Confusion in the Matrix:

going beyond the ROC curve

Stephen M Borstelmann MD & Saurabh Jha MD

SMB: www.ai-imaging.org & www.n2value.com

SJ: Department of Radiology, University of Pennsylvania

ABSTRACT:

Artificial intelligence algorithms are being created on both investigationally and commercially. Evaluation of their performance is important for developers, investigators, clinical physicians, and regulatory agencies. No clear consensus exists on what metrics are best for algorithmic evaluation for AI and ML applications in radiology. We review the basics of the confusion matrix, continue to single number summary values such as accuracy, F1 score, and Φ coefficient, and then discuss Receiver Operator Curves and their derivatives, Precision Recall Curves, and Cost Curves. Recommendations are made for potential future directions and what currently may be best practices in algorithmic evaluation metrics.

INTRODUCTION:

The increasing interest in Artificial Intelligence (**AI**) and Machine Learning (**ML**) algorithms for patient care is plainly apparent to those following developments in the academic and commercial space. Applications include risk stratification, prognosis evaluation, data mining of text reports, and of course imaging suitable for use in Diagnostic Radiology.

Prognostications by technology pundits like Vinod Khosla in 2012 that “Technology will replace 80% of what doctors do” were not considered credible at the time by most academic or clinical radiologists.[1] IBM Watson’s early announcement of a move into healthcare related fields with the purchase of Merge Healthcare in 2015 was noted.[2] In 2017, Arterys was 510k FDA-approved for its Cardio DL program[3] and shortly thereafter the CheXNet paper was published by Pranav Rajpurkar and Andrew Ng et. al. from the Stanford Group.[4] Vinod Khosla

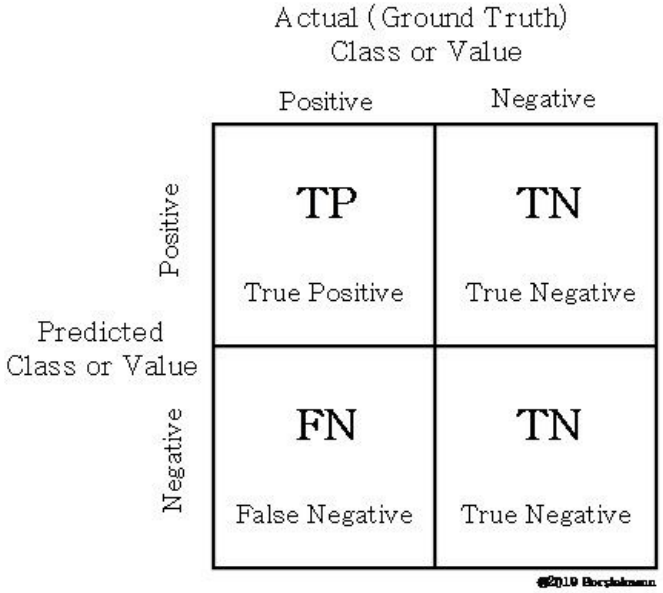
doubled down, pontificating that “Radiologists would be obsolete in five years.”[5] Suddenly, AI was at the forefront of many radiologist’s minds.

There were revisions of the CheXNet paper following discourse around the internet, with some authors focusing on the limitations of the Wang dataset[6] and others focusing on the reporting methodology[7]. As commercial interests starting moving more rapidly into the space, and investigators started releasing papers at conferences and in journals, there is a general confusion and no immediate consensus on how to properly evaluate these AI algorithms. Fortunately, the answer lies in one of the Radiologists’ fortés - diagnostic testing. A test for the presence of HIV antibodies, an abdominal CT scan to r/o ureteral stone, or an AI algorithm to detect pneumonia all share the same commonality - to positively identify the presence or absence of disease.

THE BASICS:

Diagnostic medical testing is a large portion of the average physician’s day and the radiologists’ lifeblood. Fundamentally, every test results in either a normal or abnormal result; a positive or negative. While every effort in medicine is made to try to minimize error, each test does have an associated inherent error rate – that is, sometimes the test will be **Falsely Positive** (**FP**), in the absence of abnormality, or **Falsely Negative** (**FN**) in the presence of abnormality. We term the accurate positives **True Positive** (**TP**) and the accurate negatives **True Negative** (**TN**). These values can be displayed as a two by two matrix, termed a confusion matrix or contingency table.

Figure 1 - Confusion Matrix



For physicians trying to diagnose disease it is helpful to know how good the test is in detecting abnormal results. After all, a test which doesn't catch most of the cases of what you are interested in is not much good at all, unless there is no other alternative. To gauge how good the test is, we can look at the ratio where an abnormality was detected and was real, compared to the same cases plus those that should have been detected but weren't (Type I Error).

In other words, we can calculate the **Sensitivity** of a test as : **TP/TP+FN**. Sensitivity is also called **recall**, or the **True Positive Rate** (**TPR**). And thus the **Specificity** of the test becomes : **TN/TN+FP**, allowing us to understand the fraction where the test was truly negative compared to the same plus cases which were detected, but ultimately weren't abnormal. Specificity is also termed **selectivity** or **True Negative Rate** (**TNR**).[8] Those involved in MQSA reporting in the past will be intimately familiar with these terms.

Clinicians struggle with, but need to know and understand these measurements, so that they can most accurately diagnose and treat patients. The average clinician looks for tests with high sensitivity and specificity to decrease false negative misses, and **Positive Predictive Value (PPV)**, calculated as: **TP/(TP+FP)**, also termed **precision** – usually without considering pre-test probability. This is because summary statistics are relatively complex.[9]

Sensitivity as a measure excludes TN and FP, and is biased toward screening, finding as many positives in a population as possible. Most clinicians follow a positive high sensitivity test with a test of high specificity. Specificity omits TP and FN, so if a high specificity test is positive, one can be reasonably certain of 'ruling in', but if it has been the only test performed, the test does not 'rule out'.

Sensitivity and Specificity alone may not be sufficient, so other measures have been proposed for use.

SINGLE VALUE SUMMARIES:

One of the most common measures used in ML is **Accuracy** (**ACC**).

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

Accuracy is relatively intuitive, measuring correctly predicted observations compared to all observations. However, accuracy can fail as a predictive measure because of a large difference between the size of FP and FN. This is also known as a **class imbalance** problem, and arises frequently in ML, but is largely outside the scope of this review.[10] Consider Breast Cancer Screening, in which the number of FP's (overcalls, up to 20%) will hopefully exceed FN's (missed cancers, 0.1%). A mammography model which fails to detect any cancers at all, TP or FP, could still result in a high accuracy.[11]

The **F1** Score has also been used.

$$F1 = \frac{2TP}{2TP + FP + FN}$$

It takes both false positives and false negatives into account, and is better than accuracy with an uneven distribution of members of the confusion matrix (also referred to as **classes**), as seen in the above example. F1 will also seek a balance between Sensitivity and Specificity.

Finally, the **Matthews correlation coefficient**, Φ coefficient, may be measured. It takes into account all positives and negatives, and can be used in cases of class imbalance. Its output is a scalar from -1 to 1, with 1 representing a perfect prediction, 0 no better than random prediction (null hypothesis) and -1 complete disagreement between observation and prediction.

$$\phi = \frac{TPxFN - FPxFN}{\sqrt{(TP + FP)(FT + FN)(TN + FP)(TN + FN)}}$$

Perhaps the greatest issue with single value metrics is that the same values can correspond to very different test or model performance.[8] The full range of statistics obtainable from the confusion matrix is displayed in figure 2.

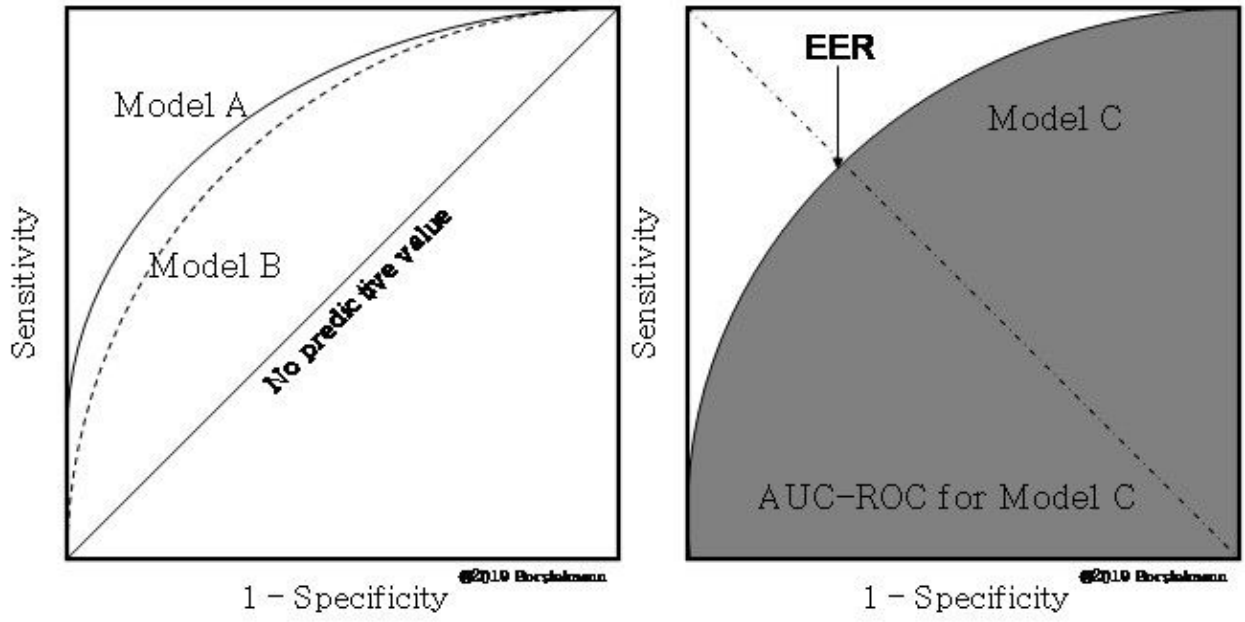
Figure 2 - Confusion Matrix Statistics

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{TPR}{FPR}$	Diagnostic odds ratio (DOR) = $\frac{LR+}{LR-}$	F_1 score = $2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{FNR}{TNR}$		

THE ROC CURVE, AUC-ROC, and derivatives

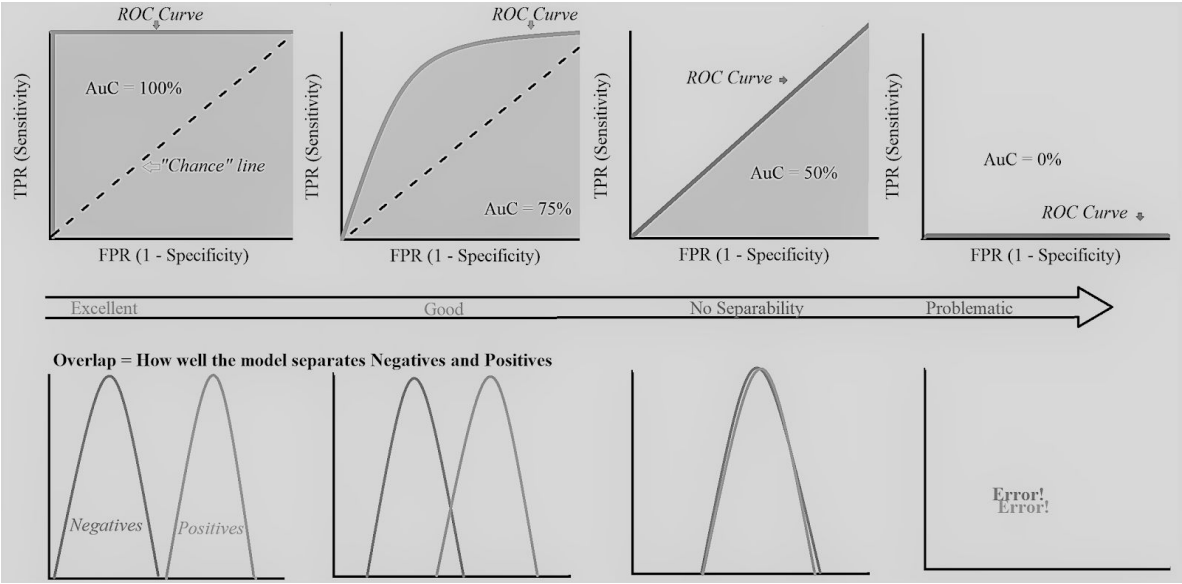
The **receiver operator curve** (**ROC)** has its humble origins in the Royal Air Force's early warning radar systems during World War II. The radar operator could pick up enemy aircraft, or be fooled by flocks of geese. As a plot of TP vs FP, expressed by plotting Sensitivity on the Y-axis, compared with 1-Specificity on the X-axis, the radar *receiver operator* could be evaluated on their ability to maximize enemy aircraft detection (TP) and minimize geese detection (FP). The plot provided a representation of sensitivity vs. specificity. 1-specificity is also known as the **fall-out** or **False Positive Rate** (**FPR)**.

figure 3a and 3b - ROC and AUC-ROC



Generally, a test (or model) which lies more to the upper left on the ROC curve (fig 3a model A) without crossing is better. The ROC curve is constructed by rank ordering test thresholds and the sensitivities and specificities for each threshold. The slope of the tangent line at a given threshold gives the **likelihood ratio** (**LR**) for that threshold. The Area Under the ROC Curve (**AUC-ROC** or simply **AUC**) measures the chance that a randomly selected TP will rank above a randomly selected TN, and thereby gives a graphical and numerical representation of the test's discriminative ability.

figure 4 – interpreting ROC curves[16]



Comparing ROC curves is a considered a quick way to a better test. But is it? If the curves cross, this indicates that one test is superior to the other in some circumstances, but inferior in others. In this manner, just because two different tests or models have equal AUC-ROC's, they may not be equally good for the same purpose.[13] Furthermore, In clinical practice, the radiologist chooses their single operating threshold level, whereas multiple thresholds exist in the ROC. AUC ROC will include performance over non-clinically relevant and possibly illogical thresholds.[14] Since ROC AUC treats both sensitivity and specificity equally, a test or model with a lower AUC ROC could clinically outperform a higher AUC ROC for purposes like screening. AUC also can suffer from a similar problem as accuracy - with few TP, the class imbalance issue can cause AUC to extrapolate inaccurately.[15]

The Error Equal Rate or Crossover Rate has also been suggested and is used more in biometrics. It is simply the ROC curve at the threshold where FP=FN; frequently at the intersection of the curve with a diagonal line inverse to the null hypothesis line on the ROC curve. As a scalar value, it is simplistic and does not take into account the actual operating threshold on the ROC curve. See figure 3b.

Different schema have been suggested for improving the ROC AUC, involving weighting. A weighted formula for CT colonography screening where benefit of early disease detection outweighs the theoretical cost of a missed cancer was proposed and termed the Net benefit function, where W is defined as the user assigned weight, and p the prevalence of abnormality in the defined population.[13] No quantitative method for establishing W has been proposed.

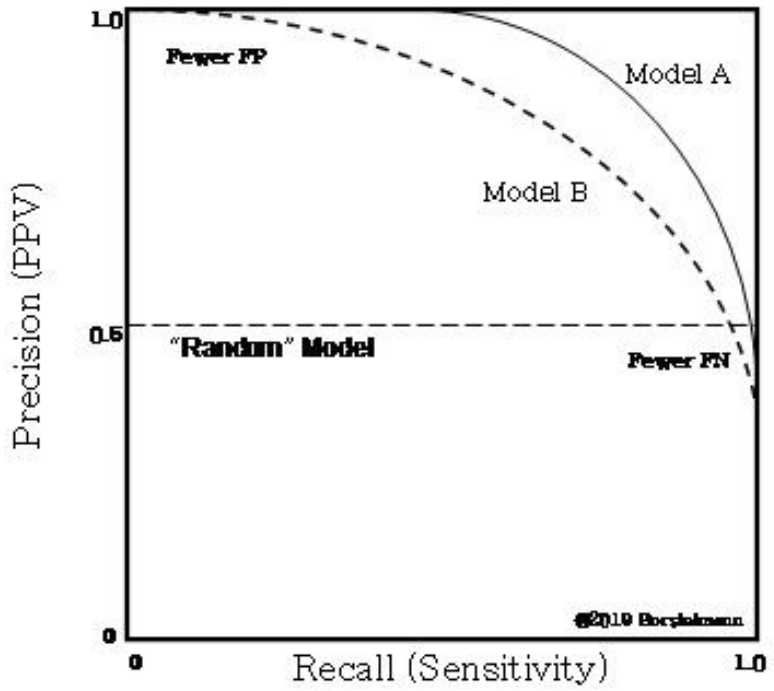
$$\text{NetBenefit} = \text{sensitivity} + \left(\text{specificity} * \left(\frac{1}{W}\right)\left(\frac{(1-p)}{p}\right)\right)$$

PRECISION-RECALL CURVE, AUC-PRC, and derivatives

Precision-Recall (**PRC)** curves are plots of sensitivity vs. PPV. One of the chief advantages of the PRC is it provides additional multi-threshold information that can be visually assessed. The closer to the upper right the curve moves, the better. The **AUC-PRC,** also termed **average precision,** can also be calculated through an integral and allows for a single value summary comparison between models or tests.[17]

A test or model with a strong ROC and AUC-ROC need not necessarily have a similarly strong PRC, and ROC optimization may not improve the PRC. However, a model or test with better ROC, AUC-ROC, PRC and PRC-ROC than another can confidently be evaluated as better. Unlike AUC, PRC is useful for imbalanced classes particularly when one is most concerned with the positive class.[18] Additionally, for high-value AUC-ROC models with a similar visual appearance, the PRC may allow more confident discrimination between the two on a visual basis. For this reason, some authors prefer it to ROC.[15]

Figure 5 - PRC curve



NEW MEASURES

The **cost curve** (**CC**) has been proposed as an improvement over the ROC curve, but has not received widespread use.[19] Perhaps this is because its calculation is more complex than a ROC, but more likely because the word 'cost' has so many meanings in the ML space, often used interchangeably with 'loss', and that the cost curve in Economics and Business research and related publications arise so frequently that meaning (and discoverability) are lost in the noise. Perhaps the cost curve could benefit from a rebranding to the **Drummond Cost Curve**?

Each point on the ROC space describes a line (format Y = Sx+b) defined by:

$$Y = (FN - FP) * p_{positive} + FP$$

Where $p_{positive}$ is the probability between 0 and 1 of a positive example in the sample, also expressible by (TP+FP)/(TP+FP+FN+TN) - really just the positive fraction.

A line in ROC space with slope S and y-intercept TP_0 then maps to CC space through the following equations, and an example of the conversion is shown in figure 6:

$$X = p_{positive} = \frac{1}{1+S}$$

$$Y = \text{error rate} = (1 - TP_0) * p_{positive}$$

Figure 6 - Cost Curves[19]

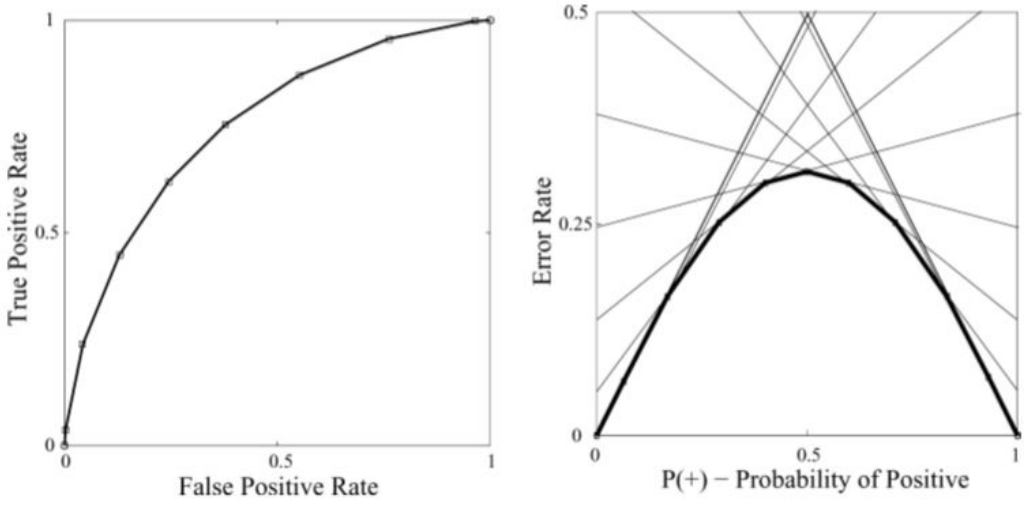


Fig. 4 (a) Ten ROC points and their ROC convex hull — (b) Corresponding set of cost lines and their lower envelope

Visually, a CC shows lines forming a lower envelope curve pulling down and away from the large apical triangle, where trivial cases of discrimination exist in the lower left and right corners of the Figure 4b plot. Better is down and away from the triangle's boundaries. Classifiers (ML models) can be compared in this manner, and potentially optimized for various different criteria. Unlike a ROC curve, which are independent of conditions, cost curves are designed for a specific performance measure. Misclassification costs, similar to that attempted in the Net Benefit weighted model above can be included. By using a bootstrap method[20] on the confusion matrix, confidence intervals can be created for the CC, and significance testing can also be performed. It is suggested that the CC method gives most of the benefits of ROC analysis, with extra benefits not available through ROC. One area where CC underperform is also in the setting of imbalanced data.[15] CC-AUC could provide a similarly comparable scalar, but more experience with that measure would be necessary.

FUTURE DIRECTIONS

Probably the biggest thing that would help solidify evaluation metrics in diagnostic imaging machine learning is a consistent effort for authors to publish multiple of these metrics so that we can review them across multiple algorithms and datasets, identifying those that have the most utility and are best in day-to-day use and evaluation. Too many investigators have insufficient experience with these metrics beyond sensitivity and specificity and perhaps the ROC curve.

As has been shown by multiple authors, class imbalances can have significant effects on loss of performance in classifier systems. To document this, investigators have proposed the use of a **class imbalance ratio** as a summary statistic on data. This may have utility in model evaluation, particularly on a formal basis by regulatory agencies, as such information would lend context to disclosures of Accuracy, F1, and ROC scores.

It should be noted that the confusion matrix reduces to a simple Positive-Negative, yes/no model. While currently most AI systems introduced use a similar binary classification, multiclass AI systems will require a more complex approach. This problem was approached in the 2000's and multiclass ROC is calculable; there is most likely still significant work to be done statistically in this area.

Further experience with the Cost Curve method would have to be performed to decide if this method was superior to the ones mentioned.

DISCUSSION

It is worth remembering the admonitions of Drummond and Holte : "a single, scalar performance measure cannot capture all aspects of the performance differences between two (classifiers)."[19] As physicians, we are used to specificity and sensitivity, but if we are to work

with AI and ML models, we must be cognizant of other model evaluation metrics. Single summary scores like ACC, MCC, and F1 are useful, but do not give the whole picture. At this time, only ROC and AUC-ROC measures are sufficiently diffused within the radiology literature and community to enjoy more widespread use. Experience is hard-earned and comes with both familiarity and frustration. With that said, the authors would like to suggest the following recommendations:

1. We admit that we don't know what we don't know. There is, collectively, insufficient experience with AI model evaluation and subsequent real-world followup assessment in clinical practice. Therefore, multiple measures for any AI model should be presented. At a minimum: Accuracy, F1, MCC and ROC.
2. Disclosure of the class imbalance ratio for any dataset or AI model should be strongly encouraged, particularly as we extend our reach into evaluation of multiple classes.
3. For tests that have imbalanced data or a bias toward screening, serious consideration should be given to use of PRC over ROC, and probably both should be routinely provided. A superior model will likely have both higher AUC-ROC and AUC-PRC.
4. Consideration to further development of the Drummond Cost Curve and new discriminative metrics should be given. Further basic applied statistical research would help here.
5. A perfect ML paper would include not only the confusion matrix, but the class imbalance ratio, ACC, F1, MCC, ROC, ROC-AUC, PRC and PRC-AUC statistics.
6. Considerations relating to external validation, spectrum bias, model drift, and edge cases are in no way minimized by the foregoing.

BIBLIOGRAPHY:

1. Khosla, V. (2012). Technology will Replace 80% of what doctors do. Retrieved from https://fortune.com/2012/12/04/technology-will-replace-80-of-what-doctors-do/
2. Forrest, C. (2015). IBM Watson bets 1 billion on healthcare with merge acquisition. Retrieved from TechRepublic website: https://www.techrepublic.com/article/ibm-watson-bets-1-billion-on-healthcare-with-merge-acquisition/
3. Crosti, N. (2017). Startup Arterys wins FDA clearance for AI-assisted cardiac imaging system. Retrieved from MedCity News website: https://medcitynews.com/2017/01/arterys-fda-ai/
4. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T.,Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, MP., Ng, AY. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. In *arXiv.org*. Retrieved from https://arxiv.org/abs/1711.05225

5. Farr, Christina. (2017). Some doctors will be “obsolete” in five years. Retrieved from https://www.cnbc.com/2017/04/07/vinod-khosla-radiologists-obsolete-five-years.html
6. Gichoya, J. & Borstelmann, S. (2018). Are computers better than doctors? What we learnt from the ChexNet paper for pneumonia diagnosis. Retrieved from https://n2value.com/blog/
7. Oakden-Rayner, L. (2018). CheXNet: an in-depth review.
8. Lever, J., Krzywinski, M., & Altman, N. (2016). Classification evaluation. *Nature Methods*. https://doi.org/10.1038/nmeth.3945
9. Reitsma, H., Rutjes, A., Whiting, P., Vlassov, V., Leeflang, M., & Deeks, J. (2009). Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy. Chapter 11. In *The Cochrane Collaboration*.
10. Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., & Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*. https://doi.org/10.1016/j.neunet.2007.12.031
11. Afonja, T. (n.d.). Accuracy Paradox. Retrieved from Towards Data Science website: https://towardsdatascience.com/accuracy-paradox-897a69e2dd9b
12. Wikipedia editors. Retrieved from: https://en.wikipedia.org/wiki/Confusion_matrix
13. Halligan, S., Altman, D. G., & Mallett, S. (2015). Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach. *European Radiology*. https://doi.org/10.1007/s00330-014-3487-0
14. López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*. https://doi.org/10.1016/j.ins.2013.07.007
15. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0118432
16. Glen, S. (2019) ROC Curve explained in one picture. *Datasciencecentral* From : https://www.datasciencecentral.com/profiles/blogs/roc-curve-explained-in-one-picture
17. McCann, S. (n.d.). Average Precision. Retrieved from https://sanchom.wordpress.com/tag/average-precision/
18. Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *AC International Conference Proceeding Series*. https://doi.org/10.1145/1143844.1143874
19. Drummond, C., & Holte, R. C. (2006). Cost curves: An improved method for visualizing classifier performance. *Machine Learning*. https://doi.org/10.1007/s10994-006-8199-5
20. Borstelmann S. (2019) Machine Learning Principles for Radiology Investigators. *Academic Radiology*. In Press.