Performance metric curve analysis framework to assess impact of the decision variable threshold, disease prevalence, and dataset variability in two-class classification

Heather M. Whitney[a,b,*] **Karen Drukker**[a] **and Maryellen L. Giger**[a]
[a]University of Chicago, Department of Radiology, Chicago, Illinois, United States
[b]Wheaton College, Department of Physics, Wheaton, Illinois, United States

Abstract

Purpose: The aim of this study is to (1) demonstrate a graphical method and interpretation framework to extend performance evaluation beyond receiver operating characteristic curve analysis and (2) assess the impact of disease prevalence and variability in training and testing sets, particularly when a specific operating point is used.

Approach: The proposed performance metric curves (PMCs) simultaneously assess sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), and the 95% confidence intervals thereof, as a function of the threshold for the decision variable. We investigated the utility of PMCs using six example operating points associated with commonly used methods to select operating points (including the Youden index and maximum mutual information). As an example, we applied PMCs to the task of distinguishing between malignant and benign breast lesions using human-engineered radiomic features extracted from dynamic contrast-enhanced magnetic resonance images. The dataset had 1885 lesions, with the images acquired in 2015 and 2016 serving as the training set (1450 lesions) and those acquired in 2017 as the test set (435 lesions). Our study used this dataset in two ways: (1) the clinical dataset itself and (2) simulated datasets with features based on the clinical set but with five different disease prevalences. The median and 95% CI of the number of type I (false positive) and type II (false negative) errors were determined for each operating point of interest.

Results: PMCs from both the clinical and simulated datasets demonstrated that PMCs could support interpretation of the impact of decision threshold choice on type I and type II errors of classification, particularly relevant to prevalence.

Conclusion: PMCs allow simultaneous evaluation of the four performance metrics of sensitivity, specificity, PPV, and NPV as a function of the decision threshold. This may create a better understanding of two-class classifier performance in machine learning.

 [DOI: 10.1117/1.JMI.9.3.035502]

Keywords: artificial intelligence; AUC; machine learning; performance assessment; radiomics; repeatability.

Paper 21326GR received Dec. 13, 2021; accepted for publication May 11, 2022; published online May 31, 2022.

1 Introduction

Receiver operating characteristic (ROC) analysis is a widely used and accepted method for performance assessment of two-class, i.e., binary, classification tasks. It investigates true positive fraction (i.e., sensitivity or recall) and false positive (FP) fraction (1-specificity) pairs over the range of the decision variable.[1] The area under the ROC curve (AUC) is commonly used as a summary metric of performance for binary classification tasks including those addressed by many artificial intelligence models in computer aided diagnosis (CADx) in medical imaging.[2] The AUC summarizes in a single value (with an uncertainty such as a 95% confidence interval,

*Address all correspondence to Heather M. Whitney, hwhitney@uchicago.edu

CI) the classification performance in the context of the entire range of thresholds for the decision variable. It has been reported, however, that ROC analysis alone may not be sufficient to thoroughly assess performance.[3] In addition, in the case of imbalanced datasets, i.e., datasets with very low or very high disease prevalence, methods such as precision–recall analysis[4] may be preferred. At the same time, a disadvantage of precision–recall analysis is that the precision (also known as the positive predictive value, PPV) and the area under the precision–recall curve depend on the disease prevalence in the dataset. In addition, some authors have noted that while PPV and the negative predictive value (NPV) can be investigated in a framework analogous to ROC analysis, these curves and prevalence-based metrics are not as easily interpreted.[5]

Moreover, when considering clinical use of CADx methods, the selection of an operating point (or operating points) is important.[6] The selected threshold value for the decision variable trades off type I (FP) and type II (false negative, FN) errors. Reducing the number of FP and FN decisions is of clinical importance in medical imaging and in other fields.[7–9] In ROC analysis, several methods have been developed for the identification of a single "optimal" decision threshold for this purpose.[10,11] Most are constructed in the context of a "preferred" sensitivity or specificity (such as 95% for either), a combination of these,[12] or their extension into cost basis or utility.[2,13,14] The "preferred" target values for sensitivity or specificity depend on disease prevalence, the clinical task, and the tolerance for (or cost of) type I and type II errors. However, the methods to optimize the decision threshold in ROC-like analysis using prevalence are not yet well developed.[5,15–18]

In this work, we therefore propose a framework of performance metric curves (PMCs) to simultaneously view sensitivity, specificity, PPV, and NPV as a function of the decision variable threshold. We investigated six different commonly used choices of operating points in the context of disease prevalence and variability in training and test sets. While our framework is applicable to performance evaluation of two-class classification tasks in general, we illustrate its use by applying it to the evaluation of classification performance for the distinction between malignant and benign breast lesions imaged with dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI), using both a clinical dataset and simulation studies.

2 Methods

2.1 *Datasets*

The clinical dataset consisted of images of 1885 unique breast lesions imaged at 3.0 T using DCE-MRI protocols. The details of the clinical dataset, including image acquisition protocol and clinical characterization, have been previously published.[19–23] Images were separated into two subsets by year of acquisition; one subset of lesions imaged in the period of 2015–2016 served as the training set and the other subset of lesions imaged in the year 2017 served as the independent test set (Table 1). This longitudinal separation in a training and test cohort mimics how CADx would be deployed in the clinic after development.

The lesions had been automatically segmented after the indication of the lesion center[24] by an experienced breast imaging technologist and radiomic features had been extracted using a

Table 1 Description of the dataset: number of lesions (and percent of total) by lesion type.

Type of lesion	Training set: 2015 to 2016 number of lesions (percent of total)	Test set: 2017 number of lesions (percent of total)
Benign	370 (26%)	111 (26%)
Malignant	1080 (74%)	324 (74%)
Total	1450	435

Table 2 Radiomic features used in the study. A complete description of the features can be found in prior publications as noted.

Feature category	Feature name
Morphology[25]	Irregularity
	Surface area-to-volume ratio
	Margin sharpness
Texture[26]	Energy
	IMC1
	Sum average
Kinetic curve enhancement[27]	Maximum enhancement
	Time to peak
	Washout rate
	Volume of most enhancing voxels

dedicated in-house workstation. In this work, we used 10 features for each imaged lesion, describing the shape and morphology,[25] texture,[26] and kinetic curve enhancement[27] of the lesions (Table 2). Our previous work identified these features as useful for distinguishing between malignant and benign breast lesions in a study using this dataset.[19]

2.2 *Clinical Dataset Study*

For the analysis of the clinical dataset (Table 1), cases were sampled to generate 1000 training and corresponding test folds. To simulate variability in both the training and test sets, the classifier was trained and tested in a bootstrap-like fashion in which the training folds and test folds were obtained from the training and test set, respectively, by stratified sampling with replacement (1000 folds each with the desired disease prevalence maintained). Each feature was standardized to zero mean unit variance within each training fold and used to train a linear discriminant analysis (LDA) classifier. The features in the test folds were standardized using the mean and variance obtained from the corresponding training fold and formed the input to the trained classifier to obtain the posterior probability of malignancy (range 0 to 1) for lesions within the test folds. The posterior probability of malignancy served as the decision variable in the PMC example application.

2.3 *Simulated Datasets Study*

To further understand and demonstrate the utility of PMCs in the performance evaluation of CADx, simulation studies were conducted. These simulation studies were based on the clinical dataset described above. To obtain simulated feature sets, the values of each of the 10 features were tabulated for the benign and the malignant lesions in the training set and in the test set separately (Table 1) and subsequently fit to a normal distribution (resulting in a total of 40 distributions, a set of four for each feature). From these fitted distributions, random samples were drawn to obtain training and test sets with five different matching disease prevalences: 0.1, 0.25, 0.5, 0.75, and 0.9. (Note that the prevalence of 0.75 was by design very similar to the prevalence of the clinical data of 0.74.) The bootstrap-like sampling was performed in the same way as for the clinical datasets albeit now sampling from the simulated (fitted) training and test sets (2000 samples each for training and testing, 1000 iterations). Note that this is bootstrapping with replacement, but the training and test sets are taken from the separate training and test sets.

Table 3 Example operating points investigated in the study. A complete description of the operating points can be found in the references as noted when applicable.

Operating point
(A) Maximum sensitivity at minimum FP rate[30]
(B) A "preferred" target sensitivity of 0.95
(C) A "preferred" target specificity of 0.95
(D) An "optimal" sensitivity and specificity that maximized sensitivity + specificity − 1 (based upon the Youden index)[12,31]
(E) An "optimal" sensitivity and specificity point that minimized the distance of the ROC curve from perfect sensitivity and specificity (i.e., the Euclidean distance)[12,32]
(F) The maximum of mutual information *I*, the amount of information expected to be gained by a diagnostic test [5,18,33]

2.4 *Statistical Analysis*

The overall performance in the task of classifying (actual or simulated) lesions as malignant or benign was evaluated using ROC analysis and the nonparametric Wilcoxon method to estimate the AUC for each test fold. The median and 95% CI for AUC were derived empirically from all test folds. Note that no .632+ bootstrap correction of AUC[28,29] was needed in this work since the training and test folds were each sampled from completely separate training and test sets.

To demonstrate the utility of our proposed PMC framework, the performance metrics of sensitivity, specificity, PPV, and NPV were obtained at thresholds over the range of the decision variable at intervals of 0.01. (Note that intervals smaller than 0.01 did not provide any numerical advantage or change figure interpretations.) Six example desired operating points on the ROC curve were investigated (Table 3), and all thresholds for the decision variable corresponding to the desired operating points were determined using each training fold and then applied to the corresponding test fold. That is, this training set-determined decision threshold was applied to the cases in the corresponding test fold. The example operating points were not selected to be exhaustive but rather a selection of operating points used in the literature. In addition, the "best" choice of operating point(s) from a clinical perspective depends on many factors including the clinical task, population, and risk/benefits, and we did not investigate the superiority of any of these operating points in this classification task.

To construct the PMCs: (1) the values for the performance metrics were aggregated for the training and test folds, and median values and 95% CIs were empirically calculated and (2) the median values and 95% CI of the threshold of the decision variable corresponding to each operating point type were determined (Table 4). By combining these performance metrics in a single

Table 4 Characteristics of performance metric curves.

Characteristic	For the training set	For the test set
Performance metrics	Median sensitivity, specificity, PPV, and NPV	Median sensitivity, specificity, PPV, and NPV
Classifier threshold	Median classifier threshold for the desired operating point(s)	n/a
Error estimates	95% CI of the decision threshold(s) corresponding to the desired operating point(s)	95% CI of all performance metrics across the range of the decision variable

CI, confidence interval.

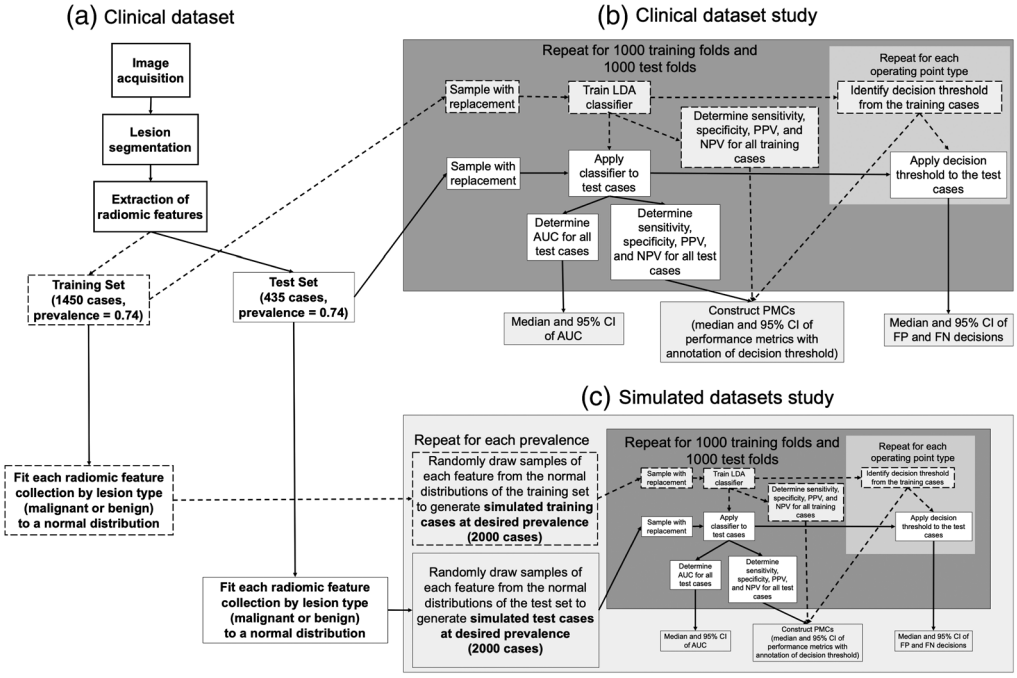

Fig. 1 Workflow depiction of the preparation of the datasets and the construction of the PMCs. (a) Separation of the unique cases into training and test sets by year of acquisition after image acquisition, lesion segmentation, and extraction of radiomic features, (b) the clinical dataset study, and (c) the simulated datasets study. Dotted lines are used to indicate dataset and PMCs preparation using the training set, contrasted with using the test set. LDA, linear discriminant analysis; PPV, positive predictive value; NPV, negative predictive value; AUC, area under the receiver operating characteristic curve.

figure, PMCs allow for simultaneous assessment of sensitivity, specificity, PPV, and NPV over the range of a decision variable, and of the impact of decision threshold value selection for both the training set and the test set.

In addition, to gain insight in the actual number of type I and type II errors, the median and 95% CI of the number of FP decisions and FN decisions were determined for each operating point of interest. A workflow figure describing the preparation of the datasets and the analysis is shown in Fig. 1.

3 Results

3.1 *Clinical Dataset Study*

3.1.1 *ROC results from the clinical dataset study*

The results in terms of ROC for the clinical test set ($N = 435$ cases, cancer prevalence 74%) demonstrate the difference of operating points for a single prevalence in terms of the uncertainty of sensitivity and specificity for each operating point, as measured by both the median and the 95% CI of these (Fig. 2).

3.1.2 *Performance metric curves from the clinical dataset study*

The results in terms of PMC for the clinical test set ($N = 435$ cases, cancer prevalence 74%) demonstrate the difference of operating points for a single prevalence in terms of the decision threshold for each operating point, as measured by both the median and the 95% CI of these (Fig. 3).

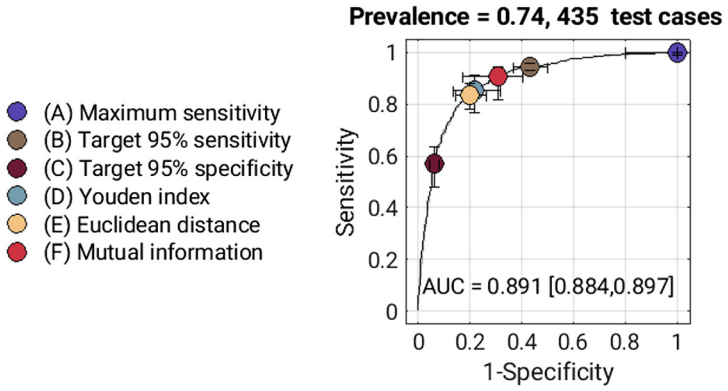

Fig. 2 Clinical dataset study ROC curves (solid line) with 95% CI (shading) and example operating points (see Table 3) with 95% CIs for the clinical dataset test folds (sampled from 435 clinical test cases) in the task of distinguishing between benign and malignant lesions. The area under the ROC curve (AUC), median and 95% CI, across the bootstrap folds is given as well. (Note that the shading (95% CI) along the ROC curve is very narrow, reflected in the narrow CI of the AUC).

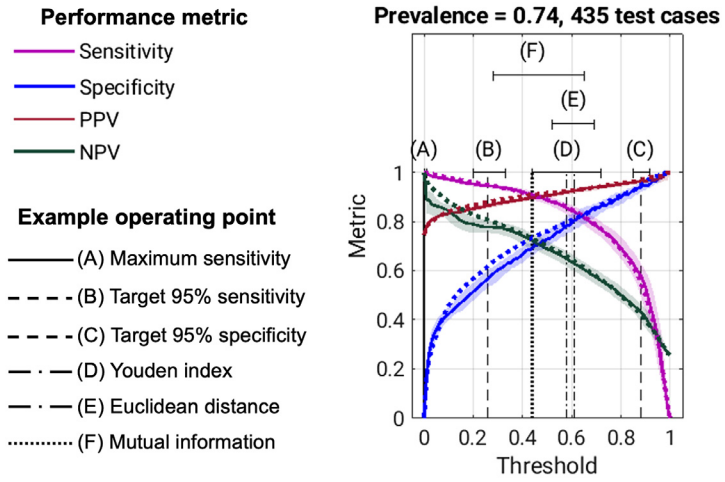

Fig. 3 Clinical dataset study (435 test cases) PMC showing sensitivity, specificity, PPV, and NPV as a function of decision threshold for the training set (median, dashed lines) and the test set cases (median, solid lines, and 95% CI, shading). Vertical lines indicate the median classifier threshold from the training set needed to attain the example operating points (Table 3), with the 95% CI of the threshold shown with error bars above the lines. The decision threshold is across the range of the classifier output, which is the posterior probability of malignancy.

3.1.3 *Performance metric curves interpretation on the clinical dataset study*

The characteristics of PMCs (Table 4) can be interpreted for the clinical dataset (Fig. 3) as follows:

Performance metrics (training set and test set). The median sensitivity, specificity, PPV, and NPV of the test set were found to be mostly similar to those of the training set across the decision variable. Exceptions were NPV over the range of ~0 to 0.3 and specificity over the range of ~0.1 to 0.6 for the decision threshold. There may be some degree of overfitting in these ranges, and operating points that utilize these metrics in these decision threshold ranges may be negatively impacted.

Classifier threshold (training set). The median decision threshold associated with the Youden index and Euclidean distance was similar for this clinical dataset (~0.6).

Error estimates (training set). The classifier threshold for the Youden index varied more than for the Euclidean distance. Using the Euclidean index may be preferable over the Youden index due to the reduced error from training set variability if an operating point of their type is desired. There was substantial variability in the classifier threshold associated with the mutual information operating point.

Error estimates (test set). The 95% CI of NPV, specificity, and sensitivity in the test set was substantial over the range of 0 to 0.3, 0.1 to 0.6, and 0.6 to 0.8, respectively. On the other hand, it was particularly small for sensitivity and PPV in the range of the target 95% sensitivity operating point.

3.1.4 *Extension of performance metric curves into clinical impact from the clinical dataset study*

As noted above, PMCs can be extended into clinical impact by identifying type I and type II errors, and their uncertainty, in terms of operating point selection. For this dataset, the trade-offs in sensitivity, specificity, PPV, and NPV across the decision threshold range can be viewed in terms of impact on resulting FP (type I errors) and FN (type II errors) decisions (Fig. 4).

This can specifically be seen for the different example operating points, (A)–(F) (Table 3), which demonstrate the impact on the median number of FP and FN decisions as well as their corresponding uncertainties (Fig. 5).

For example, Fig. 5 shows the trade-off between using the Youden index and the Euclidean distance as the operating point in the clinical dataset for type I errors (median and 95% CI: 48 [28, 77] and 54 [39, 61], respectively). The median number of FP decisions was similar between the two operating points, but the precision was shown to be higher when the Euclidean distance was used. In addition, Fig. 5 demonstrates that when the target 95% specificity operating point was used for this dataset, while there is a very low number of type I errors (7 [5, 9]), the uncertainty in the already high number of type II errors was substantial (140 [120, 169]). In contrast, the 95% CI of the sensitivity was low in the decision threshold region for the target 95% sensitivity classifier, which resulted in high precision in type I and type II errors for this operating point (48 [41, 55] and 18 [14, 22], respectively).

3.2 *Simulated Datasets Study*

3.2.1 *ROC results from the simulated datasets study*

As expected, the ROC curves and AUC values in the task of distinguishing between malignant and benign cases in the simulated test set showed little difference by prevalence (Fig. 6).

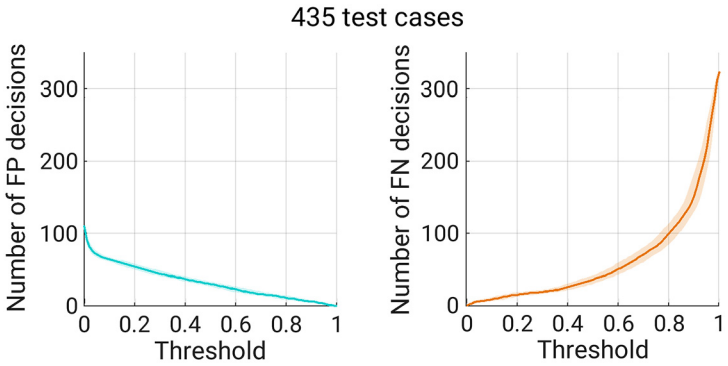

Fig. 4 Clinical dataset study (435 test cases) median number of (left) FP decisions (type I errors) and (right) FN decisions (type II errors) across the range of the decision threshold for each (lines) and 95% CI thereof (shading) for the test cases. The decision threshold is across the range of the classifier output, which is the posterior probability of malignancy.

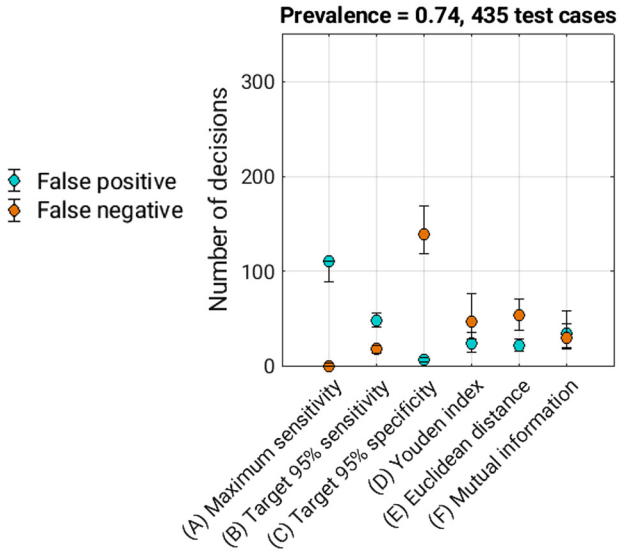

Fig. 5 Clinical dataset study (435 test cases) example operating points and error bars (95% CI) of the number of FP decisions (type I errors) and FN decisions (FN, type II errors).

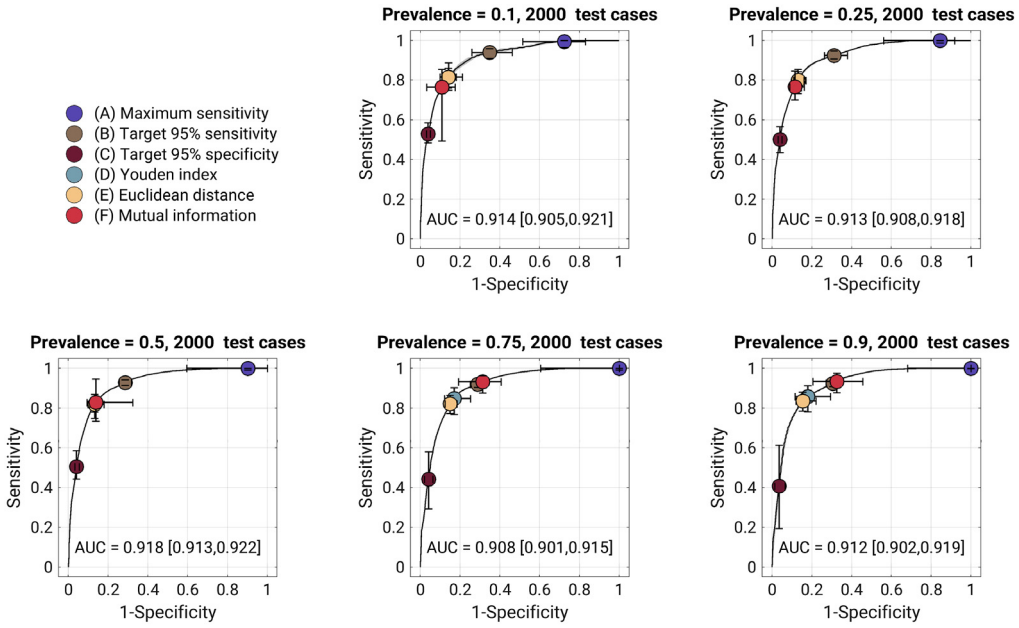

Fig. 6 Simulated datasets study ROC curves (solid lines) with 95% CI (shading) plus example operating points (see Table 3) with 95% CIs for the simulation test folds (sampled from 2000 simulated test cases) in the task of distinguishing between benign and malignant lesions. The area under the ROC curve (AUC), median and 95% CI, across the bootstrap folds is given as well. (Note that the shading (95% CI) along the ROC curve is very narrow, reflected in the narrow CI of the AUC). For prevalence = 0.1 and 0.25, the operating point for the Youden index is obscured by marker for the operating point for the Euclidean distance. For prevalence = 0.5, the operating points for the Youden index and the Euclidean distance are obscured by the marker for the operating point for mutual information.

For example, AUC (median, [95%CI]) = 0.914 [0.905, 0.921] for a prevalence = 0.1, while AUC = 0.912 [0.902, 0.919] for a prevalence = 0.9. Concurrently, the median (1 − specificity, sensitivity) of most example operating points demonstrated different levels of uncertainty among them (as indicated by the width of their 95% CIs), a result of the variability of the training set. The most substantial uncertainty across the simulated prevalences was observed

in sensitivity for the mutual information operating point at a prevalence = 0.9 (95% CI: 0.428).

3.2.2 *Performance metric curves from the simulated datasets study*

The associated proposed PMCs (Fig. 7) show that while sensitivity and specificity as a function of decision threshold were similar across prevalence, PPV and NPV depended on prevalence. These observations were as expected but their simultaneous presentation gives insight into the impact of decision variable threshold selection on type I and type II errors by prevalence (unnecessary biopsies and missed cancers, respectively, in our example). Moreover, the additional presentation of the decision variable threshold obtained from the training folds (median and 95% CI) for the selected operating points, (A)–(F), provides valuable insight into how reliable the CADx/AI output is for a given operating point, and how it may change with prevalence for some operating points (e.g., 95% specificity and mutual information) more than others. For example, the median decision variable threshold for 95% specificity was 0.79, 0.82, 0.85, 0.91, and 0.95 as prevalence increased, while for mutual information it was 0.61, 0.59, 0.54, 0.28, 0.25 as prevalence increased. In contrast, the median decision variable threshold for 95% sensitivity was fairly stable at 0.28, 0.3, 0.3, 0.32, and 0.28 as prevalence increased.

3.2.3 *Performance metric curves interpretation on the simulated datasets study*

The characteristics of PMCs (Table 4) can be interpreted for the simulated dataset (Fig. 7) as follows:

Performance metrics (training set and test set). For most prevalences, the median sensitivity, specificity, PPV, and NPV of the test set were similar to that of the training set, with

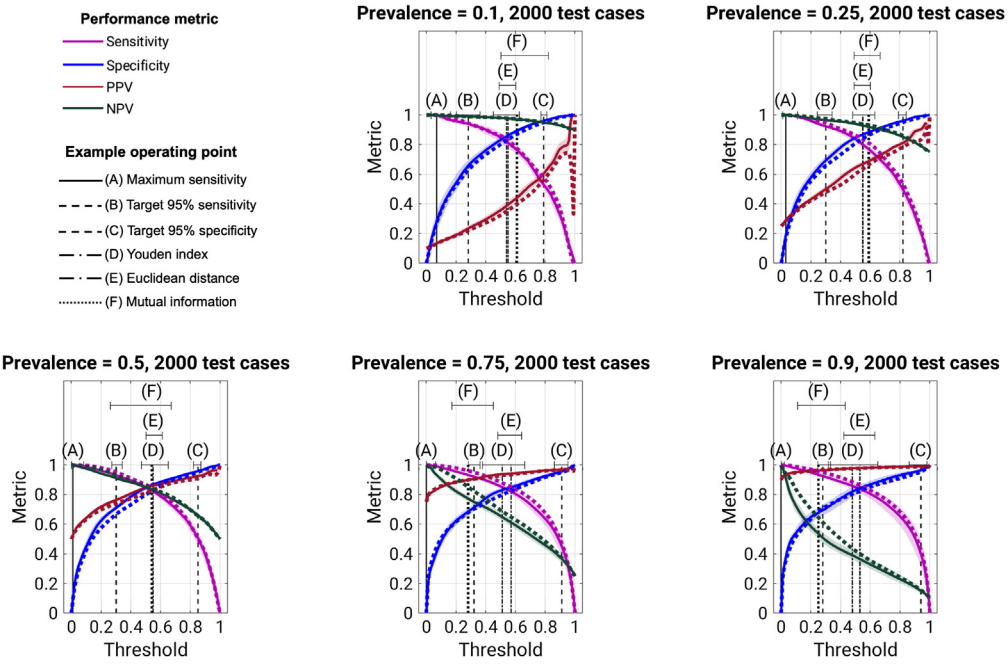

Fig. 7 Simulated dataset study performance metric threshold (PMC) curves of sensitivity, specificity, PPV, and NPV as a function of the decision threshold (obtained from the training folds) for the simulated training folds (dashed lines, median) and the simulated test folds (solid lines, median, with shading, 95% CI). The decision threshold is across the range of the classifier output, which is the posterior probability of malignancy. Vertical lines indicate the median classifier threshold from the simulated training folds needed to attain the example operating points (Table 3), with the 95% CI of the threshold shown with error bars above the lines.

the exception of NPV over the range of 0 to 0.5 for the decision threshold. This effect increased with prevalence. There may be some degree of overfitting in this range.

Classifier threshold (training set). The median decision threshold remained similar across prevalence for the operating points of maximum sensitivity, 95% sensitivity, the Youden index, and Euclidean distance. The median decision threshold associated with the mutual information operating point changed substantially with prevalence, while that of the target 95% specificity operating point did so to a lesser extent.

Error estimates (training set). The classifier threshold for the Euclidean distance varied more than for the Youden index, with increasing 95% CI as prevalence increases. Using the Youden index may be preferable over the Euclidean distance due to the reduced error from training set variability if an operating point of their type is desired. There was substantial variability in the classifier threshold associated with the mutual information operating point.

Error estimates (test set). The 95% CI of NPV and sensitivity in the test set was substantial over the range of 0 to 0.5 and 0.6 to 0.8, respectively. It improved for sensitivity as prevalence decreases.

3.2.4 *Extension of performance metric curves into clinical impact from the simulated datasets study*

As noted above, PMCs can be extended into clinical impact by identifying type I and type II errors and their uncertainty, in terms of operating point selection. The number of FP decisions (type I errors) and FN decisions (type II errors) was different by prevalence across the decision thresholds, as was the uncertainty (95% CI) thereof (Fig. 8). The uncertainty was largest for very low disease prevalence (FP decisions) or very high disease prevalence (FN decisions), as expected.

For a given example operating point, the absolute numbers of FP and FN decisions differed with prevalence in terms of both median values and 95% CI (Fig. 9). For example, the number of FP and FN decisions decreased and increased with prevalence, respectively, for most operating points, especially maximum sensitivity and target 95% specificity.

Notably, the 95% CI of type I errors for the maximum sensitivity operating point ranged from 572 FP decisions (out of 2000 test cases) at a prevalence = 0.1 to 64 FP decisions at a prevalence = 0.9, while the 95% CI of type II errors for the target 95% specificity operating point ranged from 20 FN decisions at a prevalence = 0.1 to 760 FN decisions at a prevalence = 0.9. However, these trends were not observed for the mutual information operating point.

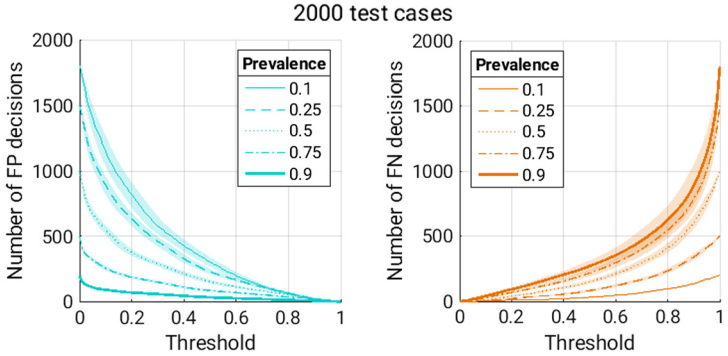

Fig. 8 Simulated datasets study median number of FP decisions (type I errors, left) and FN decisions (type II errors, right) across the range of the decision threshold (posterior probability of malignancy) for each prevalence investigated (lines) and 95% CI thereof (shading) for the 2000 simulated test cases.

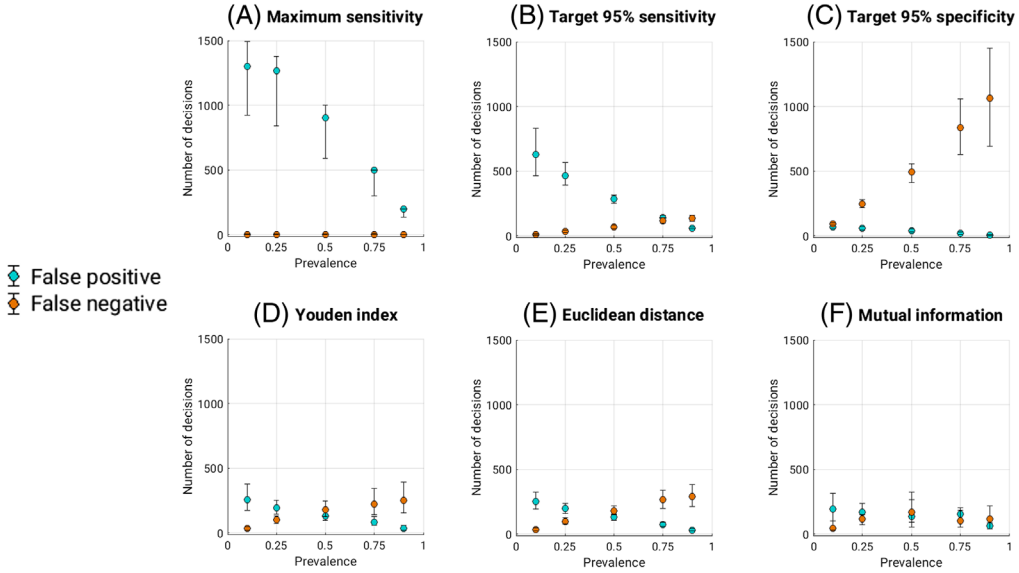

Fig. 9 Simulated datasets study example operating points and error bars (95% CI) of the number of FP decisions (type I errors) and FN decisions (type II errors) at each prevalence investigated.

4 Discussion

There is great interest in the use of graphical tools to support understanding of classification performance and association with decision thresholds. Some of these include likelihood ratios,[34,35] F-measure curves to convey information related to precision–recall,[36] a variety of methods to display information related to cost[37–39] and utility[14] (including indifference curves[40–42]), and combinations of these, such as that of prevalence and cost of misclassification into "prevalence value accuracy" (PVA) curves.[43] Most of these rely on display of the performance in terms of pairs of metrics. Some, like the PVA curve, which plots minimum cost of misclassification based upon the number of FPs and FNs as a function of values of prevalence and unit cost ratio, are three-dimensional. The authors who proposed PVA curves noted that the three-dimensional nature of the figures could inhibit their usability. The PMC graphical presentation, however, maintains two-dimensional data visualization while still aiding in identification of the number of type I and type II errors at example operating points and the associated uncertainty in those errors in the test set. The two-dimensional nature of PMCs has an advantage of close association with metrics given in traditional ROC curves, reducing difficulty in transferring the conceptual framework to a new visualization.

The proposed PMC method utilizes bootstrapping in both the training set and the test set so in the graphical presentation the curves of sensitivity, specificity, PPV, and NPV are displayed as their median and 95% CI from the test set and are easily viewed with respect to the median and 95% CI of the thresholds for operating points from the training set, similar to our presentation of sensitivity and specificity in this format in previous work.[44,45] Interpreting performance metrics and example operating points in the context of variability enhances the interpretation of them and supports their translation to a clinical decision-making process.

The decision threshold associated with the Youden index was consistently associated with the intersection of sensitivity and specificity. Greiner et al. proposed that the intersection of sensitivity and specificity as a function of decision threshold, called a "two-group ROC" curve,[46] can be used to graphically identify a useful decision threshold. Our proposed method expands upon this concept by incorporating additional information from prevalence, the PPV and NPV. Similar to these "two-group ROC" curves, if desired, PMCs can additionally provide a reference point where sensitivity is equal to PPV and where specificity is equal to NPV. Simple algebra demonstrates that at this crossing point of sensitivity and PPV, the number of FN decisions is equal to the number of FP decisions: FN = FP. The same relationship is true for the decision threshold where specificity is equal to the NPV, i.e., FN = FP. This decision threshold is simple and intuitive: for decision thresholds less than the crossing point, the number of type I errors will begin to

be greater than the number of type II errors. On the other hand, for decision thresholds greater than the crossing point, the number of type II errors will begin to be greater than the number of type II errors. While this decision threshold does not take into account different weights or costs for type I and type II errors, it could potentially serve as an additional intuitive, simple reference point to observe on PMCs.

The use of mutual information to determine a desired operating point is a method that can also incorporate prevalence information via PPV and NPV since it intrinsically depends upon prevalence, in contrast to the other operating points used in this study, which are independent of prevalence by definition. Hughes et al.[18] recently investigated this operating point outside of the context of CADx but additionally note that the concept was introduced by Metz et al. as the term "information capacity."[33] Their work suggests the utility of predictive ROC curves (PROC), which display PPV as a function of 1-NPV, as a graphical complement to traditional ROC curves. An index similar to the Youden index, called the PSEP index (which stands for "simple index of separation"[16]) by Altman and Royston, which measures an optimal position on the PROC curve has been previously proposed by separate investigators,[16,17] also outside of CADx. In the study presented here, implementing the maximum of mutual information as a criterion for decision threshold determination was associated with substantial uncertainty, as indicated by the 95% CI of the decision thresholds and the resulting numbers of cases that were classified as FN or FP. This may be due to the increased number of terms which are used for the measure. We were initially interested in combining the Youden index and PSEP as a possible, beneficial optimal decision threshold. However, preliminary results suggested that for our dataset, the decision thresholds associated with this were binormally distributed, at times corresponding to the threshold from the Youden index and at others to the threshold from the PSEP index, which reduced the usefulness of the approach. This result was not entirely unexpected because the optimal thresholds for the Youden index and for the PSEP index will most likely be different.[18] Future studies will investigate other potential methods for combining information from the Youden index and the PSEP index.

There were some limitations to our study. For example, we demonstrated PMCs on both clinical and simulated data from a classification task related to the use of radiomic features extracted from MR images of breast lesions. While the utility of PMCs is expected to easily translate to other applications, the application of this method to radiomic features extracted from other modalities, such as full-field digital mammography, or to transfer learning from deep learning networks is a topic of future work. Second, we did not make quantitative comparisons of the impacts of training set variability, operating point selection, and prevalence for this classification task and this clinical dataset. Doing so will likewise be of interest in future work as we apply PMCs to other applications and classification tasks. Third, in our study, the disease prevalence in the training and test sets was the same. Prevalence scaling may be necessary in studies with training and test sets of different prevalence.[47] Fourth, the trends of the performance metrics in the training and test sets were very similar, due to the homogeneous nature of the clinical dataset. This similarity in trends will not be seen in more heterogeneous datasets, but this emphasizes the benefits of using PMCs in assessing classification. Finally, PMCs are information-dense and this could potentially reduce their utility somewhat. It could be useful to create a graphical user interface which displays example operating points one at a time and overlays their median and CI on the PMC figure itself.

5 Conclusion

We have presented a graphical method and interpretation framework to extend performance evaluation in ROC analysis beyond AUC. PMC analysis

(1) provides a framework for at-a-glance evaluation of the certainty of the decisions by CADx/AI by providing simultaneously the performance metrics of sensitivity, specificity, PPV, and NPV in an easy-to-grasp visual format, along with their error estimates;
(2) allows for a direct comparison of the performance attained for the training set and that for the test set giving immediate feedback on the potential presence of overtraining (overfitting);

(3) gives insight in the range of the decision variable itself (such as the posterior probability of malignancy in our example application) associated with a desired operating point, including its dependence on disease prevalence and dataset variability;

(4) can be extended further into clinical impact by identifying the number of FP and FN decisions and the uncertainty of these numbers in terms of operating point selection.

Thus, PMC figures and related analysis may support more comprehensive consideration of machine learning model design, enhancing traditional two-class classification performance evaluation methods.

Disclosures

M. L. Giger is a stockholder in R2 Technology/Hologic and was a cofounder and equity holder in Quantitative Insights (now Qlarity Imaging). M. L. Giger receives royalties from Hologic, GE Medical Systems, MEDIAN Technologies, Riverain Medical, Mitsubishi, and Toshiba. K.D. receives royalties from Hologic. It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interest that would reasonably appear to be directly and significantly affected by the research activities.

Acknowledgments

The authors express their sincere gratitude to Yu Ji and Peifang Liu (Department of Breast Imaging, Tianjin Medical University Cancer Institute and Hospital, National Clinical Research Center for Cancer, Tianjin Medical University, Tianjin, China), and Hui Li, Alexandra Edwards, and John Papaioannou (Department of Radiology, The University of Chicago) for the collection and preparation of the clinical dataset for radiomic analysis, as described in previous publications. This work was supported in part by the National Institutes of Health National Cancer Institute (NIH NCI) under Grant No. U01 CA195564 and Grant No. R15 CA227948, the NIH under Grant No. S10 OD025081, and NIBIB COVID-19 Contract 75N92020D00021.

References

1. D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics*, pp. xi, 455, John Wiley, Oxford (1966).
2. C. E. Metz, "Basic principles of ROC analysis," *Semin. Nucl. Med.* **8**(4), 283–298 (1978).
3. D. Berrar and P. Flach, "Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them)," *Brief. Bioinf.* **13**(1), 83–97 (2012).
4. T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS One* **10**(3), e0118432 (2015).
5. G. Hughes, "On the binormal predictive receiver operating characteristic curve for the joint assessment of positive and negative predictive values," *Entropy* **22**(6), 593 (2020).
6. C. A. Gatsonis, "Receiver operating characteristic analysis for the evaluation of diagnosis and prediction," *Radiology* **253**(3), 593–596 (2009).
7. K. G. M. Moons et al., "Application of treatment thresholds to diagnostic-test evaluation: an alternative to the comparison of areas under receiver operating characteristic curves," *Med. Decis. Making* **17**(4), 447–454 (1997).
8. K. G. M. Moons et al., "Quantifying the added value of a diagnostic test or marker," *Clin. Chem.* **58**(10), 1408–1417 (2012).
9. D. Curran-Everett, "CORP: minimizing the chances of false positives and false negatives," *J. Appl. Physiol.* **122**(1), 91–95 (2017).
10. J. Jund et al., "Methods to estimate the optimal threshold for normally or log-normally distributed biological tests," *Med. Decis. Making* **25**(4), 406–415 (2005).

11. N. A. Obuchowski and J. A. Bullen, "Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine," *Phys. Med. Biol.* **63**(7), 07TR01 (2018).
12. X. H. Zhou, N. A. Obuchowski, and D. K. McClish, "Comparing the accuracy of two diagnostic tests," in *Statistical Methods in Diagnostic Medicine*, 2nd ed., pp. 165–194, John Wiley & Sons, Ltd (2011).
13. E. J. Halpern et al., "Comparison of receiver operating characteristic curves on the basis of optimal operating points," *Acad. Radiol.* **3**(3), 245–253 (1996).
14. Y. Wu et al., "Developing a utility decision framework to evaluate predictive models in breast cancer risk estimation," *J. Med. Imaging* **2**(4), 041005 (2015).
15. J. Vermont et al., "Strategies for graphical threshold determination," *Comput. Methods Prog. Biomed.* **35**(2), 141–150 (1991).
16. D. G. Altman and P. Royston, "What do we mean by validating a prognostic model?" *Stat. Med.* **19**(4), 453–473 (2000).
17. S.-Y. Shiu and C. Gatsonis, "The predictive receiver operating characteristic curve for the joint assessment of the positive and negative predictive values," *Philos. Trans. R. Soc. A* **366**(1874), 2313–2333 (2008).
18. G. Hughes, J. Kopetzky, and N. McRoberts, "Mutual information as a performance measure for binary predictors characterized by both ROC curve and PROC curve analysis," *Entropy* **22**(9), 938 (2020).
19. Y. Ji et al., "Independent validation of machine learning in diagnosing breast cancer on magnetic resonance imaging within a single institution," *Cancer Imaging* **19**(1), 64 (2019).
20. H. M. Whitney et al., "Harmonization of radiomic features of breast lesions across international DCE-MRI datasets," *J. Med. Imaging* **7**(1), 012707 (2020).
21. H. M. Whitney et al., "Comparison of breast MRI tumor classification using human-engineered radiomics, transfer learning from deep convolutional neural networks, and fusion methods," *Proc. IEEE* **108**(1), 163–177 (2020).
22. H. M. Whitney et al., "Multi-stage harmonization for robust AI across breast MR databases," *Cancers* **13**(19), 4809 (2021).
23. Q. Hu et al., "Improved classification of benign and malignant breast lesions using deep feature maximum intensity projection MRI in breast cancer diagnosis using dynamic contrast-enhanced MRI," *Radiol.: Artif. Intell.* **3**(3), e200159 (2021).
24. W. Chen, M. L. Giger, and U. Bick, "A fuzzy C-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images," *Acad. Radiol.* **13**(1), 63–72 (2006).
25. K. G. A. Gilhuijs, M. L. Giger, and U. Bick, "Computerized analysis of breast lesions in three dimensions using dynamic magnetic-resonance imaging," *Med. Phys.* **25**(9), 1647–1654 (1998).
26. W. Chen et al., "Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images," *Magn. Reson. Med.* **58**(3), 562–571 (2007).
27. W. Chen et al., "Automatic identification and classification of characteristic kinetic curves of breast lesions on DCE-MRI," *Med. Phys.* **33**(8), 2878–2887 (2006).
28. B. Efron and R. Tibshirani, "Improvements on cross-validation: the .632+ bootstrap method," *J. Am. Stat. Assoc.* **92**(438), 548–560 (1997).
29. B. Sahiner, H.-P. Chan, and L. Hadjiiski, "Classifier performance prediction for computer-aided diagnosis using a limited dataset," *Med. Phys.* **35**(4), 1559–1570 (2008).
30. K. Drukker et al., "Combined benefit of quantitative three-compartment breast image analysis and mammography radiomics in the classification of breast masses in a clinical data set," *Radiology* **290**(3), 621–628 (2019).
31. E. F. Schisterman et al., "Optimal cut-point and its corresponding Youden index to discriminate individuals using pooled blood samples," *Epidemiology* **16**(1), 73–81 (2005).
32. M. Coffin and S. Sukhatme, "Receiver operating characteristic studies and measurement errors," *Biometrics* **53**(3), 823 (1997).
33. C. E. Metz, D. J. Goodenough, and K. Rossmann, "Evaluation of receiver operating characteristic curve data in terms of information theory, with applications in radiography," *Radiology* **109**(2), 297–303 (1973).

34. A. I. Bandos, H. E. Rockette, and D. Gur, "Use of likelihood ratios for comparisons of binary diagnostic tests: underlying ROC curves," *Med. Phys.* **37**(11), 5821–5830 (2010).
35. B. J. Biggerstaff, "Comparing diagnostic tests: a simple graphic using likelihood ratios," *Stat. Med.* **19**(5), 649–663 (2000).
36. R. Soleymani, E. Granger, and G. Fumera, "F-measure curves: a tool to visualize classifier performance under imbalance," *Pattern Recognit.* **100**, 107146 (2020).
37. C. Drummond and R. C. Holte, "Cost curves: an improved method for visualizing classifier performance," *Mach. Learn.* **65**(1), 95–130 (2006).
38. J. Hernández-Orallo, P. Flach, and C. Ferri, "Brier curves: a new cost-based visualisation of classifier performance," in *Proc. 28th Int. Conf. Mach. Learn.*, Bellevue, Washington (2011).
39. J. Hernández-Orallo, P. Flach, and C. Ferri, "ROC curves in cost space," *Mach. Learn.* **93**(1), 71–91 (2013).
40. F. Y. Edgeworth, *Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences*, C. Kegan Paul & Co., London (1881).
41. C. K. Abbey, M. P. Eckstein, and J. M. Boone, "An equivalent relative utility metric for evaluating screening mammography," *Med. Decis. Making* **30**(1), 113–122 (2010).
42. R. J. Irwin and T. C. Irwin, "A principled approach to setting optimal diagnostic thresholds: where ROC and indifference curves meet," *Eur. J. Internal Med.* **22**(3), 230–234 (2011).
43. A. T. Remaley et al., "Prevalence-value-accuracy plots: a new method for comparing diagnostic tests based on misclassification costs," *Clin. Chem.* **45**(7), 934–941 (1999).
44. K. Drukker, L. Pesce, and M. Giger, "Repeatability in computer-aided diagnosis: application to breast cancer diagnosis on sonography," *Med. Phys.* **37**, 2659–2669 (2010).
45. A. Van Dusen et al., "Repeatability profiles towards consistent sensitivity and specificity levels for machine learning on breast DCE-MRI," *Proc. SPIE* **11316**, 113160I (2020).
46. M. Greiner, D. Sohr, and P. Göbel, "A modified ROC analysis for the selection of cut-off values and the definition of intermediate results of serodiagnostic tests," *J. Immunol. Methods* **185**(1), 123–132 (1995).
47. K. Horsch, M. L. Giger, and C. E. Metz, "Prevalence scaling: applications to an intelligent workstation for the diagnosis of breast cancer," *Acad. Radiol.* **15**(11), 1446–1457 (2008).

Heather M. Whitney is an associate professor of physics at Wheaton College. Her experience in quantitative medical imaging has ranged from polymer gel dosimetry to radiation damping in nuclear magnetic resonance to radiomics. She is interested in investigating the effects of the physical basis of imaging on radiomics, the repeatability and robustness of radiomics, the development of methods for task-based distribution, and bias and diversity of medical imaging datasets. She is a member of SPIE.

Karen Drukker is a research associate professor at the University of Chicago, where she has been involved in medical imaging research for 20+ years. She received her PhD in physics from the University of Amsterdam. Her research interests include machine learning applications in the detection, diagnosis, and prognosis of breast cancer and, more recently, of COVID-19 patients, focusing on rigorous training/testing protocols, generalizability, and performance evaluation of machine learning algorithms. She is a member of SPIE.

Maryellen L. Giger is the A. N. Pritzker Distinguished Service Professor of Radiology, Committee on Medical Physics, and the College at the University of Chicago. She conducts AI research including computer- aided diagnosis, quantitative image analysis (radiomics), and deep learning in the areas of breast cancer, thyroid cancer, thoracic diseases, and COVID-19. She is a member of NAE, a fellow of SPIE, AAPM, AIMBE, IEEE, was the 2018 SPIE President, and is contact PI of MIDRC.