging serves as a primary application context; well-established benchmarks from general computer vision provide methods for evaluation of boxes, instances and key points. No context offers a universally optimal metric.

The contribution is a cross-task review with traceable references and reproducible examples. Ten worked examples are calculated rather than based on patient data. Existing examples are cited in their original context. Proposed checklist serves as a reporting tool and was not validated by consensus. The boundaries allow distinguishing mathematical illustrations, source-reported observations and methodological recommendations throughout the manuscript.

2 Review methods and verification

Scope and evidence collection

The evidence snapshot date is 19 September 2026. The collection includes metric literature, statistical evaluation methods, probability scoring and calibration, image benchmarking protocols, applications in medicine, reporting guidelines, and some recently published developments. The inclusion criterion of this systematic review has been relevance to the topic of classification, segmentation, or localization task measurement. Architecture papers have been included exclusively in cases where there was an evaluation context, such as conditional prompting or detection instance segmentation. The total number of references has been set at the requested 150 references; it does not mark the end of comprehensive eligibility selection but works as a curating coverage goal instead.

To retrieval, 99 title or identifier targeted searches have been used, along with 10 title corrective searches, 30 verifications of the original publications at arXiv page level, and direct verification of the original journal or conference pages. Relevant records from a previously verified segmentation reference collection have been re-assessed and reused with their verification source. Four recent Europe PubMed Central query families covered metric evaluation, localization, calibration, and supplied author surname.

For each recent query, the publication dates have been limited to 1 January 2024 to 19 September 2026, and results – to 20 returned records each. The returned record numbers, caps, and query strings are preserved in the evidence package.

Such procedures form the basis for systematic review. They are not comprehensive coverage of all database sources, languages, or publication places. Due to Google Scholar access rate limit, the supplied profile publication list has been used for searching candidates. Authorship and publication identity of retrieved records has been verified against the candidate publications on that list. The surname-only recent search has produced several unrelated authors, which are not considered matches. Profile membership alone was insufficient for inclusion.

What bibliographic verification establishes

Verification involved ensuring that the title, authorship and DOI/canonical page of publication aligned. Metadata was obtained from the Crossref, Europe PMC, original proceedings page and arXiv metadata repositories. Titles and identifiers that led to different papers were either rejected or corrected. Duplicate canonical identifiers were also considered. Where there was a discrepancy between the print volume date and the online-first date, the respective citation style was captured. Indexed conference proceedings which were treated as journal articles were relabeled when the publication form was known.

The resulting corpus contains 100 journal articles, 20 conference papers, 7 conference/book chapters, 22 explicitly labeled preprint papers, and one conference paper captured via its arXiv version. The labels used here simply refer to the versions used in this analysis. They do not in any way indicate that no other versions of the paper exist. Verifying the bibliographic details only identifies the paper as the source but does not confirm the validity of the study, verify any errors or reproduce any experiment. This is different from the verification of the entire text and supplementary data.

Targeted full text retrieval provided metric data and published exemplars. Not all the papers had full-text access. Twenty-four quantitative records were extracted from four source papers where full text was available, capturing the table, label of the metric, units and the reported uncertainty. The extraction process was assisted by an AI system but not independently verified. Ambiguities in the sources were recorded without making silent corrections. The authors need to verify the records against the sources prior to submission.

Synthesis and reproducible outputs

The synthesis groups metrics by their estimand and necessary inputs, rather than by popularity or the year of a model. Comparative tables summarize the conditions under which alternatives answer different questions. The metric catalog records formulas, direction of improvement, interpretation, limitations and supporting references. Ten deterministic examples are generated by an accompanying Python script from explicit synthetic arrays or confusion counts. No participant data were collected and no models were trained.

A completed systematic review would require a prespecified eligibility protocol, reproducible database exports, complete deduplication, independent screening and extraction, and a defensible appraisal

strategy. PRISMA and PRISMA-S provide reporting structures for that work; they should not be invoked to imply that those procedures occurred here. No PRISMA flow counts, registration, ethics approval or formal risk-of-bias ratings have been invented. [10, 11]

3 Task definitions and evaluation units

Classification segmentation and localization

Classification assigns one or more labels to an image, examination or patient. Binary classification specifies one positive event. Multiclass classification usually assigns one label among mutually exclusive categories. Multilabel classification permits several labels simultaneously. An ordinal task adds a meaningful order, so a one-category error may have a different consequence from a much larger error. The label ontology determines which comparisons are valid. [3, 12]

Semantic segmentation assigns a class to each pixel or voxel. Instance segmentation additionally distinguishes individual objects of the same class. Panoptic segmentation combines these ideas by representing individual countable objects and amorphous background categories in a single output. The two-class foreground measure ignores this difference: the combination of three neighboring nuclei might have a lot of foreground intersection but at the same time change the object number. Localization goes beyond object detection. A system could localize a lesion through a bounding box, a point for a suspicious location, an anatomical landmark, or transform for alignment of images. Detection involves the counting of objects and usually their types. Landmark localization could be based on a pre-defined set of anatomical identities. No detection and landmark offset are accounted for separately although both produce coordinates. [16, 17, 18, 19]

Table 1. Task definitions and the unit that must be preserved during evaluation.

Task	Typical output	Principal evaluation unit and requirement
Classification	Class or class probabilities	Image, examination or patient; define event and threshold
Semantic segmentation	Labeled pixels or voxels	Case and class; preserve geometry and empty-target policy
Instance segmentation	Separate labeled objects	Object plus mask; define correspondence, splits and merges
Box or point detection	Scored locations and categories	Eligible object; count misses, duplicates and false alarms
Landmark localization	Named coordinates	Corresponding landmark and subject; physical tolerance and failures
Image registration	Spatial transformation	Independent target point; distinguish fit error from target error

The unit belongs in the claim

The unit of prediction, unit of evaluation and unit of statistical independence need not be identical. A model can predict a probability for each frame, derive a result for each examination, and be tested on

patients who each contribute several examinations. Pooling frames answers a frame-weighted question and gives longer videos greater influence. Treating every frame as an independent statistical observation also understates uncertainty when frames are correlated.

A similar issue occurs with bilateral anatomy. Side-level evaluation may be appropriate for a side-specific intervention, whereas a patient-level outcome may be appropriate for referral. Labeling systems and criteria for aggregating them could generate different estimates but do not necessarily mean that either of them is wrong in their own right. A study involving mammographies reveals the effect of the level of assessment that could be performed but does not necessarily suggest the same numeric differences in all cases and cannot be taken as a general correction factor. [20, 21]. The review uses TP, FP, FN, and TN as abbreviations for true positives, false positives, false negatives, and true negatives, respectively. N denotes the number of evaluated units; K denotes the number of classes or labels. P and G denote predicted and reference foreground sets. Probabilities are denoted by p, binary outcomes by y, and decision or matching thresholds by their stated value. Unless otherwise specified, geometric distances are measured after transformation into physical coordinates.

4 The measurement framework

Estimands and reference standards

An estimand is the quantity a study intends to estimate. In an imaging study it includes the population, target event, prediction procedure, reference standard and reduction over observations. Mean organ Dice per patient and Dice after pooling all organ voxels are different estimands. Lesion sensitivity at one false-positive mark per scan differs from sensitivity at a confidence threshold chosen for maximum F1. A precise estimand should be written before selecting software functions.

Reference standards are observations made under a process, not necessarily error-free truth. Pathology, follow-up, reader consensus and a single annotation can have different limitations. In segmentation, boundary disagreement may reflect genuine ambiguity, image quality or inconsistent instructions. STAPLE illustrates statistical fusion of multiple annotations, but a fused mask does not remove the need to describe annotators, assumptions and uncertainty. [22, 2]

For landmark studies, inter-reader variability should be measured in the same coordinates and units as algorithm error. For classification, uncertain or indeterminate reference categories need an explicit inclusion policy. Removing every difficult case can make a classifier appear more reliable while changing the population it addresses. Where labels are incomplete, an unannotated object should not automatically be treated as a verified negative without a justified protocol.

Thresholds and matching rules

The setting of a threshold is necessary for many measures. The threshold might transform a probability into a class, determine that two bounding boxes are matching, indicate an admissible distance for boundaries, or determine those instances to be sent to a human. Different questions are answered by the different thresholds. A confidence threshold of 0.5 and an IoU threshold of 0.5 are not the same idea. One deals with the score of the detector; the other one deals with geometric accuracy relative to the reference.

Threshold tuning should be performed on development sets or a distinct calibration set. Optimal threshold tuning based on test labels turns the test into a selection tool. The same holds true for selecting the measure after having seen which measure favors a particular approach. Nested evaluation is useful if model selection and performance evaluation would otherwise leak information. The design of nested evaluation must respect patient or subject populations. [23, 24] In addition to this, detection and instance segmentation need a correspondence scheme. The evaluator needs to define class compatibility, one-to-one mapping, tolerance, treatment of duplicates, ignored areas, and bounds for predictions. All this affects both the numerator and the denominator of the measure. Reporting the measure alone is not sufficient [5, 13].

A complementary metric set

A realistic assessment will select an appropriate primary measure, together with a small number of ancillary measures. The ancillary measures should be able to characterize different failure modes instead of just repeating the same thing under different names. Accuracy and top 1 error are complementary measures in the sense of arithmetic. Binary Dice and foreground F1 computed over the same units will always have the same value. Dice and IoU are similar overlap measures, but a physical boundary measure can help detect another failure mode.

Not all-important aspects of a model are captured by any single scalar measure. A model can have increased sensitivity with increased false positive burden, improved overlap even when allowing for a clinically significant deviation from ground truth or maintained accuracy by looking at more samples. Direct illustration of these tradeoffs is clearly necessary. Any aggregated measure should be justified before seeing the outcomes [1, 25].

5 Classification at a fixed operating point

Accuracy sensitivity specificity and predictive values

The accuracy is measured as the fraction of decisions that agree with the reference standard. The interpretation of accuracy is simple where the class distribution represented and the cost of making decisions is relevant to the situation being addressed. The accuracy measure may be problematic in situations where a dominant class allows a classifier to have high accuracy without detecting the target event. For example, suppose we had a population where positives were only one percent; the always negative classifier would score 99 percent in terms of accuracy but no sensitivity at all. The confusion matrix shows this clearly. Both should be accompanied by the number of eligible cases in their denominators. A point estimate of 100 percent sensitivity based on three positive cases provides much less information than the same estimate based on several hundred independent positives. Precision and negative predictive value instead condition on the prediction, so their denominators represent what a user encounters after a positive or negative output. [26, 3, 27]

Predictive values change with prevalence even when class-conditional performance is unchanged. For positive prevalence pi, sensitivity Se and specificity Sp, Bayes' rule gives PPV = Se pi / [Se pi + (1 - Sp)(1 - pi)]. This is useful for understanding why an enriched research dataset cannot directly supply the

expected precision in a screening population. The calculation assumes that sensitivity and specificity transport to the new setting, which may fail under acquisition or case-mix shift.

$$\text{Accuracy} = (TP + TN) / N \quad \text{Sensitivity} = TP / (TP + FN)$$

$$\text{Specificity} = TN / (TN + FP) \quad \text{Precision} = TP / (TP + FP)$$

$$F1 = 2TP / (2TP + FP + FN) \quad \text{Balanced accuracy} = (\text{Sensitivity} + \text{Specificity}) / 2$$

Balanced accuracy F scores and correlation

Balanced accuracy gives equal weight to class recalls. In a binary task it is the mean of sensitivity and specificity; in a multiclass task it is commonly the mean class recall. This avoids weighting directly by the empirical class frequency. It does not, however, establish that equal class weighting represents the relevant clinical or operational costs.

F1 is the harmonic mean of precision and recall and can be written as 2TP divided by 2TP + FP + FN. True negatives do not enter this expression. This makes it useful when recovery of a positive class and the reliability of positive predictions are central, but it also makes F1 unsuitable as a complete description of a binary decision system. The threshold maximizing F1 depends on the score distribution and the evaluated batch. It should not be assumed to define the best intervention threshold. [28, 29]

The Matthews correlation coefficient uses all four confusion counts and has a binary range from minus one to one. It offers a concise measure of association with generalizations to multiclass problems. The denominator becomes zero if any of the marginals needed in the formula are zero. Even though the software may adopt some convention in this case, it must be clear in its report that this convention is separate from what constitutes mathematically defined correlation. High values of MCC do not compensate for class-specific errors or consequences. [30, 31, 32]

The Youden J index is simply sensitivity plus specificity minus one. It puts emphasis on equal separation between classes. Likelihood ratio is the change in odds of the event based on a positive or negative test result in the considered conditions. Neither gives posterior probability of the event without incorporating base-rate. The choice among these statistics will depend on the decision setting rather than the fact that any of them are inherently robust to imbalance.

Comparing errors in practice

In the context of a triage classifier, sensitivity could be crucial; nevertheless, it is imperative that the downstream team can check false positives. Sensitivity should be reported at a pre-defined level of specificity or vice versa, depending on whether it is pertinent to the intended use case. In terms of a confirmatory classifier, precision and false-positive actions could be of importance. The above points give examples in designing the studies but not absolute clinical cutoffs.

It is necessary for the confusion matrix to maintain distinct subclasses. The consolidation of different biological diseases under one positive group would provide a good binary outcome, but it would also hide problems related to the rare subclass. It is necessary to make subgroup reporting possible and reflect

potential mechanisms of failure and provide sufficient sample size. Post hoc identification is a way of generating hypotheses, but it needs verification [33, 34].

Table 2. Choosing a fixed-threshold classification summary. Complements are part of interpretation, not interchangeable alternatives.

Metric	Useful question	Required complement or caution
Accuracy	How many decisions are correct in this population?	Class prevalence and confusion counts
Sensitivity	How many eligible positives are recovered?	Specificity or false-alert workload
Precision	How trustworthy are positive predictions?	Recall and represented prevalence
Balanced accuracy	How do recalls perform with equal class weights?	Class-specific recalls and decision costs
F1	How do precision and recall combine for a target class?	TN behavior, threshold and averaging
MCC	How strongly do labels and predictions agree?	Marginal counts and zero-denominator convention
Weighted kappa	How severe is disagreement on an ordered scale?	Weight matrix, confusion table and class marginals

6 Ranking metrics and precision recall analysis

ROC curves and AUROC

A receiver operating characteristic curve traces sensitivity against the false-positive rate as a score threshold varies. AUROC can be interpreted as the probability that a randomly selected positive has a higher score than a randomly selected negative, with half credit for ties. It summarizes discrimination rather than the correctness of a particular decision threshold. An AUROC of 0.90 does not mean that 90 percent of patients will be classified correctly. [35, 36]

The ROC representation conditions separately on the positive and negative classes. Under a pure prevalence change with stable class-conditional score distributions, the ROC curve remains the same. Real deployment shifts rarely guarantee these assumptions. Differences in severity, device, referral pathway or annotation can alter score distributions within each class and change AUROC. A high internal AUROC therefore cannot establish transportability.

An overall area also averages across operating regions that may be irrelevant to use. Partial AUROC can focus on a prespecified false-positive interval, but bounds and normalization must be stated. An unnormalized partial area and a standardized partial area have different scales. Hand's analysis of AUC highlights the need to make decision assumptions explicit rather than treating all threshold summaries as context-free measures of usefulness. [37]

Precision recall curves and average precision

Precision-recall analysis conditions directly on positive predictions and recovered positives. It is often informative when the positive class is rare and false-alert burden matters. Its baseline and magnitude depend on prevalence, so precision-recall results from differently enriched datasets should not be compared without qualification. The relationship between ROC and precision-recall curves is defined by the same confusion counts but changes with the class distribution. [38, 39]

Average precision is commonly a sum of precision values weighted by increases in recall. A trapezoidal area under sampled precision-recall points is a different numerical procedure. Object-detection evaluators can additionally use an interpolated precision envelope, fixed recall samples, class averaging and geometric matching. Consequently, the labels AP and AUPRC do not identify a unique estimator. The evaluator, interpolation rule and handling of ties belong in the methods. [38, 29, 5]

Precision-recall-gain curves provide another representation intended to address properties of conventional precision-recall analysis. Their existence reinforces a general point: a curve's axes and integration rule express a choice about what is being averaged. A review should explain that choice rather than place every area-like quantity on a common performance scale. [29]

Multiclass discrimination

Multiclass AUROC requires a reduction from several classes to binary comparisons. One-versus-rest compares a class with the mixture of all remaining classes. One-versus-one compares pairs, and Hand and Till describe a multiclass generalization based on pairwise discrimination. The following steps involve macros, weighted, and micro aggregations, followed by introducing more weighting. Without these details, the multiclass AUC reported will be inadequate for replication [40]. Class-based AUCs and class-based supports help interpret an aggregation. In the one versus rest approach, the negative class consists of a mix that changes with different data sets. In micro aggregation, the aggregation will be driven by more common classes and possibly a huge number of negative class labels. In macro aggregation, all classes receive the same weighting; however, this will not be reliable if some of the classes have very few independent instances.

7 Multiclass multilabel and ordinal evaluation

Averaging is a scientific choice

Generally speaking, macro F1 is the arithmetic average of per-class F1 scores. There have been instances where the harmonic average of macro precision and macro recall is used. It should be noted that these two operations are generally non-commutative and therefore may lead to divergence of values. Explicitly mentioning the way the calculation is done becomes very important especially in cases where one compares a paper and a software which uses a different definition for the same term [41]. Micro F1 calculates label decisions before calculating the F1 score. Under exhaustive and single-label multiclass classifications, every mistake will lead to one false positive and one false negative case, and as such micro F1 is the same as accuracy. Mentioning both metrics does not bring any additional dimension to the performance. Weighted F1 takes into consideration the class support and may mask bad performance on

rare classes. The recall, precision and confusion matrix per class are necessary to spot problems [3]. The unit of averaging can be the patient, center, image, lesion or class. Macro averaging over patients answers a completely different question than macro averaging over classes. The manuscript must explicitly mention both operations when they are used at the same time. First, calculate the class-specific measure per patient, second, average over all patients and third use the pre-specified weights for the classes.

Multilabel tasks

Multilabel images include several diagnoses that could happen together. Hamming loss determines the rate of wrong predictions for the labels, and subset accuracy requires all the labels for a certain image to be true. They solve different problems: a low Hamming loss does not necessarily mean high subset accuracy if most images have at least one wrong label. The interpretations of these measures depend much on label prevalence and on the number of labels.

Thresholds chosen separately for each label would be more appropriate than a common threshold when label distribution differs; but all threshold choosing must be done before testing on the final set. It is inappropriate to treat missing labels automatically as negative. If a dataset includes just some findings, evaluation needs either negatives confirmed or a positive-unlabeled approach to be specified. Otherwise, a correct additional diagnosis in a clinical sense would count as a false positive.

There are co-occurrence issues as well. Two plausible labels could make an impossible combination, or the lack of one label could change the meaning of the other label. Therefore, the multilabel evaluation could include violations in consistency and complete-set evaluation alongside per-label discrimination. Implementing these improvements requires an explicit ontology and cannot be done after the analysis of errors just to improve a certain model.

Ordinal outcomes and agreement

Ordinal classes have intrinsic order without necessarily implying equal distances between classes. Weighted kappa assigns partial disagreement based on a weight matrix, which is determined using either linear or quadratic weights, giving preference to different types of errors. The value of kappa is dependent on the marginal distribution of the labels and on the type of chance model used. It is desirable to report kappa along with the observed confusion matrix and the clinical importance of adjacent or remote misclassification. [42, 12]

The ordinal class being turned into a numeric quantity with mean absolute error reported requires an implicit distance between the classes. Although it might be justified in some cases, it cannot turn an ordinal scale into a physical measure. Exact agreement, within one category agreement, and the error patterns may give more explicit and complementary measures. In the case of quantitative results obtained from images, the measures of agreement that will be discussed next are more meaningful compared to correlation.

8 Probability scoring and calibration

Proper scores measure probabilistic performance

The probability forecast is an additional claim compared to classification output. The binary Brier score can be understood as the average square distance between the predicted probability and the event occurred. The log loss can be seen as the negative average logarithm of the probability given to the true outcome. The two loss functions have been proven to be proper in their standard versions, which means that the true conditional probability gives the minimum loss. However, their realizations take into account several criteria of performance [43, 44].

Brier score is bounded for binary probabilities and has decompositions that distinguish reliability, resolution and uncertainty under specified conventions. It should not be described as a pure measure of calibration. A model predicting the cohort prevalence for every case can be calibrated in a limited sense while providing little discrimination. Conversely, a model with useful ranking can produce poorly scaled probabilities. [45, 46, 47]

Log loss penalizes confident errors heavily and becomes infinite when a model assigns zero probability to an event that occurs. Numerical implementations therefore often clip probabilities. Clipping limitations have to be disclosed whenever it is necessary for precise replication. With regard to multiclass Brier scores, the decision to either sum or average the squared errors in classes changes the scale; similarly, the binary one-probability rule does not apply to the sum of probabilities for classes.

Calibration is defined relative to an event

Calibration determines how often predictions with probability p show the occurrence of the event with probability p. The considered events can be the presence of a certain disease, prediction accuracy, voxel membership in a particular region, or segmentation acceptance. However, they are different. Calibration of confidence in the top class score does not imply the calibration of all class probabilities. Also, voxel-level calibration does not imply patient-level calibration of the failure rate [48, 49]. The reliability diagram depicts the comparison of predicted and observed frequencies. The ECE normally summarizes absolute differences on the level of bins. This number depends on decisions regarding equal-width or equal-mass binning, number of bins, classwise or confidence-wise evaluation, finite sample noise, etc. Two ECEs cannot be compared directly unless all these choices coincide. It is better to report the number of bins, the diagram itself, and its uncertainty rather than just a single number. [50, 51, 52] The maximum calibration error focuses only on the largest difference between bins and can be sensitive to the finite sample size, especially when there are few examples in a bin. Calibration intercept and slope are two statistics that describe systematic differences in probabilities under a logistic recalibration model; their target values are zero and one, respectively. These statistics do not reveal every local deviation from calibration. Kernel-based statistics and calibrated validation approaches can offer alternative statistics and theoretical guarantees, although with other assumptions [47, 53, 54].

Calibration methods and independent assessment

Temperature scaling rescales logits using a scalar multiplier, while Platt scaling, isotonic regression, Bayesian binning, and Dirichlet calibration offer increasing levels of flexibility. Flexibility in calibration leads to overfitting where the calibration data is small. The fitting data needs to be independent from the evaluation data, and choosing different calibrations also implies an additional validation step [50, 55, 56]. A strictly increasing function does not change AUROC as the ordering of probabilities will be retained, but it will affect the Brier score, log loss, and decisions made at a certain threshold for binary classification. In a multiclass softmax model, temperature scaling retains the top-class probability but does not preserve every one-versus-all class probability ordering. Hence, statements regarding unchanged discrimination need to refer to the specific score and task being analyzed. The calibration technique needs to be revisited for any shifts in acquisitions, populations, and times since the method used to recalibrate in one situation cannot be relied upon in another situation. The goal here is to generate a probability or failure score that helps make an intended decision. Visualization of uncertainty alone does not accomplish this goal [57, 58].

Table 3. Discrimination probability scoring and calibration answer different questions.

Quantity	What it evaluates	Principal limitation
AUROC	Pairwise score ranking	Does not establish calibration or a useful threshold
AP or PR area	Precision-recall behavior	Prevalence and estimator conventions affect values
Brier score	Squared probability error	Combines calibration, resolution and outcome uncertainty
Log loss	Probability assigned to observed outcomes	Confident errors and numerical clipping strongly affect values
ECE or MCE	Binned calibration gaps	Binning, sample size and class reduction matter
Calibration slope/intercept	Systematic probability scaling	Does not capture every local calibration failure
Decision net benefit	Utility under a specified action threshold	Requires justified harms and an appropriate population

9 Semantic segmentation and overlap

Dice and intersection over union

Dice and intersection over union (IoU) measure agreement between predicted and reference foreground regions. Dice is two times the intersection of the sets divided by the sum of the set sizes, and IoU is the intersection divided by the union. For a single binary comparison with a non-empty union, Dice = 2 IoU / (1 + IoU). These measures encourage spatial coincidence and ignore the true negative background region.

Though their numerical scale differs, the ranking of individual measurements by either metric is identical [59, 4]. The mathematical relationship does not normally make it possible to convert from an average Dice to an average IoU. Transformation followed by averaging is non-commutative. The same constraint applies to comparison between a mean class IoU and a frequency-weighted Dice, or to conversion from a volumetric measurement to a mean slice measure. A review study should report the originally aggregated results unless the per case predictions or values permit a valid calculation. Overlap is especially dependent upon target size, because the same absolute distance of the boundary is relatively more important for a smaller structure than a larger one since a larger portion of the volume is shifted. It is not because there is some additional physical error made in predicting the smaller target. Results should be reported in clinically significant size categories if size is relevant to the application, defined a priori before final analysis [2].

Empty masks and background

The absence of a target leads to certain edge cases. In the situation when both masks are void, the Dice formula results in dividing zero by zero. Therefore, implementations can return one, zero, undefined result, or avoid considering this situation at all. In the case when the reference is void but a false positive mask exists, the foreground Dice is zero, but zero does not tell us how big the false positive burden was – either one voxel or a big spurious lesion.

Scoring correctly empty cases as one can be defensible for a task that includes absence detection, but it changes the interpretation of the mean. A dataset dominated by negative slices can achieve a high all-slice Dice even when positive-slice delineation is weak. Excluding all empty-reference cases instead can hide false positives. A transparent report presents positive-target quality, correct absence detection and false-positive burden separately, with the full denominator retained.

Background pixel accuracy is especially vulnerable to dominance by irrelevant image area. Cropping tightly around a lesion changes the background denominator and can improve or worsen pixel accuracy without changing the lesion mask. The evaluation field of view and region of interest therefore form part of the metric definition. A target-aware crop based on reference information must be disclosed as additional information supplied to the evaluation.

Volume agreement and class weighting

Volumetric similarity and relative volume error quantify agreement in size. They can be useful when a downstream measurement is volume, but equal volumes do not imply correct location. Two disjoint masks with equal volume can achieve perfect volumetric similarity. Signed volume errors can also cancel when averaged, so absolute errors and their distribution should accompany any mean signed bias. [60, 4]

Generalized Dice and related weighted objectives change the contribution of classes or regions. Inverse-volume weighting can emphasize small classes, but its behavior depends on whether weights are estimated from a batch, a case or a fixed reference collection. These are important training choices and must not be silently substituted for the standard evaluation Dice used by a benchmark. Loss-function studies demonstrate the need to distinguish the optimized surrogate from the final reported endpoint. [61, 62]

Table 4. Segmentation metric families and the errors they can leave unresolved.

Family	Sensitive to	Can miss
Dice and IoU	Foreground overlap	Boundary location, topology, small remote objects
Volume measures	Size agreement and bias	Wrong location with the same volume
Surface distances	Physical contour displacement	Clinical importance of the displaced region
Tolerance surface measures	Fraction within a chosen tolerance	Errors hidden by a loose tolerance
Object metrics	Missed, spurious, split and merged objects	Detailed boundaries inside matched objects
Topology measures	Connectivity, components and holes	Spatially misplaced but topologically plausible anatomy

10 Boundary and surface metrics

Physical distances

Boundary metrics define conflict zones between the surfaces. Given a point on one surface, its distance to the closest point on the other defines the unidirectional difference. The symmetric summaries combine both distances into one. The largest distance represents the Hausdorff metric; the mean and percentiles describe different aspects of the distribution. Not all metrics are the same: while some describe average error, others can detect outliers [63, 64]. Physical distance is crucial in medicine. A one-voxel displacement through a thick CT or MRI slice can exceed a several-pixel in-plane displacement. Computing distance in array indices can therefore reverse the apparent ranking of two predictions. Coordinates must incorporate the correct spacing, orientation and transformation back from resampled model input. The evaluation should occur in a declared common physical space.

The surface representation also matters. Counting boundary voxels weights them differently from weighting mesh triangles or surface elements by area. Anisotropic spacing can amplify the discrepancy. Average symmetric surface distance is sometimes implemented as the mean of two directed means and sometimes as a pooled mean over all surface samples. Those definitions differ when the surfaces have different numbers or weights of samples. [4, 65]

Hausdorff distance and HD95

Hausdorff distance is sensitive to a single remote island or annotation outlier. That sensitivity can be useful when any large spatial failure is unacceptable, but can also make a result unstable. HD95 reduces the influence of a small fraction of extreme distances. It may consequently miss a small but consequential false-positive structure. Reporting both a tail distance and object-level false positives can provide a clearer account of this situation.

HD95 itself is not uniquely specified by its name. One implementation takes the 95th percentile of pooled bidirectional distances; another takes the larger of two directional 95th percentiles. Percentile interpolation, connectivity and distance weighting can further alter results. Empty predicted or reference surfaces require an explicit failure policy, because the usual nearest-neighbor distance may be undefined. These cases should not disappear from the analysis simply because a numerical library returns a nonfinite value. [2]

Tolerance based measures

Normalized surface Dice measures the fraction of the combined boundary surface that lies within a specified tolerance of the other boundary. The tolerance can make the result more closely reflect acceptable contour disagreement, but only if it is justified for the anatomy and use. Selecting the tolerance after inspecting model rankings defeats that purpose. A tolerance in millimeters is not equivalent to a fixed number of pixels across scans with different spacing. [66]

Boundary precision, recall and F score similarly require a contour-matching tolerance and a matching procedure. Boundary IoU evaluates overlap of specified boundary bands and can expose errors diluted by large object interiors. Band width, discretization and image resolution are part of its definition. None of these measures automatically establishes that a contour is safe for a treatment plan or accurate for a downstream measurement. [67, 68]

A useful boundary evaluation reports an interpretable central or average error, an appropriate tail measure and a tolerance-based success fraction when the tolerance has a meaningful basis. The specific combination should remain small and prespecified. Correlated variants of the same surface statistic add less information than a direct measure of the decision that uses the contour.

11 Instance segmentation and topology

Separating objects

Instance segmentation has two intertwined problems: recovering the correct objects and delineating each recovered object. Unioning all nuclei into one foreground mask removes the distinction between a correct set of nuclei and a merged cluster. However, the representations used by methods like StarDist and HoVer-Net are different, and therefore the assessment must preserve the object identities rather than just measure foreground overlaps. One-to-one object pairing avoids the problem of counting the same reference object multiple times. False negatives occur when there are no matching references, whereas false positives occur when there are no matching predictions. Splits and merges require an explicit criterion because depending on how matching candidates are paired, they mean different things. Hungarian matching, greedy score-matching, and overlap-based criteria can result in different pairings. When comparing scores from a particular benchmark, the benchmark's pairing criterion must be implemented exactly. The Aggregated Jaccard Index (AJI) measures the sum of intersections and unions between matches minus a penalty term for the unmatched predicted area. However, the original AJI and the one-to-one variations that follow it can differ; the actual version must be determined, especially in

highly crowded pathology images. Size mismatch can also affect the aggregation because larger objects have more pixels.

Panoptic quality and conditional quality

Panoptic quality combines segmentation quality and recognition quality. Under the conventional formulation in matched instance settings, PQ is equal to the sum of matched IoUs divided by TP + 0.5 FP + 0.5 FN. This expression can be factorized as a product of the meaning matched IoU and the object level F1. This factorization helps to differentiate the problem of bad segmentation from the problems of missing or false objects. However, the matching threshold and the way of dealing with categories and ignored regions still play an important role [13]. A conditional metric like the mean IoU of the matched instances may rise due to non-detection of hard objects; this does not contradict since the evaluation sample becomes easier. Thus, the recognition quality, including recall and the number of the eligible reference objects, needs to accompany the conditional segmentation quality. The same applies to localization error restricted to successfully detected landmarks or bounding boxes. Partition metrics provide another angle on the problem. Rand-type indexes check consistency of assignment of pairs of objects. Adjusted Rand index is considered the random model. Information theory-based variation of information distinguishes information about split and merge operations. Large object size, background, and the unit in the contingency table affect these measures, so they cannot be automatically considered as substitutes for the clinically relevant number of objects [71, 72, 73].

Connectivity and anatomical structure

Tubular objects require careful attention to connectivity. A small discontinuity in the object can have a trivial effect on volume overlap but still break a measure or path for planning. The measure clDice evaluates the consistency between skeletons and masks. Topological loss functions based on persistent homology are similar means to encourage certain topological properties in the training process. However, assessment also needs to ensure the right spatial location: an object can be connected, but its connection could be at the wrong spot [74, 75]. Betti numbers represent such properties as connected components and holes. But having the right numbers of such features does not guarantee that the predicted connectivity matches the anatomy of interest. Skeletonization, connectivity choice, and discretization processes affect topology too. An assessment of topology in combination with overlapping, boundary distance, and specific connectivity or endpoint for a particular task would be helpful.

12 Bounding boxes and object detection

Detection correctness requires assignment

The prediction in a box-detection task consists of the class label, confidence score, and the coordinates of a bounding box. The success of the detection is determined by the correctness of both predicted class label and by some spatial criterion, which most often is the Intersection over Union (IoU) between a prediction and a reference box. Predictions are matched according to benchmark-specific policy. The duplicated high-confidence box is considered a false positive after the match with a reference. Assigning

all overlapping boxes as correct would result in the inflation of results by rewarding redundant predictions [16, 76]. IoU criterion is dependent on the object size and box shape – the fixed translation affects the score in a larger relative manner for small boxes, and elongated objects might be prone to make mistakes along their short axis. There are different notions of overlap for axis-aligned and rotated boxes. In case of 3D boxes, it is preferable to use either physical coordinates or some consistent coordinate system which preserves the geometry of the box. It does not make sense to evaluate the 2D slice box as recovery of the entire 3D lesion. The error types in classification and localization can interfere with each other – correctly localized but misclassified box counts both as false positive of the predicted class and missed positive of the ground truth class. Error analyses proposed by Hoiem et al. and TIDE help to separate errors in localization, classification, duplicates, background, and object errors. Such analysis can be more helpful for model development than any minor improvement in the aggregated metric like AP [77, 78].

AP and benchmark conventions

Detection average precision summarizes a precision-recall curve constructed after matching predictions. AP50 uses an IoU threshold of 0.50; AP75 uses 0.75. A detector can find the correct region sufficiently well for AP50 while placing boundaries too loosely for AP75. The difference is therefore evidence about the chosen localization criterion, not an inconsistency between metrics.

COCO-style headline AP averages across IoU thresholds from 0.50 through 0.95 in steps of 0.05 and across categories, with specific recall sampling and evaluator conventions. Size ranges, limits on detections, crowd annotations and ignored objects are also part of the benchmark. An implementation described only as mAP may instead average AP at one IoU threshold. The exact evaluator and version are necessary for a defensible comparison. [17, 5]

Average recall depends on the allowed number of detections/proposals. An increase in this budget will help detect more objects but will also lead to more reviews. Recall calculated without any restriction to the budget is not a replacement for precision or false positive rate. It is essential that reports specify the budget as well as the practical operating point.

Alternative detection metrics

Localization recall precision error provides a decomposition of localization error, false positives and false negatives. Its optimal-threshold variant identifies a confidence operating point under that objective. The threshold must be selected using permitted development information when estimating future performance. LRP complements AP by making components of error explicit; it does not remove the need to justify the localization tolerance. [79, 80]

Probability-based quality measures consider uncertainty in space and labeling in the probabilistic context. This measure evaluates a different question than the overlap threshold-based AP measure. A detector which produces only deterministic bounding boxes cannot provide an evaluation of spatial probability. The work done in the probabilistic detection domain also highlights the significance of evaluating uncertainty under changing circumstances instead of relying on confidence scores which have undefined semantics [81, 82].

Generalized IoU and distance IoU account for enclosing-box geometry or center displacement in addition to overlapping. Complete IoU includes further geometric factors. These measures are often used as regression losses. They should not be reported as if they replace object recall, duplicate handling or a benchmark's AP definition. A training surrogate and a system-level evaluation metric occupy different roles. [83, 84, 85]

Table 5. Detection comparisons require both geometric correctness and object accounting.

Measure	Main comparison	Parameters that must accompany it
AP50 / AP75	Confidence-ranked detection at one IoU threshold	Class matching, interpolation, duplicates and ignored objects
COCO-style AP	Average across categories and IoU thresholds	Evaluator version, recall sampling, size ranges and maximum detections
Average recall	Recovery within a detection budget	Budget and IoU averaging
FROC sensitivity	Lesion recovery at a false-alert burden	Eligible lesions, match tolerance and false marks per case
LRP error	Localization, false positives and false negatives	Confidence selection and localization threshold
PDQ	Spatial and class probability quality	Required probability representation and assignment
GIoU / DIoU / CIoU	Box geometric agreement or training surrogate	Formula and box geometry; object recall remains separate

13 Lesion localization and free response evaluation

Why lesion detection differs from ordinary classification

A patient or scan can contain no lesion, one lesion or several lesions. An image-level classifier may correctly identify that disease is present while missing most lesions. Conversely, a detector may identify a small lesion accurately but generate several distracting false alarms elsewhere. Lesion-level sensitivity and false-positive marks per scan preserve these distinctions.

Free-response receiver operating characteristic analysis plots lesion localization sensitivity against a false-positive rate expressed per image, scan or case. It permits multiple marks. The localization rule may use a distance tolerance, overlap, a reference region or a lesion-specific criterion. Both the mark rule and the denominator must be reported. A voxel false-positive rate is not the same quantity as false lesions per scan. [86, 87]

Challenge protocols such as LUNA16 make operating points explicit by evaluating lesion sensitivity at specified false-positive rates. The selected rates express the benchmark's purpose; they are not universal clinical acceptance limits. A study that averages sensitivity at a different set of rates produces a different summary, even if both reports call the result a FROC score. [88]

Reference lesions and ambiguous findings

Eligibility criteria define the reference lesion population. Size thresholds, lesion type, visibility, completeness of annotation and reader agreement can materially change sensitivity. The LIDC-IDRI resource illustrates the importance of multi-reader lesion annotation. Evaluation should state how disagreements and uncertain nodules are handled, rather than silently presenting one derived label set as error-free. [89]

In pathology, the slide, lesion and patient are nested units. Detection of a metastatic region may support a patient-level diagnostic decision, but the two endpoints remain distinct. The CAMELYON-related diagnostic study provides a useful example of lesion and clinical classification evaluation in a shared application. Any comparison with another study must preserve its slide selection, reference process and patient grouping. [90]

A detector can be tuned to increase sensitivity by emitting many candidates. Reporting the number of false marks makes the resulting burden visible. For a workflow, false positives should be measured at the point where a user encounters them, after the same grouping and postprocessing that deployment would apply. Slice-level marks merged into one lesion should be evaluated through the actual merge procedure.

LROC AFROC and uncertainty bands

Localization ROC analysis combines correct localization with case-level detection behavior under a specified marking design. Alternative FROC methods and newer extensions address how free-response marks and case-level false-positive events are summarized. These curves should not be merged under one generic AUC label, because their axes and observational units differ. Recent nonparametric work extends LROC to free-response tasks and develops confidence-band procedures. [87, 91]

Curve uncertainty should respect clustering of lesions within patients and readers within cases when applicable. Resampling isolated lesions can break the within-patient dependence. A reader study may require a design that accounts for both reader and case variation. Where the raw data are unavailable, a review should retain source-reported uncertainty rather than invent a pooled confidence interval from published means.

14 Landmark localization pose and registration

Distance and normalization

In case of corresponding predicted and reference landmarks, the Euclidean distance becomes a natural metric of choice. Mean radial error represents the average distance, whereas the median and some higher percentiles give a sense of the most typical and tail distance values, correspondingly. Root mean square error place more weight on the large deviations. Units are necessary to define, and all distances must be calculated in the coordinate system that is meaningful for the purpose of application [18, 19]. Normalized mean error is expressed as a ratio of some distance divided by the scale selected, such as the head size, interocular distance, or image diagonally. Even though this makes comparison of objects of different sizes possible, different scales result in different metrics and hence direct numeric comparison of the

performance of one normalization technique with another normalized error is meaningless without proper conversion. Normalization does not help with detecting significant errors in absolute terms where physical tolerance is of interest. In the case of anisotropic voxels, physical scaling and proper orientation transformation are needed. Scaling an image affects coordinate error in pixels even if the physical prediction stays the same. Evaluation requires transforming predictions into the original physical coordinate system or any other predefined one. Transformation and origin conventions need to be verified using known coordinates prior to evaluation.

Threshold based landmark success

Detection success rate measures the number of landmarks lying inside a specified tolerance zone, measured in millimeters when applied to medicine. Percentage of correct keypoints (PCK) uses the similar idea but with explicit normalization; PCK and PCKh are distinguished by the scale against which the tolerance is measured. Visualization of results in terms of performance versus tolerance is a good way to demonstrate the stability of the conclusion, while the initial tolerance value needs to be chosen about the target use case before the experiment (references [92, 93, 18]). All eligible failure cases must be counted in the denominator. Average distance calculation only for those landmarks that are visible and correctly predicted gives a conditional measure of performance. Not providing coordinates for especially difficult to predict landmarks make the average distance artificially low in case of ignoring such cases. It is recommended to report localization availability or detection rate, distance distribution conditionally and the approach to handling complete failures.

Object keypoint similarity is measured through distances normalized against the object scale and keypoint-specific tolerance parameters. The COCO evaluation tool allows performing thresholded matching and aggregating results in a manner similar to average precision. The key elements to reproduce the result are visibility flags, object area scale, and keypoint-specific constants. Using human pose constants for different body landmarks without justifying it will change the meaning of the allowed error [17].

Registration accuracy

Registration aligns coordinate systems, so fitted landmark error and independent target error differ. A small fiducial registration error indicates close agreement at landmarks used in fitting. It does not guarantee a small target registration error at independent locations, particularly outside the spatial support of the fiducials. Evaluation should use independent targets appropriate to the downstream procedure. [19]

Procrustes alignment or other post hoc transformations can deliberately remove translation, rotation or scale before computing pose error. That may be appropriate for shape comparison, but it removes components of absolute localization error. Reports should state whether alignment is part of the algorithm, part of the metric or an oracle step using the reference. Absolute and aligned errors should not be interchanged.

Table 6. Landmark and registration measures should preserve units and eligible failures.

Measure	Interpretation	Required specification

Measure	Interpretation	Required specification
Mean / median radial error	Average or typical coordinate discrepancy	Physical units, correspondence and failure policy
RMSE / high percentile	Large-error emphasis or tail behavior	Aggregation, sample count and outlier policy
Normalized mean error	Error relative to anatomical scale	Exact normalization denominator
SDR / PCK / PCKh	Fraction within a tolerance	Physical or normalized threshold and visibility
OKS	Scale and keypoint normalized similarity	Keypoint constants, area and visibility convention
Target registration error	Error at independent targets after alignment	Target independence and transformation model

15 Localization explanations and image fidelity

Saliency is not a lesion mask

A saliency or attention map identifies image regions associated with a model's output under a particular explanation method. It is not automatically a calibrated probability of lesion membership. Thresholding it into a region introduces a new selection procedure. A visually plausible highlighted area cannot by itself establish correct reasoning, faithful explanation or successful detection. Sanity-check studies show why explanations need validation against controlled changes to model and data. [94]

Weakly supervised localization can be evaluated against spatial references, but the choice of threshold and access to localization labels must be disclosed. An image classifier that receives localization labels for threshold tuning has received additional supervision. Work on evaluating weakly supervised localization emphasizes this distinction. Comparisons should report the classification task, localization criterion and tuning information separately. [95]

A pointing-style criterion asks whether one selected peak falls inside a reference region. This may be useful for a narrowly defined attention question, but it does not measure full lesion coverage, contour quality or false-positive extent. Overlap, detection or landmark metrics should be selected according to the spatial output that the system actually claims to produce. A review should avoid relabeling explanation visualizations as clinically validated localization.

Fidelity metrics answer another question

Methods for image synthesis and restoration often utilize pixel-wise measures of error such as PSNR, SSIM, and perceptual distances based on learned feature representations (LPIPS, FID). The latter measure is based on distribution in a feature space according to some approximation. Although useful for evaluation of reconstruction or generative methods, they do not provide any means of proving correctness for classification, segmentation, or localization tasks [96, 97, 98].

A synthetic image might look good despite removal of small lesions or change in boundaries. Alternatively, an image with poor perceptual fidelity might still retain a measurement required by a task at hand. For cases where synthetic images serve as an aid to downstream analysis, it is necessary to evaluate the downstream task without regard to generated images, including pathological instances. In our examples, the cross-modality MRI synthesis is pertinent to the first case, while the skin lesion synthesis is pertinent to the second case. Neither example should be used for comparison of image fidelity to clinical validity [99, 100].

16 Uncertainty selective prediction and conformal evaluation

Uncertainty must predict a consequential event

Ambiguity in the image or reference, model knowledge limitations, distributional changes, or prediction instability can result in uncertainty. Both the use of Monte Carlo dropout and ensembles offer independent mechanisms for generating different types of uncertainty. These need to be evaluated considering the relevant metric, which could be error detection, individual failure, or the ability to make referrals. Uncertainty alone is not enough proof of good calibration [101, 102, 57].

For segmentation, probabilistic models can represent multiple plausible masks. Evaluating only the mean or most likely mask discards part of that claim. Distributional measures such as generalized energy distance compare diversity and agreement using a chosen base distance and reference annotation samples. The mathematical properties depend on that distance, and a small number of annotators may poorly represent the true range of acceptable masks. [103, 104]

Uncertainty can be assessed at voxel, object or case level. A high-entropy boundary map may reflect expected ambiguity without identifying a completely missed lesion. Conversely, a model can confidently segment the wrong anatomical structure. The evaluation should therefore ask whether uncertainty identifies the failure that matters in the intended workflow, including confident errors and absent predictions. [49]

Risk coverage and referral

A selective predictor is capable of either withholding automated predictions or referring to human review. Coverage stands for the fraction of cases kept for automated predictions. Selective risk is defined as the mean loss of the task associated with the kept cases. The curve connecting risk and coverage describes the trade-off between these two quantities. The area under this curve is employed as the summarizing measure, but along with this measure, one should clearly provide the corresponding loss, numerical integration, and the task error rate [105, 106, 107].

The subset of rejected cases falls into the domain of the system; the number of referrals, the composition of these referrals, the effort needed for resolution, and the effects of the remaining errors in the subset of accepted cases should all be reported. To become operationally effective, the system that demonstrates high retained accuracy with almost all cases being referred cannot be considered as such [108].

Recent research on selective segmentation expands the notion of accepting and rejecting the cases to segmentation candidates. It becomes evident that there exist several stages of referral, but candidate acceptance is not patient-level safe. The unit of selection, the downstream action, and the approach to partial failures should all be included in the protocol [109].

Conformal guarantees and their limits

The conformal predictions or outputs come with guarantees of coverage or risk under certain assumptions, most commonly exchangeability or a more plausible assumption. Coverage and other measures of efficiency, such as set sizes, must be reported in the evaluation, as well as the difference between marginal coverage and coverage conditional on groups. The nominal 90 percent coverage guarantee does not ensure that all subtypes of the disease and all hospitals have 90 percent coverage [110]. Conformal risk control and risk-controlling prediction sets involve loss or decision considerations into the framework, and its assumptions, calibration sample and admissible losses play a role. For a segmentation problem, it is important to state whether the controlled event refers to the pixels, region, contour, or whole case. The high pixel-level coverage may coincide with the omission of a small lesion, which means that the mathematical guarantee must match the objective. [111, 112] Distribution shift, multiple adaptation and reuse of calibration labels can make the guarantee complicated. The reports should clarify the calibration partition and sample size, target risk and selection procedure, as well as the conditions of independent test. Conformal predictions will be relevant if these aspects are specified; they cannot be regarded as a safety certificate.

17 Training objectives and evaluation metrics

Similar names do not imply identical quantities

Training losses are minimized on minibatches and generally depend on differentiable surrogates of the objective function. Evaluation metrics aggregate the predictions following a certain protocol. Soft Dice loss, class-weighted batch Dice loss, and hard patient-wise Dice at some threshold denote different quantities. Claiming that a network was trained on Dice does not necessarily indicate the criterion that was used for the final evaluation of its segmentation [61, 62].

Tversky loss reweighs false positives and false negatives; focal loss and compound losses can focus on difficult samples or alleviate class imbalance, which can help to optimize networks for a particular task. Nevertheless, these weightings are not necessarily indicative of clinically meaningful costs of errors. Comparing losses necessitates controlled splitting of data, tuning budget, sampling strategy, and postprocessing. Positive results obtained using certain configurations do not mean that the loss performs universally better [113, 114]. Boundary losses and Hausdorff losses aim at optimizing spatial error which can be neglected by overlap losses. The differentiable surrogates of these losses do not necessarily coincide with the metric used for the evaluation of the surface in question. Physical geometry, class absence, and distance transforms can affect both optimization and evaluation of these losses. The paper should distinguish between the surrogate used for training and the evaluation metric explicitly [115, 116].

Optimizing the complete pipeline

The final score of the model will depend on the thresholding, connected-component filtering, ensembling, test-time data augmentation, and prompting. All of the above steps can utilize development data. All of those steps must be fixed before the final evaluation and controlled in equal conditions. For example, reference boxes provide target data to the promptable segmentation technique and should not be regarded as an automatic localization algorithm [8, 9, 117]. A good loss ablation involves the modification of just one intended element, keeping all other variables of the experiment unchanged. Multiple random seeds can be used to measure training stability, but this cannot fix leakage or bias in reference samples. The resource usage is another factor of comparison if the methods significantly differ in their inference and human intervention needs. The entire pipeline of the prediction creation process is the proper experimental unit.

Table 7. Training objectives and the evaluation claims they do not establish by themselves.

Objective family	Training emphasis	Additional evaluation needed
Cross entropy	Probability fit over training units	Independent calibration, operating point and class errors
Soft / generalized Dice	Foreground or class-weighted overlap	Hard case-level overlap, empty targets and probability quality
Tversky / focal / compound	Asymmetric or difficult errors	Prespecified false-positive and false-negative consequences
Boundary / Hausdorff surrogate	Contour discrepancy	Exact physical boundary metric and failure behavior
Topology objective	Structural constraints	Spatial agreement and clinically relevant connectivity
GIoU / DIoU / CIoU	Bounding-box regression geometry	Complete matching, detection precision and recall

18 Conditions for valid comparisons

What must match

Interpretation of numerical comparison depends on proper alignment of target population, reference definition, data splitting, prediction information, implementation of the metric, and its aggregation. The mere presence of a benchmark designation does not imply sufficient alignment. Separate releases can involve different cases or labels, while one approach can be trained on data used only by another approach as test data. Pre-training overlap affects evaluations of foundation models if the description of the original training data corpus is not complete [118, 119]. Examples of BraTS benchmark and the Medical Segmentation Decathlon show the importance of defining label definitions and the task itself together with the results. Brain tumor regions, organ masks, and lesion targets for a specific task require different error structures. Even the identical metric name does not mean that the obtained scores for those

tasks represent the same difficulty [120, 121]. Paired evaluations on the same held-out data are more informative compared to comparisons of average scores on completely different cohorts. In paired comparisons, the equality of training information and reasonable tuning budgets are still important. If an approach uses manual bounding boxes for prediction, the evaluation should be performed under an established interaction budget. The detector, which gets a target crop, solves a different task compared to a detector scanning a whole volume. Controlled comparison highlights this difference rather than hides it in the approach name. Scores should be reported in their natural units. Slice-level Dice, volumetric Dice and lesion-level recall cannot be combined into one performance column. AUPRC calculated on the enriched case-control cohort cannot be compared to AUPRC based on representative screening cohort. Mean landmark error after reference-based alignment cannot be compared to physical error without unalignment. Conditions like this are recorded in evidence registry instead of generating a universal leaderboard.

Domain shift and subgroup performance

Generalization involves performance under some defined changes in conditions. Changes in scanner, reconstruction, staining, acquisition protocol, operator, institution, time-related aspects, and disease distribution can cause each change separately. Artificial corruption studies investigate changes but do not imitate all aspects of the transport in clinical practice outside the laboratory. Differentiate between the frozen testing from outside the laboratory on the one hand and target-domain adaptation on the other [58, 122]. Subgroup analysis must be associated with the mechanism of possible failure and intended application. Class, lesion size, anatomy, acquisition machine, and demographic population can be some of the relevant attributes. However, the analysis of few samples provides inaccurate estimates of parameters. Describe the denominators and confidence intervals for subgroups, as well as exploration subgroup analysis. A pooled estimate might hide the failure common to a smaller but meaningful subgroup [33, 34]. There are several interpretations of fairness concepts. The equal opportunity relates to the class-conditional error, while calibration refers to the interpretation of probabilistic prediction. Several fairness criteria can conflict in real life. The description of a small difference does not prove the overall fairness. The choice of a fairness criterion must match the use case and affected populations [123, 124].

Benchmark rankings

Rankings depend both on the magnitude of the effect size as well as the associated variance such that small changes can affect hierarchy even if practical importance is low. The selected ranking may also depend on missing data, class weighting, aggregation, metrics chosen, and the treatment of failures. Issues with reporting and analysis of rank instability make the distribution of interest as opposed to the highest-ranked as the universal truth [25, 119]. A comparison table should indicate whether the findings are directly comparable, replicated using the same methodology, from comparable sources, or simply provided for context. The present review uses the last category only where indicated. An aggregated performance estimate for the entire field is not given, nor does it claim a particular architecture performs well overall.

19 Reproducible worked comparisons

These next examples use purposefully designed counts, probabilities, or masks. These serve to demonstrate metric behavior under certain assumptions. However, they are neither patient data nor model results, and the values obtained from them cannot be viewed as clinical performance indicators. The evidentiary material contains the input data arrays, numeric output in their entirety, and a script recreating all ten examples.

Rare positives and apparently high accuracy

Example E1 contains 10,000 hypothetical cases, of which 100 are positive. An always-negative model has 99 percent accuracy, zero sensitivity and balanced accuracy of 0.50. A second model with sensitivity and specificity both equal to 0.90 has lower accuracy, 90 percent, but detects 90 of the 100 positives. Its 990 false-positive predictions give precision of only 0.0833. Both the recovered positives and the large false-alert workload are relevant; neither accuracy nor sensitivity alone captures the whole result.

The always-negative model's ordinary MCC formula is undefined because one predicted marginal is zero. The script records that result as undefined rather than replacing it with an apparently measured correlation. If an implementation maps it to zero, the report should state the convention. F1 is zero in this example because positive reference cases exist and all are missed.

Prevalence changes precision

In E2, sensitivity and specificity remain fixed at 0.90 while positive prevalence changes from 1 to 50 percent. Precision is 0.0833 at 1 percent prevalence, 0.3214 at 5 percent, 0.50 at 10 percent and 0.90 at 50 percent. The calculation shows the effect of changing only prevalence. It does not assert that sensitivity and specificity will remain fixed when a system is deployed in a different population.

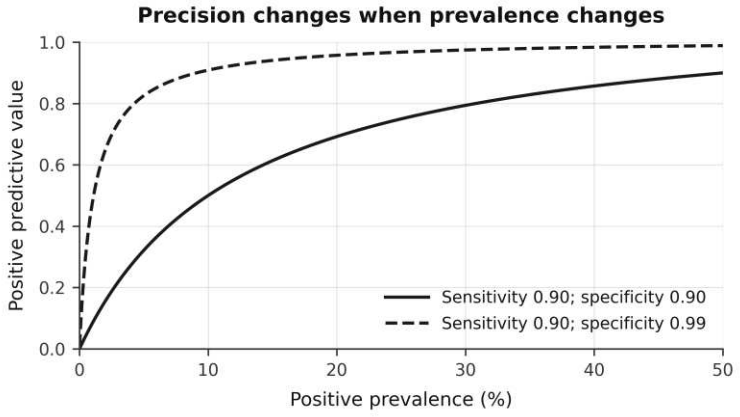

Figure 1. Constructed prevalence analysis. Sensitivity remains 0.90 in both curves; specificity is 0.90 or 0.99. Class-conditional performance is assumed unchanged. The graph is an analytical illustration, not clinical data.

The same ranking with worse probability scores

E3 uses six labels with associated probabilities. As a result of sharping the logits using a positive temperature of 0.25, the order of decisions based on temperature 0.5 is maintained. The accuracy is

unchanged at 0.6667, as is the AUROC at 0.8889. On the other hand, the Brier score moves up from 0.1367 to 0.2324, while the log loss increases from 0.4149 to 0.6021 due to higher confidence in wrong predictions. This is an example that demonstrates the difference between discrimination and probability scoring, where only six artificial observations are used.

Aggregation reverses the preferred model

Scenario E4 has two hypothetical segmentations. Model A gives Dice values of 0.9000 and 0.1818, and Model B gives 0.7000 and 0.8000. Using equal-case averaging, Model B would be preferred because it gives the average Dice value of 0.7500 in B while A gets 0.5409. On the other hand, using the sum of foreground confusion count averaging, Model A will give a better Dice value of 0.8626 in contrast to Model B which has 0.7091. Both models are the same and the average metric will overweigh the larger case. This scenario also shows that an incorrect transformation takes place where the average IoU for Model A is 0.4591 while the transformed mean Dice, calculated as Dice over 2 minus Dice, results in 0.3707. The mathematical equation is true for each binary comparison, but not for the arbitrary averages. Therefore, reviewing papers should not estimate the mean IoU from mean Dice.

Table 8. Constructed example E4. The reduction changes the ranking; no predictions change.

Model	Case 1 Dice	Case 2 Dice	Mean case Dice	Pooled Dice
A	0.9000	0.1818	0.5409	0.8626
B	0.7000	0.8000	0.7500	0.7091

Identical Dice and a remote false positive

Prediction E5 results in a square reference mask as well as two predictions having the same area. In the first case, the square is translated into two pixels, while in the second case, the same amount of foreground is moved to an isolated island. Both predictions have a Dice score of 0.9375. Given that the boundary-pixel implementation of the algorithm is used, the HD95 metric equals 2.00 pixels for the former prediction and around 39.20 pixels for the latter.

This illustration uses unit pixel spacing, four-connected boundary extraction and a pooled bidirectional percentile. It is not an area-weighted three-dimensional surface calculation. The script and figure specify those choices so that the result is reproducible and cannot be mistaken for a universal HD95 convention.

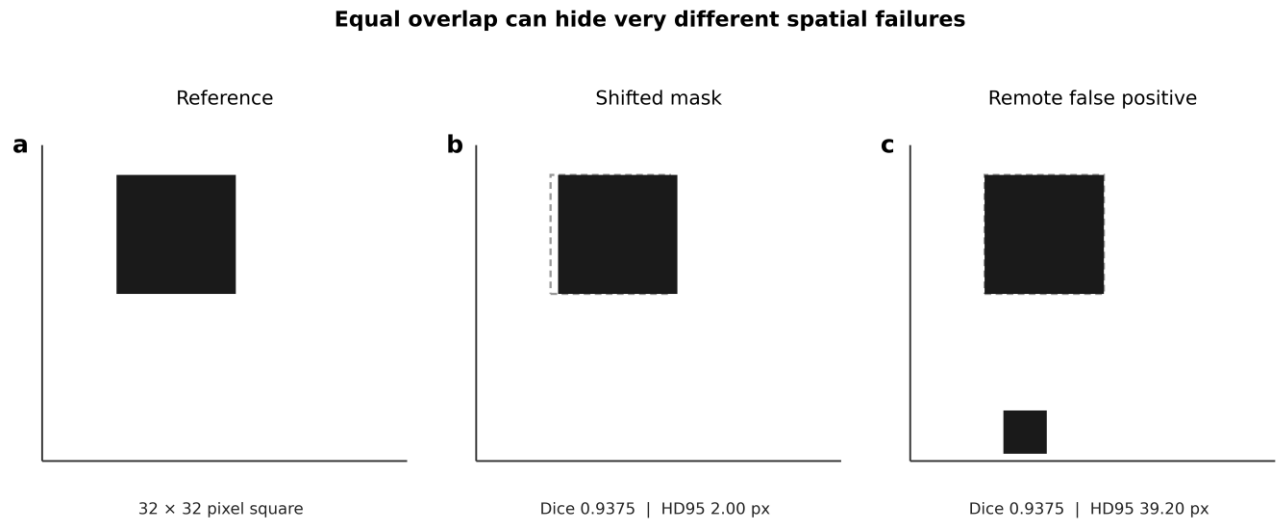

Figure 2. Constructed example E5. Equal Dice coexists with very different boundary error and component count. The dashed contour shows the reference. HD95 uses pooled bidirectional distances between four-connected boundary pixels. Values are in pixels on a unit-spaced grid.

Matching changes detection AP

E6 contains three reference objects and five predictions sorted by confidence. At IoU 0.50, the status sequence is true positive, duplicate false positive, true positive, background false positive and true positive. At IoU 0.75, the third prediction fails the localization requirement. Noninterpolated recall-weighted AP is 0.7556 at the lower threshold and 0.4667 at the higher threshold; recall changes from 1.00 to 0.6667.

These are deliberately specified AP calculations, not COCO's 101-point estimator. Their purpose is to show why the IoU threshold and duplicate policy belong in the reported metric. Ignoring the duplicate would increase apparent precision without improving object recovery. Evaluating only matched boxes would additionally omit the missed reference at the higher threshold.

Pixel error can reverse physical error

In E7, voxel spacing is 0.5 by 0.5 by 4.0 millimeters. Model A is displaced by three voxels along the first axis, giving a physical error of 1.5 millimeters. Model B is displaced by one voxel along the third axis, giving an error of 4.0 millimeters. An index-space comparison favors B; a physical-distance comparison favors A. The intended geometric claim determines which quantity matters.

Referral improves a conditional result

E8 sorts eight theoretical instances into order based on confidence values. In the scenario where all eight instances are retained, there is an error probability of 0.50. Acceptance of the first four instances followed by referral of the remaining four results in a decreased error probability of 0.25. Error committed that occurs within the accepted instances is classified as a confident error. With the help of the given discretization, the right-step area under the risk-coverage graph for the example instance is about 0.2807.

Empty masks inflate an aggregate

E9 contains ten positive scans with Dice 0.80 and ninety correctly predicted empty scans. Assigning a score of one to both-empty cases yields an all-scan mean Dice of 0.98. Evaluating positive scans alone gives 0.80. Neither value should be described without its denominator and policy. The all-scan value includes success at recognizing absence; the positive-only value does not describe false-positive behavior on negatives.

Utility needs a decision threshold

E10 contains 1,000 hypothetical cases with 100 positives. A model leads to 80 true-positive and 90 false-positive actions. At a specified threshold probability of 0.10, net benefit is 0.070, compared with zero for the treat-none and treat-all strategies under these constructed assumptions. This value represents a particular weighting of actions. It is not a recommended clinical threshold and cannot be generalized without specifying the action, harms and population.

Table 9. Ten reproducible constructed examples. These are mathematical demonstrations, not estimates from patients or trained models.

ID	Mechanism	Illustrative result
E1	Rare positives	Always-negative accuracy 0.99; sensitivity 0.00
E2	Prevalence	PPV 0.0833 at 1% prevalence and 0.90 at 50%
E3	Probability scaling	Same AUROC 0.8889; Brier 0.1367 versus 0.2324
E4	Aggregation	Mean-case ranking favors B; pooled ranking favors A
E5	Spatial failures	Equal Dice 0.9375; HD95 2.00 versus 39.20 pixels
E6	Detection matching	Noninterpolated AP 0.7556 at IoU50 versus 0.4667 at IoU75
E7	Physical spacing	Three voxels = 1.5 mm; one thick-slice voxel = 4.0 mm
E8	Referral	Risk 0.50 at full coverage; 0.25 at half coverage
E9	Empty masks	Positive-only mean Dice 0.80; all-scan mean 0.98
E10	Utility	Net benefit 0.070 at the specified hypothetical threshold

20 Published examples and the supplied author papers

Principles for author associated evidence

Ten papers whose retrieved authorship records include Behnam Kiani Kalejahi were retained from the supplied publication list. They cover multiclass radiograph classification, multimodal thyroid diagnosis, plant-image classification, image synthesis, brain-tumor segmentation, uncertainty-aware endoscopy and a segmentation review. Their inclusion is based on relevance to the present metric questions. They are

evaluated with the same evidentiary expectations as other publications and are not used to construct a favorable cross-study ranking.

Four accessible papers provide 24 selected numeric records. This is a purposive illustration set. It cannot estimate how frequently imaging research uses appropriate metrics or validation, and it is not a representative sample of the literature. Reported confidence intervals, seed variation and image or patient units are retained as different quantities.

Table 10. Relevance of the ten verified papers associated with the supplied profile.

Publication	Application	Use in this review
2026 [125]	Multiclass bone-tumor classification	Accuracy versus multiclass AUC and class support
2026 [126]	Multimodal thyroid classification	Discrimination, Brier score, thresholds and reporting consistency
2026 [127]	Plant-image classification	Cross-domain classification context and metric reduction
2021 [100]	Skin-lesion synthesis and classification	Image fidelity versus downstream classification
2026 [99]	Cross-modality MRI synthesis	Fidelity versus preservation of task-relevant anatomy
2023 [128]	3D GAN segmentation	Original metric labels and region definitions
2023 [129]	Auxiliary-classifier GAN segmentation	Segmentation evaluation and comparator conditions
2026 [130]	2.5D brain-tumor segmentation	Positive-only versus all-slice aggregation
2026 [108]	Uncertainty-aware endoscopy	Shift, retained accuracy, coverage and referrals
2026 [131]	Brain-tumor segmentation review	Robustness, uncertainty and translation context

Thyroid classification separates metric roles

The multimodal thyroid study reports a 77-patient holdout comparison in Table 1. Both v1 and v2 have a reported AUROC of 0.977, while Brier scores are 0.083 and 0.042. At threshold 0.5, reported sensitivity is 0.955 versus 0.977, and specificity is 0.818 versus 0.909. These are source-reported values with their original intervals retained in the extraction file. They illustrate that similar discrimination can accompany different probability scores and operating-point decisions. They do not establish significance for every difference. [126]

The source also describes identical patient rankings while reporting AUPRC values of 0.984 and 0.979 in that table. Identical rankings, ties, labels and AP definitions would produce identical rank-based AP. Rounded equality of AUROC alone would not require identical AUPRC. Consequently, the original predictions and exact AUPRC implementations should be checked to reconcile the broader wording with the table. The review preserves this issue instead of silently changing the numbers or inferring a resolution.

Endoscopy exposes the referral denominator

Table 3 of the supplied endoscopy paper reports GI accuracy under resolution shift at several coverage levels. At full coverage, 5,909 images are retained and accuracy is 0.683. At 25 percent coverage, 1,477 are retained, 4,432 are deferred and accuracy on the retained subset is 0.852. The larger retained-subset value does not describe performance on all original images, and the deferred work is part of the result. [108]

The numerical records identify the GI evaluation unit as images. They should not be converted into a claim of independent patient-level external validation. A resolution perturbation tests a specified acquisition stress and does not by itself reproduce a new clinical site's case mix. This example motivates reporting accuracy or risk jointly with retained and deferred counts.

GDNet demonstrates aggregation sensitivity

GDNet Table 1 reports test results under positive-only and all-slice conventions. For enhancing tumor, the values are 0.6536 and 0.9189, a difference of 0.2653 or 26.53 percentage points. Whole-tumor and tumor-core results also differ between the conventions. These are changes in the reported aggregation for the same model, not improvements obtained by a new architecture. [130]

The table reports variation across two training seeds. That variation is not a patient-level confidence interval. The review also preserves the slice aggregation unit and does not compare these numbers directly with subject-level volumetric BraTS challenge Dice. The practical lesson is to define empty-target treatment and averaging before interpreting a high aggregate.

Bone classification and related contributions

Bone-CNN Table 3 reports test accuracy of 96.52 percent and macro-AUC of 0.9989 for the proposed model. These quantities describe different aspects of the source experiment. They should not be compared as if the higher numerical value were the stronger performance result. Per-class support, AUC reduction, held-out unit and reference process are needed to interpret the aggregate. No superiority across independently published models is inferred here. [125]

The auxiliary-classifier GAN and 3D GAN papers provide segmentation contexts in which label names, region definitions and reported metric formulas should be reconciled before numerical comparison. The related brain-tumor review contributes discussion of robustness and translation. The plant-image classification paper broadens the application context beyond medicine, while the synthesis papers motivate separate evaluation of image fidelity and downstream task performance. [129, 128, 131, 127, 99, 100]

Table 11. Selected source-reported comparisons. They are illustrative and do not form a cross-study ranking. Full records preserve source uncertainty.

Source	Original setting	Selected values and interpretation
Thyroid [126]	77-patient holdout, Table 1	AUROC 0.977 for both; Brier 0.083 and 0.042. Original prediction/ranking wording requires reconciliation.

Source	Original setting	Selected values and interpretation
Endoscopy [108]	GI resolution shift, Table 3	Full coverage: accuracy 0.683 on 5909 images. 25% coverage: 0.852 on 1477 retained; 4432 deferred.
GDNet [130]	Slice aggregation, Table 1	Enhancing-tumor Dice 0.6536 positive-only versus 0.9189 all-slice. Seed variation is not a patient CI.
Bone-CNN [125]	Radiograph test split, Table 3	Accuracy 96.52%; macro-AUC 0.9989. Different estimands; no cross-study superiority inferred.

21 Statistical inference and uncertainty intervals

Confidence intervals must follow the sampling unit

A point estimate does not show how precisely performance has been estimated. For a fixed binary operating point with independent cases, binomial intervals such as Wilson intervals can describe uncertainty in sensitivity or specificity. Their sample sizes are the number of reference positives and negatives, respectively. They are not the total number of pixels, patches or augmentations used during training. [27]

Bootstrap methods resample observed units to estimate sampling variability. In imaging, resampling should usually preserve the relevant clusters, such as patients with several scans or slides with many patches. Paired model comparisons should resample the same cases for both models and compute the difference within each replicate. This retains the correlation between models evaluated on the same examples. The resampling design and handling of rare classes must be specified. [132, 21]

A bootstrap interval conditional on one trained model does not include every source of uncertainty in training, data collection or annotation. Repeating training seeds measures a different variation component. Reports should distinguish patient sampling intervals, seed standard deviations, cross-validation variability and predictive uncertainty for an individual output. Using the same plus-minus notation for all four obscures their meaning.

Paired tests and practical importance

The approach by DeLong considers AUROCs that are correlated within the same statistical framework. McNemar’s test considers paired binary decisions through discordant pairs at a particular threshold. These tests answer different questions since a nonsignificant difference in AUROCs does not indicate whether decisions based on thresholding differ, and vice versa. If overlap, distance, or loss differences need to be analyzed, then a paired comparison will need to maintain pairing at the level of patients. An interval for the effect size could provide more information than just the p-value. The metric of interest and the comparison itself need to be pre-specified, and adjustments need to be made for multiplicity when there are multiple models, targets, or thresholds being analyzed. A change that is statistically significant might be practically insignificant. The numerical difference between two published means does not form a proper paired significance test. Covariance matrices and distributions of errors are normally not available

without the raw paired outcomes. Using p-values calculated from means and intervals unrelated to the mean can be highly misleading. The results of this review only include arithmetic differences when explicitly mentioned.

Cross validation and multiple datasets

The cross-validation folds are correlated because of having overlapping training sets. Using the fold scores as independent repeats in an easy t-test may underestimate the amount of uncertainty. Bengio and Grandvalet have found constraints for universal unbiased variance estimation of k-fold cross-validation, whereas Dietterich explores approximate tests for comparison of learning algorithms. No test emerges from these results that are applicable automatically to all possible imaging experiments [135, 136]. Separation of model selection and performance estimation by using an actually held out set or a proper nested design is necessary. Iterative model optimization by using a public leaderboard also involves a selection effect even with hidden test labels. Provenance of data splits and a final clean evaluation set are worth more than application of some complicated interval to a contaminated experiment [23, 24, 118]. In comparative studies across different datasets, non-parametric comparison of techniques becomes relevant if the datasets themselves form units of generalization. Demšar describes such experimental designs and their assumptions. These assumptions cannot be used to justify the use of several related instances of a dataset as independent evidence, and ranks cannot express absolute values of differences in performance [137].

22 Clinical and operational usefulness

Measure the quantity used downstream

Masking may often be used behind the value of interest, which can include volume, diameter, function, and planning parameter. The measured value must be evaluated in its specific units of measurement. The same Dice similarity values may be associated with different volume errors, and a small boundary error near a critical structure may affect more than a large error of non-critical structure. Echocardiography analysis is an example of how image segmentation correlates with a functional outcome [138].

The Bland-Altman method is used to analyze bias and scatter of the paired quantitative measurements. Under certain conditions, the limits of agreement provide estimates of expected differences. Some problems, including heteroscedasticity, repeated measurements, and non-normal differences, require alternative solutions. Correlation alone is not enough because the method may preserve the order of measurements while overestimating all measurements [139]. The intraclass correlation requires certain model and interpretation. Agreement is not the same as consistency. The reliability of a single measurement and average measurement differ. The chosen ICC must match the expected usage for readers or devices. The high value of ICC in heterogeneous population may occur with a high absolute error of a method [140].

Human effort and workflow outcomes

Time savings from automatic output are only possible if it can be accepted or corrected by the users. The evaluation process needs to consider the whole pipeline, including loading, reviewing, correction, retrying, and escalation. A contour that has good geometric scores might still need heavy correction if there are many mistakes in several boundaries. Work on organ at risk segmentation has shown how important is to study time saving processes [141]. Human-AI work needs to clearly describe the users, their training, task order and the endpoint used. Learning of the reader, ordering of cases, carryover effect, and possibility of accepting mistakes are among the factors that might affect the study results. Accuracy study in a retrospective fashion cannot be used to conclude that the system leads to better medical decisions and patient outcomes. Early clinical evaluation and interventional reporting frameworks help to answer these broader questions [142, 143]. Time, memory and throughput measurements become relevant when they are in the same conditions of deployment. Mention hardware, precision, input size, preprocessing and auxiliary models. Latency of model might not include the most expensive operation of the whole process, especially when human interaction gives the prompt.

Reporting and maintenance

CLAIM provides a reporting approach for AI in medical imaging, while FUTURE-AI covers more than just the reporting phase of the lifecycle. A metric table needs to be accompanied by data provenance, purpose, human supervision, and model maintenance. Changes in software, thresholds, preprocessing, and definition of the reference could affect the performance measurements and need to be versioned [144, 145]. Model cards and datasheets assist in evaluating the limits and provenance. FAIR principles will increase the accessibility and reusability of research data that is allowed. This will increase the traceability but does not ensure correctness on its own. For a reproducible release, exact partitions when allowed, versions of evaluators and configurations are required to recreate the measured reductions [146, 147, 148].

23 Recommendations for metric selection

A practical sequence

Start with what the claim is supposed to demonstrate or measure. Identify the population, the gold standard and the smallest possible unit of failure. Identify the question of classification, detection, localization, demarcation, connection preservation or boundary determination and what is the required physical precision. Choose the metric only at this stage. Design single primary and additional secondary endpoints to shed light on different types of errors. Establish threshold values, criteria for match, empty-case treatment and rules of aggregation. Keep full groups of patients for analysis if the data set is split, keep the test from tuning in any way, and choose appropriate uncertainty analysis according to the sampling unit. Analyze the entire process if the claim is about practical value. The recommendations given above can be used as guidelines but not as obligatory packages of actions to be done in every scenario. Task-specific justification can require choosing other actions and it must be documented before

results are analyzed. In case of conflicting primary goals, disclose it explicitly or establish prespecified constrained goal in advance but do not hide the problem in an ad hoc weighted average.

Table 12. Scenario-specific starting points for study design. Each requires a task-specific protocol.

Scenario	Primary emphasis	Complementary evidence and condition
Binary case classification	Sensitivity and specificity at a justified threshold	AUROC, precision, confusion counts and probability assessment. Fix the positive class, prevalence and independent case unit.
Rare positive screening research	Sensitivity at a prespecified false-alert constraint	Precision-recall analysis, predictive values and workload. An enriched test set does not establish deployment precision.
Multiclass diagnosis	Per-class recall and a prespecified macro summary	Confusion matrix, class support and defined multiclass AUC. Report averaging and rare-class uncertainty.
Multilabel images	Per-label performance at fixed thresholds	Hamming loss, complete-set accuracy and label consistency. Distinguish missing annotations from verified negatives.
Organ segmentation	Per-case overlap plus physical boundary error	Volume bias, tolerance-based surface agreement and failures. Define empty masks, spacing and reduction order.
Small lesion segmentation	Lesion sensitivity and false positives per scan	Conditional delineation, size strata and boundary error. Keep missed lesions in the overall denominator.
Vessel or duct segmentation	Connectivity or topology measure plus spatial accuracy	clDice, physical boundary error and downstream path success. Topology alone does not establish correct anatomy.
Nuclei instance segmentation	Object recovery and matched delineation	PQ components, AJI variant, splits, merges and count error. Use declared one-to-one matching and slide-level grouping.
Bounding box detection	AP under a fully specified evaluator	AP50/AP75, recall budget, error types and realistic operating point. Document duplicates, ignored regions and detection limits.
Medical lesion localization	FROC or other declared localization curve	Sensitivity at fixed false marks per scan and case-level diagnosis. Define lesion eligibility, match tolerance and case clustering.
Landmark or pose localization	Physical or justified normalized distance with failure accounting	Success-rate curve, tail error, visibility and correspondence. Separate aligned error from absolute error.
Selective automated workflow	Task risk at a feasible coverage	Referral counts, rejected-case outcomes, set size and human effort. Do not treat retained accuracy as full-cohort performance.

Minimum comparison information

A comparative analysis should describe population and sample sizes, the reference process, the input to predictions, the data split for independence, metric formulations or evaluator identifiers, threshold settings, and aggregation sequence. How classes that are not represented, output values missing, duplicated detections, areas that are ignored, and failures were handled should also be included in the analysis. A table of scores that is missing these details cannot serve as reliable evidence in cross-study comparison. Effect sizes need to be accompanied by appropriate confidence intervals, classwise or targetwise, along with failure distributions. In case where the computational and annotation resource availability is different, this needs to be explained. In case of limited data sharing, the evaluation code and schema can help understand the scoring procedure without revealing patients' identities. The checklist proposed in Appendix B facilitates such reporting and does not provide any validity scores.

24 Research priorities

A shared evaluator conformance study

The proposed experiment would test multiple popular metric libraries on the same set of purposely designed masks, boxes, and labels. The experimental cases should include empty targets, anisotropic sampling, duplicate detections, ties, ignored areas, and missed detections. The noted discrepancies should be explained by clear conventions and not averaged out. The result would be a versioned conformance test suite and a conversion table between the implementations. The study would separate mathematically equivalent implementations from truly different estimands. For example, the various HD95 variants would be named rather than forced to converge using undocumented hacks. A small number of reference examples could save more comparisons from being unreproducible than introducing one more aggregation metric.

Metric sensitivity and ranking stability

A potential secondary analysis could affect the performance evaluation process without changing the predictions generated by the models. Clinically feasible aggregation criteria, tolerance for boundaries, eligibility of lesions and class weights would be manipulated within specified limits. The study will examine what ranking of models is stable and which one depends highly on the selected convention. Absolute effect sizes and ranks should be measured. The design would separate the variability arising from model prediction errors from the variability arising from the definition of performance. It should use a validation set of predictions for various tasks and reveal the details of the performance evaluation process.

Uncertainty linked to human actions

The third criterion for investigation is the assessment of methods of dealing with uncertainty through realistic referral decisions. Use a test set that is separated from the training set by either time or space, define the review capability beforehand, and compare the error rate, the nature of referrals, correction time, and the failures which continue to be confidently referred to even after correction. Evaluate

calibration based on the true failure event at the level of the individual case, not only on the voxel entropy. Conformal and non-conformal approaches must be compared under similar conditions of development and assumptions.

Meaningful geometric tolerances

The tolerances for boundaries and landmarks must relate to downstream consequences and variation in annotation. This study can involve the comparison of the level of acceptance by experts, the amount of effort required to correct errors, and measurement error for different candidate geometric tolerances based on a pre-defined analysis plan. The aim of this study would be to establish whether a given tolerance is useful, especially the anatomy and conditions under which it might not work. A one-size-fits-all tolerance defined in millimeters cannot be useful for the variation present in different organs, lesions, and procedures. For this reason, the study must assess the transferability of this metric to other tasks without claiming its validity.

Consensus and validation of reporting aids

The proposed checklist should undergo multidisciplinary review before being promoted as a standard. Readers should test whether items are understandable, whether independent extractors agree and whether the checklist identifies consequential evaluation omissions. Items should be revised on the basis of those results. A sum of checked boxes should not become a quality score without separate validation and a justified weighting model.

25 Limitations and conclusions

Limitations

This review takes a broad perspective without being exhaustive. Targeted and bounded methods of reference retrieval tend towards known foundational works and some more recent selected works. The target number of references of 150 is viewed as a curated set size and not as a measure of saturation. Availability of full texts and extent of evaluation may differ and preprints are clearly indicated. Bibliographic screening does not authenticate the studies and does not guarantee all later corrections and revisions have been accounted for.

One quantitative extraction was performed with the aid of an AI tool and needs verification by the author independently. The 24 studies found are all published and originate from four purposively selected papers related to the provided profile and are not enough to provide representative information on reporting practices. The ten illustrative examples show the mathematical implications but do not describe their occurrence rate or size in practical use. No patient-level data reanalysis, machine learning, meta-analysis or clinical validation was performed.

The metric catalog gives brief descriptions and cannot replace a full explanation of the benchmark or statistical method. Some metrics have multiple variations, and the software conventions change. The recommendations and checklist suggested are the interpretive results that need specific evaluation and, for the standardization claim, further consensus and validation.

Conclusions

Evaluation quality depends on the appropriate match between the selected measure and a well-specified research question while maintaining the information required for reproducibility. For classification, it is critical to distinguish operating point, score discrimination, and probability calibration-related decisions. In segmentation, it is vital to consider overlap, boundary, object identity, and topology considerations as needed. Localization entails matching and tolerance as well as consideration of false negatives and false positives.

The first step is specifying the unit, denominator, reference, threshold, and aggregation used whenever claims about performance are made. Independent evaluation and error analysis will then help quantify the extent to which this claim is generalizable. The most valuable comparison will reveal which errors are consistently made, who makes them, and whether the tool improves the intended measurement or workflow. This goal is best accomplished by a well-justified suite of measures rather than a leaderboard score alone.

Declarations

Funding. Not applicable.

Conflicts of interest / Competing interests. Not applicable.

Ethics approval and consent to participate. Not applicable. This review analyzes previously published information and uses constructed numerical examples; no new patient data were collected, and no ethics approval or protocol registration was required.

Consent for publication. Not applicable.

Availability of data and materials. The accompanying evidence package (reference identifiers, extraction records and the ten worked-example inputs and outputs) is available from the corresponding author on reasonable request.

Author contributions

Behnam Kiani Kalejahi: Conceptualization, methodology, literature investigation, formal analysis, software, visualization, and writing, original draft. Mohammad Javad Rajabi: Methodology, validation, and writing, review and editing. Both authors contributed to the interpretation of findings, approved the final manuscript, and agree to be accountable for the work.

Use of AI-assisted technologies. AI-assisted tools were used for literature retrieval, data extraction, drafting, calculations and document preparation. The authors are responsible for independently verifying all claims, references and numerical records prior to submission, accept full responsibility for the content of the manuscript, and disclose this assistance in accordance with the target journal's policy on the use of AI in scientific writing.

Appendix A Metric catalog

The metric suite consists of 72 metrics and corresponding evaluation measures. These equations use the notation introduced in Section 3. They offer compact descriptions instead of complete benchmarking tools. Direction is defined in the standard way assuming fixed parameters; it does not resolve trade-offs or justify cross-task comparisons. The searchable sister includes the same data with reference links.

Table 13. Metric catalog grouped by evaluation question. The companion database contains these same definitions and source links.

Classification at a fixed threshold

Metric and direction	Definition or formula	Interpretation and caution
M01 Accuracy Higher [3, 32]	(TP + TN) / N	Fraction of correct decisions. Majority-class predictions can dominate. Specify decision threshold.
M02 Sensitivity or recall Higher [26, 3]	TP / (TP + FN)	Fraction of reference positives detected. Does not measure false-positive workload. Undefined if no positives.
M03 Specificity Higher [26]	TN / (TN + FP)	Fraction of reference negatives correctly rejected. High values can coexist with many false positives when negatives are numerous.
M04 Precision or PPV Higher [38, 39]	TP / (TP + FP)	Fraction of positive predictions that are correct. Depends on prevalence; undefined with no positive predictions.
M05 Negative predictive value Higher [26]	TN / (TN + FN)	Fraction of negative predictions that are correct. Depends on prevalence and the represented population.
M06 Balanced accuracy Higher [3]	(Sensitivity + Specificity) / 2	Equal average of class recalls in the binary case. Does not encode the actual consequences or prevalence of errors.
M07 F1 score Higher [28, 41]	2TP / (2TP + FP + FN)	Harmonic mean of precision and recall. Ignores TN; class and case reductions alter interpretation.
M08 F beta score Higher [26, 29]	$(1 + \beta^2)TP / [(1 + \beta^2)TP + FP + \beta^2 FN]$	Trades recall against precision with chosen beta. Beta is not automatically a clinical cost ratio.
M09 Matthews correlation coefficient Higher [30, 32, 31]	(TP TN - FP FN) / sqrt[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]	Association of binary labels and predictions; range -1 to 1. Zero denominator requires an explicit software convention.
M10 Youden J Higher [35, 24]	Sensitivity + Specificity - 1	Equal-weight separation of positive and negative classes. Maximizing on the final test set introduces selection bias.
M11 Likelihood ratios LR+ higher and LR- lower [36, 26]	LR+ = sensitivity/(1-specificity); LR- = (1-sensitivity)/specificity	Change in odds from a binary test result under the evaluated setting. Not posterior probabilities; infinite or undefined values need care.
M12 Cohen kappa Higher [42]	(observed agreement - expected agreement) / (1 - expected agreement)	Agreement adjusted for the specified chance model. Affected by marginal distributions; not a universal fairness measure.

Metric and direction	Definition or formula	Interpretation and caution
M13 Weighted kappa Higher [12]	1 - weighted observed disagreement / weighted expected disagreement	Agreement allowing graded errors for ordinal labels. State linear or quadratic weights and label order.
M14 Macro F1 Higher [41]	Mean over class-specific F1 values	Equal class weighting. Different from the harmonic mean of macro precision and macro recall.
M15 Micro F1 Higher [3]	F1 after pooling TP FP FN over labels	Pooled label-decision performance. Equals accuracy for exhaustive single-label multiclass classification.
M16 Multilabel Hamming loss Lower [3]	Incorrect label decisions / (N times K)	Per-label error frequency. Can be small when most labels are absent.
M17 Multilabel subset accuracy Higher [3]	Fraction with every label correct	Exact correctness of an entire label set. Increasing label count makes exact matches harder.

Ranking and probabilities

Metric and direction	Definition or formula	Interpretation and caution
M18 AUROC Higher [35, 36, 37]	Pr(score positive > score negative) + 0.5 Pr(tie)	Pairwise ranking discrimination. Does not establish calibration or performance at one operating point.
M19 Partial AUROC Higher [35]	Integral of TPR over a prespecified FPR interval	Ranking performance in an operating region. Specify bounds and normalization; values are otherwise not comparable.
M20 Average precision Higher [38, 29, 5]	Sum over recall increments of precision times recall increment	A rank-based precision-recall summary. Specify interpolation, ties, matching and prevalence.
M21 Trapezoidal PR area Higher [38, 29]	Trapezoidal integral of precision against recall	Geometric area under sampled PR points. Generally differs from noninterpolated AP and can imply unattainable interpolation.
M22 Brier score Lower [43, 45, 44]	Mean[$(p - y)^2$] for binary outcomes	Squared error of predicted event probabilities. Combines calibration and discrimination; not a pure calibration metric.
M23 Log loss or NLL Lower [44]	-Mean[y log(p) + (1-y) log(1-p)]	Proper probability score strongly penalizing confident errors. Report clipping and binary versus multiclass reduction.
M24 Expected calibration error Lower [50, 48, 51, 52]	Sum over bins of bin weight times absolute calibration gap	Binned estimate of confidence-frequency disagreement. Binning, class reduction and sample size can change the value.
M25 Maximum calibration error Lower [51, 54]	Maximum absolute gap over calibration bins	Largest estimated bin-level mismatch. Unstable for small bins; report bin counts and uncertainty.
M26 Calibration intercept and slope Targets 0 and 1 [47]	logit Pr(Y=1) = a + b logit(p)	Logistic recalibration summary. Needs sufficient data and does not diagnose every local calibration failure.

Metric and direction	Definition or formula	Interpretation and caution
M27 Decision net benefit Higher at a specified threshold [149, 150]	TP/N - FP/N times pt/(1-pt)	Decision value in true-positive-equivalent units. Requires a meaningful action and harm ratio; no arbitrary thresholds.

Semantic segmentation

Metric and direction	Definition or formula	Interpretation and caution
M28 Dice coefficient Higher [59, 4]	2|P intersect G| / (|P| + |G|)	Foreground overlap; equal to foreground F1 for identical binary units. Empty masks, averaging and tiny targets materially change interpretation.
M29 Intersection over union Higher [4]	|P intersect G| / |P union G|	Intersection relative to union. Do not convert mean Dice into mean IoU.
M30 Volumetric similarity Higher [4, 60]	1 - abs(|P|-|G|)/(|P|+|G|)	Agreement of total foreground volumes. Can equal one for disjoint regions of the same volume.
M31 Relative volume error Target 0 [4]	(|P| - |G|) / |G|	Signed fractional oversegmentation or undersegmentation. Undefined for empty reference; sign cancellation can hide case errors.
M32 Hausdorff distance Lower [63, 64]	max of the two directed maximum surface distances	Largest nearest-surface discrepancy. Very sensitive to outliers; physical spacing and empty sets matter.
M33 HD95 Lower [4, 2]	Specified 95th percentile of bidirectional surface distances	Tail boundary error with reduced sensitivity to a few extremes. Pooled percentile differs from maximum of directional percentiles.
M34 Average symmetric surface distance Lower [4, 65]	Mean bidirectional nearest-surface distance under stated weights	Average boundary discrepancy. Point weighting and surface area weighting are different estimators.
M35 Normalized surface Dice Higher [66, 65]	Surface area within tolerance / total surface area	Fraction of boundaries within a justified physical tolerance. Tolerance and surface discretization must be reported.
M36 Boundary F score Higher [67]	Harmonic mean of boundary precision and boundary recall	Matched contour quality within a distance tolerance. Matching and thinning conventions alter the result.
M37 Boundary IoU Higher [68]	IoU of the specified interior boundary bands	Overlap focused on object boundaries. Band width and image resolution influence sensitivity.
M38 clDice Higher [74]	2 times topology precision times topology sensitivity / their sum	Skeleton-mask consistency for tubular structures. Depends on skeletonization and does not replace spatial accuracy.
M39 Betti number error Lower [75]	Difference between predicted and reference component or hole counts	Topological discrepancy in selected dimensions. Equal counts do not guarantee correct connectivity or location.

Instance segmentation

Metric and direction	Definition or formula	Interpretation and caution

Metric and direction	Definition or formula	Interpretation and caution
M40 Aggregated Jaccard index Higher [70, 15]	Matched intersection sum / (matched union sum + unmatched prediction area)	Combined delineation and object-detection quality under the stated matching rule. Original AJI and one-to-one variants differ.
M41 Panoptic quality Higher [13]	Sum of matched IoUs / [TP + 0.5 FP + 0.5 FN]	Joint recognition and segmentation quality. Use the stated matching threshold and category/void handling.
M42 Segmentation quality Higher [13]	Mean IoU over matched instances	Conditional quality of successfully matched objects. Can improve when difficult objects are missed.
M43 Recognition quality Higher [13]	TP / (TP + 0.5 FP + 0.5 FN)	Object-level F1 under the matching protocol. Does not reflect within-match boundary quality.
M44 Adjusted Rand index Higher [71, 72]	Chance-adjusted pairwise partition agreement	Similarity of two partitions. Background and object size can dominate pair counts.
M45 Variation of information Lower [73]	H(G|P) + H(P|G)	Information lost through partition splits and merges. Specify logarithm base, domain and background handling.

Detection and box localization

Metric and direction	Definition or formula	Interpretation and caution
M46 Box IoU Higher [16, 17]	Intersection area or volume / union area or volume	Spatial agreement of two boxes. Small shifts penalize small boxes more strongly; assignment remains separate.
M47 AP50 and AP75 Higher [16, 5]	Detection AP at IoU thresholds 0.50 and 0.75	Detection quality at a chosen localization requirement. Duplicate predictions are false positives after one-to-one matching.
M48 COCO style AP Higher [17]	Mean AP across categories and IoU 0.50 to 0.95 in steps of 0.05	Benchmark summary across localization requirements. Requires the exact evaluator including recall sampling, ignored regions and max detections.
M49 Average recall Higher [17, 5]	Mean recall over stated IoU thresholds and detection budgets	Object recovery given a proposal or detection limit. A larger maximum-detection budget can improve recall.
M50 False positives per image or scan Lower [86, 88]	Total unmatched predictions / number of images or scans	False-alert burden. Denominator and postprocessing must be declared.
M51 FROC sensitivity Higher [86, 87, 88]	Matched lesions / eligible reference lesions at a specified FP rate	Lesion detection versus false-alert workload. No universal maximum false-positive axis or universal FROC area.
M52 LROC performance Higher [91]	Correctly localized detection versus case false-positive rate	Combines case-level detection and correct location. Depends on the number of permitted marks and success criterion.
M53 Localization recall precision error Lower [79, 80]	Normalized localization error plus FP and FN penalties per evaluated object	Decomposable detection error at a confidence threshold. Report localization threshold and how the confidence threshold is selected.

Metric and direction	Definition or formula	Interpretation and caution
M54 Probabilistic detection quality Higher [81, 82]	Assignment-based combination of spatial and label quality	Quality of probabilistic object detections. Requires compatible spatial probabilities; not an ordinary box AP substitute.
M55 Generalized IoU Higher [83]	IoU - area(C minus union)/area(C)	Box relation including empty space inside the enclosing box C. Often a training objective; can be negative and is not AP.
M56 Distance IoU Higher [84, 85]	IoU - squared center distance / squared enclosing-box diagonal	Box overlap plus center displacement. Depends on box geometry; does not evaluate object recall.

Landmarks and geometric localization

Metric and direction	Definition or formula	Interpretation and caution
M57 Mean radial error Lower [18, 19]	Mean Euclidean distance between corresponding landmarks	Average positional error in pixels or physical units. Units, correspondence and failed-localization policy must be explicit.
M58 Root mean square landmark error Lower [19]	sqrt(mean squared Euclidean landmark distance)	Positional error giving more weight to large failures. A small mean may coexist with a clinically important tail.
M59 Normalized mean error Lower [93, 92]	Mean landmark distance / prespecified anatomical scale	Relative localization error across sizes. Normalization by head size or image diagonal yields different quantities.
M60 Successful detection rate Higher [18]	Fraction of eligible landmarks within tolerance tau	Localization success at a physical tolerance. Exclude neither failed predictions nor difficult eligible landmarks silently.
M61 Percentage of correct keypoints Higher [93, 92]	Fraction of keypoints within a normalized distance threshold	Pose correctness relative to a specified scale. Report PCK versus PCKh, normalization and visibility policy.
M62 Object keypoint similarity Higher [17]	Weighted average of exp(-distance^2 / scale-and-keypoint variance term)	Scale-normalized similarity using keypoint-specific tolerances. Requires benchmark sigmas, area scale and visibility conventions.
M63 Target registration error Lower [19]	Distance between corresponding independent target points after registration	Localization accuracy away from fitted fiducials. Fiducial fit error is not the same as independent target error.

Uncertainty and workflow

Metric and direction	Definition or formula	Interpretation and caution
M64 Selective risk Lower at specified coverage [105, 106, 107]	Mean task loss among accepted cases	Residual error after uncertain cases are deferred. Report coverage and errors/workload among rejected cases.
M65 Coverage Task dependent [107, 108]	Accepted cases / all eligible cases	Fraction handled automatically. High retained accuracy at very low coverage can be operationally unhelpful.

Metric and direction	Definition or formula	Interpretation and caution
M66 Area under risk coverage curve Lower [106, 107]	Integral of selective risk with respect to coverage	Quality of the ordering used for abstention. Depends on task loss, discretization and underlying error rate.
M67 Prediction set coverage Target at least nominal level [110, 111]	Fraction with true outcome inside the predicted set	Empirical coverage of prediction sets. Marginal guarantees do not automatically imply subgroup or shifted-domain coverage.
M68 Prediction set size Lower at adequate coverage [110, 112]	Mean number or volume of candidate outcomes in the prediction set	Efficiency of uncertainty sets. Small sets without adequate coverage are not a success.
M69 Generalized energy distance Lower under a valid distance [103, 104]	2 E[d(S,Y)] - E[d(S,Sprime)] - E[d(Y,Yprime)]	Comparison of predicted and reference segmentation distributions. Distance assumptions, finite annotation samples and diversity all matter.
M70 Bland Altman limits of agreement Narrower within acceptable bounds [139]	Mean paired difference ± 1.96 SD of differences under suitable assumptions	Bias and dispersion of derived clinical measurements. Check heteroscedasticity and repeated measures; correlation is insufficient.
M71 Intraclass correlation coefficient Higher [140]	Model-specific ratio of between-subject and total variance components	Reliability of repeated quantitative measurements. Specify model, absolute agreement versus consistency, and single versus average measurement.
M72 Correction time and failure rate Lower [141, 142]	Human editing time and proportion of cases failing an acceptance rule	Practical effort and unusable outputs. Time must cover the intended workflow and representative readers.

Appendix B Proposed evaluation reporting checklist

The following 32 items are intended for protocol planning and manuscript review. Responses should record evidence and a source location. Suitable categories are documented, unclear, not located in the inspected material, and not applicable with a reason. Not located should not be converted into not performed. The checklist has not undergone expert consensus, reliability testing or validation and should not be summed into a quality score.

Table 14. Proposed 32-item reporting checklist. Not a validated score, standard or certification.

Target and reference

ID	Item	Evidence to report
C1.1	Intended question	Define the decision, measurement or scientific claim.
C1.2	Eligible population	Describe inclusion, exclusions, setting and target prevalence.
C1.3	Reference process	Identify annotators or diagnostic standards, uncertainty and adjudication.
C1.4	Unit and hierarchy	State image, patient, lesion, pixel, instance or landmark units and nesting.

Partitions and information

ID	Item	Evidence to report
C2.1	Independent groups	Keep patient, subject, slide or acquisition groups intact across splits.
C2.2	Final test protection	Exclude final test labels from tuning, thresholding and calibration.
C2.3	Pretraining provenance	Disclose known overlap and uncertainty about training repositories.
C2.4	Inference information	Document crops, prompts, support examples and human input.

Metric specification

ID	Item	Evidence to report
C3.1	Primary endpoint	Prespecify the principal metric and justify its relationship to the task.
C3.2	Complementary endpoints	Select measures of distinct relevant failure modes.
C3.3	Thresholds and tolerances	State decision thresholds, matching IoUs and physical tolerances.
C3.4	Estimator details	Identify formulas, interpolation, tie rules and software versions.

Spatial and object evaluation

ID	Item	Evidence to report

ID	Item	Evidence to report
C4.1	Coordinates	Record spacing, orientation, resampling and inverse transforms.
C4.2	Correspondence	Specify class compatibility, assignment and matching rules.
C4.3	Duplicate and ignored outputs	Describe duplicate boxes, ignored regions, crowd labels and budgets.
C4.4	Structural failures	Count missing objects, splits, merges, topology errors and failed landmarks.

Aggregation and edge cases

ID	Item	Evidence to report
C5.1	Reduction order	State case, class, center and voxel aggregation in order.
C5.2	Weighting	Report class weights, patient weights and size effects.
C5.3	Undefined values	Define absent classes, empty masks and zero denominators.
C5.4	Missing predictions	Retain failures and disclose exclusions with denominators.

Statistics and comparison

ID	Item	Evidence to report
C6.1	Paired effects	Compare the same cases when possible and report effect sizes.
C6.2	Uncertainty method	Use intervals and resampling appropriate to the independent unit.
C6.3	Variability sources	Separate seed variation, patient sampling and individual prediction uncertainty.
C6.4	Selection and multiplicity	Account for tuning, repeated comparisons and exploratory analyses.

Probability and robustness

ID	Item	Evidence to report
C7.1	Calibration event	Define the event, binning or estimator and independent calibration partition.
C7.2	Selective performance	Report coverage, retained risk, referral counts and rejected-case outcomes.
C7.3	Shift and subgroups	Define relevant shifts and subgroup denominators with uncertainty.
C7.4	Guarantee assumptions	State assumptions and empirical checks for conformal or risk guarantees.

Use and reproducibility

ID	Item	Evidence to report

ID	Item	Evidence to report
C8.1	Downstream outcomes	Measure intended clinical or scientific quantities and acceptable error.
C8.2	Workflow effort	Include correction time, failures, retries and end-to-end resources.
C8.3	Reproducible release	Provide permitted predictions, evaluator configuration and provenance.
C8.4	Limitations and maintenance	Disclose scope limits, protocol deviations and version changes.

Appendix C Implementing a reproducible evaluation

Freeze the prediction archive and reference version before calculating final results. Record patient or subject identifiers in a permitted form, class mappings, spatial metadata, confidence values and every missing prediction. Keep transformations from model coordinates to evaluation coordinates explicit. Create a manifest that associates each prediction with exactly one eligible reference unit.

Run basic conformance cases before evaluating the archive. These should include perfect prediction, no prediction, all-background prediction, a shifted object, a remote false-positive object, duplicated detection, tied scores, a missing class, anisotropic coordinates and failed landmark output. The expected result or undefined-value policy should be written before the software output is inspected.

For each endpoint, save the evaluator version, parameters, per-case values and aggregation procedure. Preserve paired case-level results across models. Apply the prespecified statistical analysis and publish the denominator of every reported summary. Keep external, temporal, adapted and internal results in separately labeled comparisons. Recompute a small independent subset by hand or with another implementation when a result is sensitive to conventions.

References

[1] Maier-Hein L, Reinke A, Godau P, Tizabi MD, Buettner F, Christodoulou E, et al. Metrics reloaded: recommendations for image analysis validation. Nature methods. 2024;21(2):195-212. 10.1038/s41592-023-02151-z

[2] Reinke A, Tizabi MD, Baumgartner M, Eisenmann M, Heckmann-Nötzel D, Kavur AE, et al. Understanding metric-related pitfalls in image analysis validation. Nature methods. 2024;21(2):182-194. 10.1038/s41592-023-02150-0

[3] Marina Sokolova, Guy Lapalme. A systematic analysis of performance measures for classification tasks. Information Processing & Management. 2009;45(4):427-437. 10.1016/j.ipm.2009.03.002

[4] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC medical imaging. 2015;15:29. 10.1186/s12880-015-0068-x

[5] Rafael Padilla, Sergio L. Netto, Eduardo A. B. da Silva. A Survey on Performance Metrics for Object-Detection Algorithms. 2020 International Conference on Systems, Signals and Image Processing (IWSSIP). 2020;:237-242. 10.1109/iwssip48289.2020.9145130

[6] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods. 2021;18(2):203-211. 10.1038/s41592-020-01008-z

[7] Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick. Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020;42(2):386-397. 10.1109/tpami.2018.2844175

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, et al. Segment Anything. 2023 IEEE/CVF International Conference on Computer Vision (ICCV). 2023;:3992-4003. 10.1109/iccv51070.2023.00371

[9] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, Bo Wang. Segment anything in medical images. Nature Communications. 2024;15(1):654. 10.1038/s41467-024-44824-z

[10] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ (Clinical research ed.). 2021;372:n71. 10.1136/bmj.n71

[11] Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews. Systematic reviews. 2021;10(1):39. 10.1186/s13643-020-01542-z

[12] Jacob Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin. 1968;70(4):213-220. 10.1037/h0026256

[13] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, Piotr Dollár. Panoptic Segmentation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019;:9396-9405. 10.1109/cvpr.2019.00963

[14] Graham S, Vu QD, Raza SEA, Azam A, Tsang YW, Kwak JT, et al. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical image analysis. 2019;58:101563. 10.1016/j.media.2019.101563

[15] Verma R, Kumar N, Patil A, Kurian NC, Rane S, Graham S, et al. MoNuSAC2020: A Multi-Organ Nuclei Segmentation and Classification Challenge. IEEE transactions on medical imaging. 2021;40(12):3413-3423. 10.1109/tmi.2021.3085712

[16] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, Andrew Zisserman. The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision. 2010;88(2):303-338. 10.1007/s11263-009-0275-4

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, et al. Microsoft COCO: Common Objects in Context. Lecture Notes in Computer Science. 2014;:740-755. 10.1007/978-3-319-10602-1_48

[18] Wang CW, Huang CT, Lee JH, Li CH, Chang SW, Siao MJ, et al. A benchmark for comparison of dental radiography analysis algorithms. Medical image analysis. 2016;31:63-76. 10.1016/j.media.2016.02.004

[19] J.M. Fitzpatrick, J.B. West, C.R. Maurer. Predicting error in rigid-body point-based registration. IEEE Transactions on Medical Imaging. 1998;17(5):694-702. 10.1109/42.736021

[20] Mabrok MA. Side-level versus patient-level evaluation in four-view mammography classification: a comprehensive benchmark on the CMMD dataset. Future science OA. 2026;12(1):2681575. 10.1080/20565623.2026.2681575

[21] Nancy A. Obuchowski. Nonparametric Analysis of Clustered ROC Curve Data. Biometrics. 1997;53(2):567. 10.2307/2533958

[22] S.K. Warfield, K.H. Zou, W.M. Wells. Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation. IEEE Transactions on Medical Imaging. 2004;23(7):903-921. 10.1109/tmi.2004.828354

[23] Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC bioinformatics. 2006;7:91. 10.1186/1471-2105-7-91

[24] Gavin C. Cawley, Nicola L. C. Talbot. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. Journal of Machine Learning Research. 2010;11(70):2079-2107. Original publication record

[25] Maier-Hein L, Eisenmann M, Reinke A, Onogur S, Stankovic M, Scholz P, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. Nature communications. 2018;9(1):5217. 10.1038/s41467-018-07619-7

[26] Powers, David M. W.. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv. 2020. 10.48550/arXiv.2010.16061

[27] Edwin B. Wilson. Probable Inference, the Law of Succession, and Statistical Inference. Journal of the American Statistical Association. 1927;22(158):209-212. 10.1080/01621459.1927.10502953

[28] Lipton, Zachary Chase, Elkan, Charles, Narayanaswamy, Balakrishnan. Thresholding Classifiers to Maximize F1 Score. arXiv. 2014. 10.48550/arXiv.1402.1892

[29] Peter Flach, Meelis Kull. Precision-Recall-Gain Curves: PR Analysis Done Right. Advances in Neural Information Processing Systems. 2015;28. Original publication record

[30] B.W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure. 1975;405(2):442-451. 10.1016/0005-2795(75)90109-9

[31] J. Gorodkin. Comparing two K-category assignments by a K-category correlation coefficient. Computational Biology and Chemistry. 2004;28(5-6):367-374. 10.1016/j.compbiolchem.2004.09.006

[32] Davide Chicco, Giuseppe Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6. 10.1186/s12864-019-6413-7

[33] Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. Proceedings of the ACM Conference on Health, Inference, and Learning. 2020;2020:151-159. 10.1145/3368555.3384468

[34] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine. 2021;27(12):2176-2182. 10.1038/s41591-021-01595-0

[35] Tom Fawcett. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27(8):861-874. 10.1016/j.patrec.2005.10.010

[36] J A Hanley, B J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36. 10.1148/radiology.143.1.7063747

[37] David J. Hand. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning. 2009;77(1):103-123. 10.1007/s10994-009-5119-5

[38] Jesse Davis, Mark Goadrich. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning - ICML '06. 2006;:233-240. 10.1145/1143844.1143874

[39] Takaya Saito, Marc Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE. 2015;10(3):e0118432. 10.1371/journal.pone.0118432

[40] David J. Hand, Robert J. Till. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning. 2001;45(2):171-186. 10.1023/a:1010920819831

[41] Opitz, Juri, Burst, Sebastian. Macro F1 and Macro F1. arXiv. 2019. 10.48550/arXiv.1911.03347

[42] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960;20(1):37-46. 10.1177/001316446002000104

[43] GLENN W. BRIER. Verification of forecasts expressed in terms of probability. Monthly Weather Review. 1950;78(1):1-3. 10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2

[44] Tilmann Gneiting, Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. 2007;102(477):359-378. 10.1198/016214506000001437

[45] Allan H. Murphy. A New Vector Partition of the Probability Score. Journal of Applied Meteorology. 1973;12(4):595-600. 10.1175/1520-0450(1973)012<0595:anvpot>2.0.co;2

[46] Morris H. DeGroot, Stephen E. Fienberg. The Comparison and Evaluation of Forecasters. The Statistician. 1983;32(1/2):12. 10.2307/2987588

[47] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC medicine. 2019;17(1):230. 10.1186/s12916-019-1466-7

[48] Vaicenavicius, Juozas, Widmann, David, Andersson, Carl, Lindsten, Fredrik, Roll, Jacob, Schön, Thomas B.. Evaluating model calibration in classification. arXiv. 2019. 10.48550/arXiv.1902.06977

[49] Mehrtash A, Wells WM, Tempany CM, Abolmaesumi P, Kapur T. Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation. IEEE transactions on medical imaging. 2020;39(12):3868-3878. 10.1109/tmi.2020.3006437

[50] Chuan Guo, Geoff Pleiss, Yu Sun, Kilian Q. Weinberger. On Calibration of Modern Neural Networks. Proceedings of Machine Learning Research. 2017;70:1321–1330. Original publication record

[51] Nixon, Jeremy, Dusenberry, Mike, Jerfel, Ghassen, Nguyen, Timothy, Liu, Jeremiah, Zhang, Linchuan, et al. Measuring Calibration in Deep Learning. arXiv. 2019. 10.48550/arXiv.1904.01685

[52] Roelofs, Rebecca, Cain, Nicholas, Shlens, Jonathon, Mozer, Michael C.. Mitigating Bias in Calibration Error Estimation. arXiv. 2020. 10.48550/arXiv.2012.08668

[53] Aviral Kumar, Sunita Sarawagi, Ujjwal Jain. Trainable Calibration Measures for Neural Networks from Kernel Mean Embeddings. Proceedings of Machine Learning Research. 2018;80:2805-2814. Original publication record

[54] Kumar, Ananya, Liang, Percy, Ma, Tengyu. Verified Uncertainty Calibration. arXiv. 2019. 10.48550/arXiv.1909.10155

[55] Mahdi Pakdaman Naeini, Gregory Cooper, Milos Hauskrecht. Obtaining Well Calibrated Probabilities Using Bayesian Binning. Proceedings of the AAAI Conference on Artificial Intelligence. 2015;29(1). 10.1609/aaai.v29i1.9602

[56] Kull, Meelis, Perello-Nieto, Miquel, Kängsepp, Markus, Filho, Telmo Silva, Song, Hao, Flach, Peter. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. arXiv. 2019. 10.48550/arXiv.1910.12656

[57] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, et al. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. Advances in Neural Information Processing Systems. 2019;32. [Conference paper; arXiv version] 10.48550/arXiv.1906.02530

[58] Goetz L, Seedat N, Vandersluis R, van der Schaar M. Generalization-a key challenge for responsible AI in patient-facing clinical applications. NPJ digital medicine. 2024;7(1):126. 10.1038/s41746-024-01127-3

[59] Lee R. Dice. Measures of the Amount of Ecologic Association Between Species. Ecology. 1945;26(3):297-302. 10.2307/1932409

[60] Crum WR, Camara O, Hill DL. Generalized overlap measures for evaluation and validation in medical image analysis. IEEE transactions on medical imaging. 2006;25(11):1451-1461. 10.1109/tmi.2006.880587

[61] Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge Cardoso M. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. Deep learning in medical image analysis and multimodal learning for clinical decision support : Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, held in conjunction with MICCAI 2017 Quebec City, QC,. 2017;2017:240-248. 10.1007/978-3-319-67558-9_28

[62] Ma J, Chen J, Ng M, Huang R, Li Y, Li C, et al. Loss odyssey in medical image segmentation. Medical image analysis. 2021;71:102035. 10.1016/j.media.2021.102035

[63] D.P. Huttenlocher, G.A. Klanderman, W.J. Rucklidge. Comparing images using the Hausdorff distance. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1993;15(9):850-863. 10.1109/34.232073

[64] M.-P. Dubuisson, A.K. Jain. A modified Hausdorff distance for object matching. Proceedings of 12th International Conference on Pattern Recognition. 1994;1:566-568. 10.1109/icpr.1994.576361

[65] Yeghiazaryan V, Voiculescu I. Family of boundary overlap metrics for the evaluation of medical image segmentation. Journal of medical imaging (Bellingham, Wash.). 2018;5(1):015006. 10.1117/1.jmi.5.1.015006

[66] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, et al. Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study. Journal of Medical Internet Research. 2021;23(7):e26151. 10.2196/26151

[67] D.R. Martin, C.C. Fowlkes, J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2004;26(5):530-549. 10.1109/tpami.2004.1273918

[68] Bowen Cheng, Ross Girshick, Piotr Dollar, Alexander C. Berg, Alexander Kirillov. Boundary IoU: Improving Object-Centric Image Segmentation Evaluation. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021;:15329-15337. 10.1109/cvpr46437.2021.01508

[69] Uwe Schmidt, Martin Weigert, Coleman Broaddus, Gene Myers. Cell Detection with Star-Convex Polygons. Lecture Notes in Computer Science. 2018;:265-273. 10.1007/978-3-030-00934-2_30

[70] Kumar N, Verma R, Sharma S, Bhargava S, Vahadane A, Sethi A. A Dataset and a Technique for Generalized Nuclear Segmentation for Computational Pathology. IEEE transactions on medical imaging. 2017;36(7):1550-1560. 10.1109/tmi.2017.2677499

[71] William M. Rand. Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association. 1971;66(336):846-850. 10.1080/01621459.1971.10482356

[72] Lawrence Hubert, Phipps Arabie. Comparing partitions. Journal of Classification. 1985;2(1):193-218. 10.1007/bf01908075

[73] Marina Meilă. Comparing clusterings—an information based distance. Journal of Multivariate Analysis. 2007;98(5):873-895. 10.1016/j.jmva.2006.11.013

[74] Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, et al. clDice - a Novel Topology-Preserving Loss Function for Tubular Structure Segmentation. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021;:16555-16564. 10.1109/cvpr46437.2021.01629

[75] Clough JR, Byrne N, Oksuz I, Zimmer VA, Schnabel JA, King AP. A Topological Loss Function for Deep-Learning Based Image Segmentation Using Persistent Homology. IEEE transactions on pattern analysis and machine intelligence. 2022;44(12):8766-8778. 10.1109/tpami.2020.3013679

[76] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, Andrew Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. International Journal of Computer Vision. 2015;111(1):98-136. 10.1007/s11263-014-0733-5

[77] Derek Hoiem, Yodsawalai Chodpathumwan, Qieyun Dai. Diagnosing Error in Object Detectors. Lecture Notes in Computer Science. 2012;:340-353. 10.1007/978-3-642-33712-3_25

[78] Daniel Bolya, Sean Foley, James Hays, Judy Hoffman. TIDE: A General Toolbox for Identifying Object Detection Errors. Lecture Notes in Computer Science. 2020;:558-573. 10.1007/978-3-030-58580-8_33

[79] Kemal Oksuz, Baris Can Cam, Emre Akbas, Sinan Kalkan. Localization Recall Precision (LRP): A New Performance Metric for Object Detection. Lecture Notes in Computer Science. 2018;:521-537. 10.1007/978-3-030-01234-2_31

[80] Oksuz K, Cam BC, Kalkan S, Akbas E. One Metric to Measure Them All: Localisation Recall Precision (LRP) for Evaluating Visual Detection Tasks. IEEE transactions on pattern analysis and machine intelligence. 2022;44(12):9446-9463. 10.1109/tpami.2021.3130188

[81] David Hall, Feras Dayoub, John Skinner, Haoyang Zhang, Dimity Miller, Peter Corke, et al. Probabilistic Object Detection: Definition and Evaluation. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). 2020;:1020-1029. 10.1109/wacv45572.2020.9093599

[82] Di Feng, Ali Harakeh, Steven L. Waslander, Klaus Dietmayer. A Review and Comparative Study on Probabilistic Object Detection in Autonomous Driving. IEEE Transactions on Intelligent Transportation Systems. 2022;23(8):9961-9980. 10.1109/tits.2021.3096854

[83] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, Silvio Savarese. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019;:658-666. 10.1109/cvpr.2019.00075

[84] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, Dongwei Ren. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. Proceedings of the AAAI Conference on Artificial Intelligence. 2020;34(07):12993-13000. 10.1609/aaai.v34i07.6999

[85] Zheng Z, Wang P, Ren D, Liu W, Ye R, Hu Q, et al. Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation. IEEE transactions on cybernetics. 2022;52(8):8574-8586. 10.1109/tcyb.2021.3095305

[86] Chakraborty DP. Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data. Medical physics. 1989;16(4):561-568. 10.1118/1.596358

[87] Chakraborty DP, Winter LH. Free-response methodology: alternate analysis and a new observer-performance experiment. Radiology. 1990;174(3 Pt 1):873-881. 10.1148/radiology.174.3.2305073

[88] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas de Bel, Moira S.N. Berens, Cas van den Bogaard, Piergiorgio Cerello, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. Medical Image Analysis. 2017;42:1-13. 10.1016/j.media.2017.06.015

[89] Samuel G. Armato, Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans. Medical Physics. 2011;38(2):915-931. 10.1118/1.3528204

[90] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA. 2017;318(22):2199. 10.1001/jama.2017.14585

[91] Liu K, Zhou XH. Nonparametric inference for the localization receiver operating characteristic curve and its extension to free-response image localization tasks. Statistical methods in medical research. 2026;35(9):1932-1958. 10.1177/09622802261468028

[92] Yi Yang, Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. CVPR 2011. 2011;:1385-1392. 10.1109/cvpr.2011.5995741

[93] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, Bernt Schiele. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. 2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014;:3686-3693. 10.1109/cvpr.2014.471

[94] Adebayo, Julius, Gilmer, Justin, Muelly, Michael, Goodfellow, Ian, Hardt, Moritz, Kim, Been. Sanity Checks for Saliency Maps. arXiv. 2018. 10.48550/arXiv.1810.03292

[95] Choe, Junsuk, Oh, Seong Joon, Lee, Seungho, Chun, Sanghyuk, Akata, Zeynep, Shim, Hyunjung. Evaluating Weakly Supervised Object Localization Methods Right. arXiv. 2020. 10.48550/arXiv.2001.07437

[96] Zhou Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing. 2004;13(4):600-612. 10.1109/tip.2003.819861

[97] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018;:586-595. 10.1109/cvpr.2018.00068

[98] Heusel, Martin, Ramsauer, Hubert, Unterthiner, Thomas, Nessler, Bernhard, Hochreiter, Sepp. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. arXiv. 2017. 10.48550/arXiv.1706.08500

[99] Behnam Kiani Kalejahi, Sebelan Danishvar, Mohammad Javad Rajabi. GAN-Based Cross-Modality Brain MRI Synthesis: Paired Versus Unpaired Training and Comparison with Diffusion and Transformer Models. Biomimetics. 2026;11(3):175. 10.3390/biomimetics11030175

[100] Mutepfe F, Kalejahi BK, Meshgini S, Danishvar S. Generative Adversarial Network Image Synthesis Method for Skin Lesion Generation and Classification. Journal of medical signals and sensors. 2021;11(4):237-252. 10.4103/jmss.jmss_53_20

[101] Gal, Yarin, Ghahramani, Zoubin. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. arXiv. 2015. 10.48550/arXiv.1506.02142

[102] Lakshminarayanan, Balaji, Pritzel, Alexander, Blundell, Charles. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. arXiv. 2016. 10.48550/arXiv.1612.01474

[103] Kohl, Simon A. A., Romera-Paredes, Bernardino, Meyer, Clemens, De Fauw, Jeffrey, Ledsam, Joseph R., Maier-Hein, Klaus H., et al. A Probabilistic U-Net for Segmentation of Ambiguous Images. arXiv. 2018. 10.48550/arXiv.1806.05034

[104] Baumgartner, Christian F., Tezcan, Kerem C., Chaitanya, Krishna, Hötker, Andreas M., Muehlematter, Urs J., Schawkat, Khoschy, et al. PHiSeg: Capturing Uncertainty in Medical Image Segmentation. arXiv. 2019. 10.48550/arXiv.1906.04045

[105] C. Chow. On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory. 1970;16(1):41-46. 10.1109/tit.1970.1054406

[106] Geifman, Yonatan, El-Yaniv, Ran. Selective Classification for Deep Neural Networks. arXiv. 2017. 10.48550/arXiv.1705.08500

[107] Geifman, Yonatan, El-Yaniv, Ran. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. arXiv. 2019. 10.48550/arXiv.1901.09192

[108] Behnam Kiani Kalejahi, Sajid Khan, Murodbek Akhrorov, Mohammad Javad Rajabi, Ahmed Aziz. Uncertainty-Aware Prediction Across Endoscopic Domains: Laryngeal Narrow-Band and Gastrointestinal Imaging. Biomedicines. 2026;14(7):1616. 10.3390/biomedicines14071616

[109] Melendez-Corres G, Wahi-Anwar M, Kim J, Tada D, Hoffman J, Genender J, et al. Selective segmentation with rejection option: a machine reasoning approach and evaluation metrics in CT kidney segmentation. Journal of medical imaging (Bellingham, Wash.). 2026;13(5):054501. 10.1117/1.jmi.13.5.054501

[110] Anastasios N. Angelopoulos, Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. arXiv. 2021. 10.48550/arXiv.2107.07511

[111] Angelopoulos, Anastasios N., Bates, Stephen, Fisch, Adam, Lei, Lihua, Schuster, Tal. Conformal Risk Control. arXiv. 2022. 10.48550/arXiv.2208.02814

[112] Bates, Stephen, Angelopoulos, Anastasios, Lei, Lihua, Malik, Jitendra, Jordan, Michael I.. Distribution-Free, Risk-Controlling Prediction Sets. arXiv. 2021. 10.48550/arXiv.2101.02703

[113] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, Ali Gholipour. Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. Lecture Notes in Computer Science. 2017;:379-387. 10.1007/978-3-319-67389-9_44

[114] Yeung M, Sala E, Schönlieb CB, Rundo L. Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society. 2022;95:102026. 10.1016/j.compmedimag.2021.102026

[115] Kervadec H, Bouchtiba J, Desrosiers C, Granger E, Dolz J, Ben Ayed I. Boundary loss for highly unbalanced segmentation. Medical image analysis. 2021;67:101851. 10.1016/j.media.2020.101851

[116] Karimi D, Karimi D, Salcudean SE. Reducing the Hausdorff Distance in Medical Image Segmentation With Convolutional Neural Networks. IEEE transactions on medical imaging. 2020;39(2):499-513. 10.1109/tmi.2019.2930068

[117] Isensee, Fabian, Rokuss, Maximilian, Krämer, Lars, Dinkelacker, Stefan, Ravindran, Ashis, Stritzke, Florian, et al. nnInteractive: Redefining 3D Promptable Segmentation. arXiv. 2025. 10.48550/arXiv.2503.08373

[118] Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (New York, N.Y.). 2023;4(9):100804. 10.1016/j.patter.2023.100804

[119] Maier-Hein L, Reinke A, Kozubek M, Martel AL, Arbel T, Eisenmann M, et al. BIAS: Transparent reporting of biomedical image analysis challenges. Medical image analysis. 2020;66:101796. 10.1016/j.media.2020.101796

[120] Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE transactions on medical imaging. 2015;34(10):1993-2024. 10.1109/tmi.2014.2377694

[121] Antonelli M, Reinke A, Bakas S, Farahani K, Kopp-Schneider A, Landman BA, et al. The Medical Segmentation Decathlon. Nature communications. 2022;13(1):4128. 10.1038/s41467-022-30695-9

[122] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, et al. Shortcut learning in deep neural networks. Nature Machine Intelligence. 2020;2(11):665-673. 10.1038/s42256-020-00257-z

[123] Hardt, Moritz, Price, Eric, Srebro, Nathan. Equality of Opportunity in Supervised Learning. arXiv. 2016. 10.48550/arXiv.1610.02413

[124] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, Kilian Q. Weinberger. On Fairness and Calibration. Advances in Neural Information Processing Systems. 2017;30. Original publication record

[125] Behnam Kiani Kalejahi, Sajid Khan, Rakhim Zakirov. Bone-CNN: A Lightweight Deep Learning Architecture for Multi-Class Classification of Primary Bone Tumours in Radiographs. Biomedicines. 2026;14(2):299. 10.3390/biomedicines14020299

[126] Kalejahi BK, Rajabi MJ. Multimodal deep learning for papillary thyroid carcinoma diagnosis using ultrasound and cytology. BMC medical imaging. 2026;26(1):449. 10.1186/s12880-026-02526-6

[127] Behnam Kiani Kalejahi, Shahri Yahyayeva, Inara Abdurahmanova, Sadig Hasanzade. Lightweight deep learning models for early plant disease detection: a transfer learning approach with mobilenetv2 and resnet50. Problems of Information Society. 2026;17(1):11-23. 10.25045/jpis.v17.i1.02

[128] Behnam Kiani Kalejahi, Saeed Meshgini, Sebelan Danishvar. Segmentation of Brain Tumor Using a 3D Generative Adversarial Network. Diagnostics. 2023;13(21):3344. 10.3390/diagnostics13213344

[129] Behnam Kiani Kalejahi, Saeed Meshgini, Sebelan Danishvar. Brain tumor segmentation by auxiliary classifier generative adversarial network. Signal, Image and Video Processing. 2023;17(7):3339-3345. 10.1007/s11760-023-02555-6

[130] Behnam Kiani Kalejahi, Sajid Khan, Mohammad Javad Rajabi. GDNet: A Robust 2.5D Multimodal MRI Brain Tumor Segmentation Framework with EMA Stabilization and Tumor-Aware Sampling. Journal of Imaging. 2026;12(7):288. 10.3390/jimaging12070288

[131] Behnam Kiani Kalejahi, Mohammad Javad Rajabi. Deep learning for brain tumour segmentation on MRI with focus on robustness uncertainty and clinical translation. Discover Artificial Intelligence. 2026. 10.1007/s44163-026-01713-2

[132] B. Efron. Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics. 1979;7(1). 10.1214/aos/1176344552

[133] Elizabeth R. DeLong, David M. DeLong, Daniel L. Clarke-Pearson. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988;44(3):837. 10.2307/2531595

[134] Quinn McNemar. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. Psychometrika. 1947;12(2):153-157. 10.1007/bf02295996

[135] Yoshua Bengio, Yves Grandvalet. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. Journal of Machine Learning Research. 2004;5:1089-1105. Original publication record

[136] Thomas G. Dietterich. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation. 1998;10(7):1895-1923. 10.1162/089976698300017197

[137] Janez Demšar. Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research. 2006;7(1):1-30. Original publication record

[138] Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, et al. Video-based AI for beat-to-beat assessment of cardiac function. Nature. 2020;580(7802):252-256. 10.1038/s41586-020-2145-8

[139] J. Martin Bland, DouglasG. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet. 1986;327(8476):307-310. 10.1016/s0140-6736(86)90837-8

[140] Terry K. Koo, Mae Y. Li. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine. 2016;15(2):155-163. 10.1016/j.jcm.2016.02.012

[141] Vaassen F, Hazelaar C, Vaniqui A, Gooding M, van der Heyden B, Canters R, et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Physics and imaging in radiation oncology. 2020;13:1-6. 10.1016/j.phro.2019.12.001

[142] Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature medicine. 2022;28(5):924-933. 10.1038/s41591-022-01772-9

[143] Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nature medicine. 2020;26(9):1364-1374. 10.1038/s41591-020-1034-x

[144] Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. Radiology. Artificial intelligence. 2024;6(4):e240300. 10.1148/ryai.240300

[145] Lekadir K, Frangi AF, Porras AR, Glocker B, Cintas C, Langlotz CP, et al. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ (Clinical research ed.). 2025;388:e081554. 10.1136/bmj-2024-081554

[146] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, et al. Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019;:220-229. 10.1145/3287560.3287596

[147] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, et al. Datasheets for datasets. Communications of the ACM. 2021;64(12):86-92. 10.1145/3458723

[148] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016;3:160018. 10.1038/sdata.2016.18

[149] Andrew J. Vickers, Elena B. Elkin. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. Medical Decision Making. 2006;26(6):565-574. 10.1177/0272989x06295361

[150] Andrew J. Vickers, Ben van Calster, Ewout W. Steyerberg. A simple, step-by-step guide to interpreting decision curve analysis. Diagnostic and Prognostic Research. 2019;3(1):18. 10.1186/s41512-019-0064-7