Confusion in the Matrix: going beyond the ROC curve and down the rabbit hole for AI performance metrics

Stephen M Borstelmann MD & Saurabh Jha MD

Ai-imaging.org & N2value.com, University of Pennsylvania

Correspondence to: socmed@n2value.com

ABSTRACT:

Artificial intelligence algorithms are being created 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 the performance of 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**). Each of these cases, TP, FP, TN, FN can be considered a **class** of results. These values can be displayed as a two by two matrix, termed a confusion matrix or contingency table.

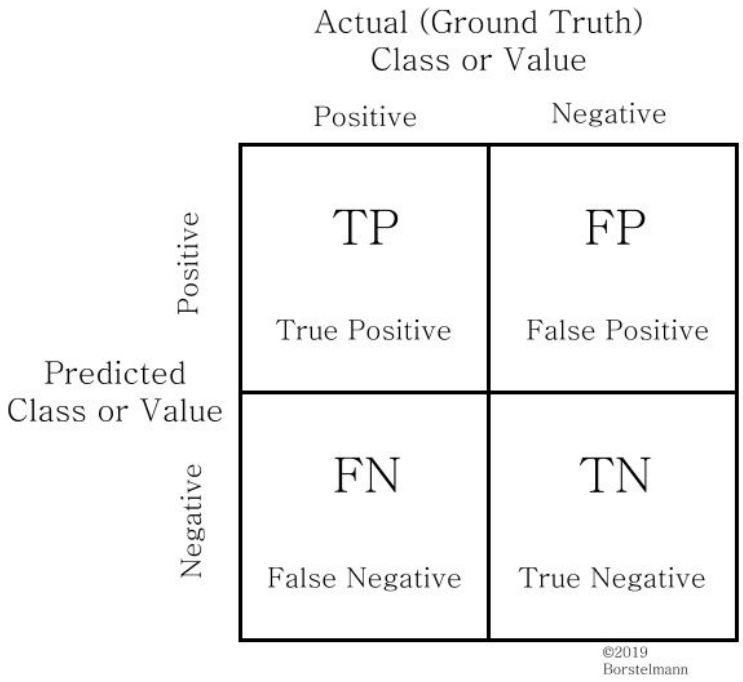


Figure 1 - The basic Confusion Matrix

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'.

Of course, if we simply add together all values, we get the total number (**N)** of positive and negative examples : **N = TP+FP+TN+FN**. Sensitivity and Specificity alone may not be sufficient, so other measures have been proposed for use.

BASIC SINGLE VALUE SUMMARIES:

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

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

Accuracy is relatively intuitive, measuring correctly predicted observations compared to all observations. However, accuracy can fail as a predictive measure when there is a large FP:FN ratio, for example 50:1. This is known as a **class imbalance** problem, and arises frequently in ML, where it carries over into end algorithmic performance as well.[10] In practice, a 4:1 ratio might not be significant, but a 100:1 ratio certainly could. 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] Data engineering through data augmentation and over or undersampling techniques can be used to address the class imbalance problem, but runs the risk of altering positive prevalence between the engineered data and real-world test data.

The **F1** Score, also known as the **DICE coefficient** has also been proposed for use.

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

It takes both false positives and false negatives into account, and can be used with an uneven distribution of classes in the confusion matrix. F1 will also seek a balance between Sensitivity and Specificity. Useful in segmentation tasks, it cannot be used to 'rule out' as it does not incorporate TN.

MORE ADVANCED SINGLE VALUE SUMMARIES:

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 and is a special case of the **Pearson correlation coefficient**, ρ . 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.

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

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.

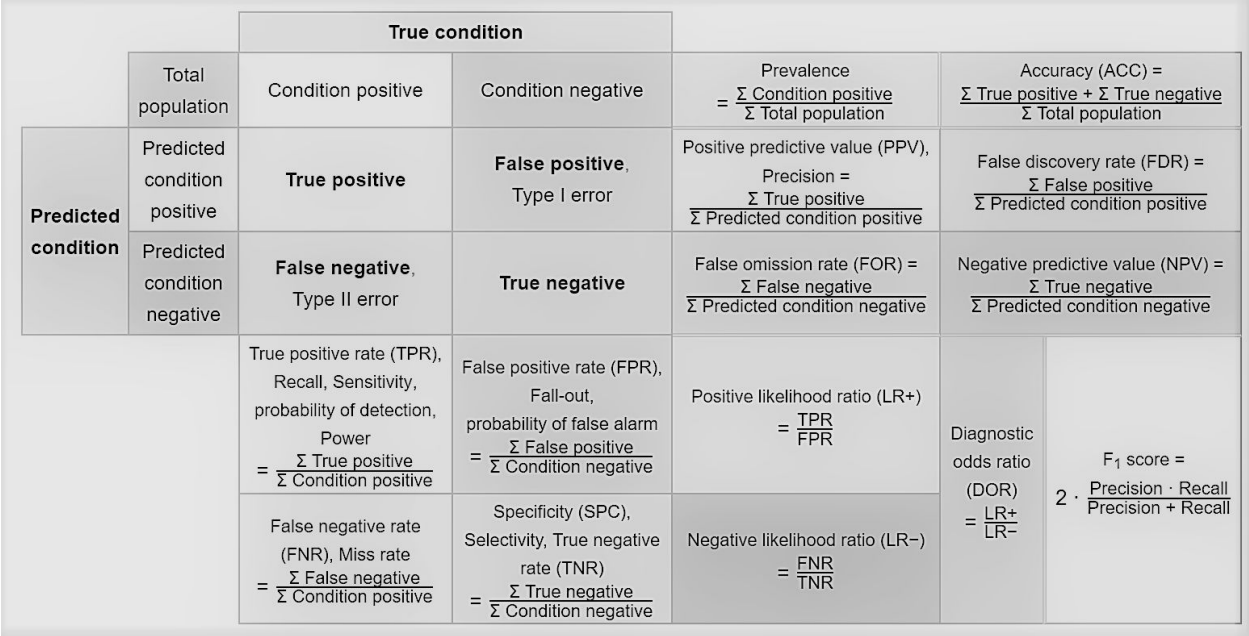


Figure 2 - Confusion Matrix Statistics[12]

The **kappa** (**K**) **statistic** is frequently used to calculate interobserver variability between radiologists. However, it is also of use in a confusion matrix and can be used as a summary statistic, considering the ground truth observations (TP & TN) to the actual observations. It takes this observed accuracy (**OA**) and compares it to expected accuracy (**EA**). The statistic returns a number from 0-1 with 1 indicating perfect agreement between the observers and 0 representing complete disagreement. Kappa can be used in settings of imbalanced data, and can also be expanded to multi-class problems. **OA** is equivalent to ACC. Expected accuracy (**EA**) is calculated as follows, but is frequently 0.5 in a 2x2 binary classification matrix:

$$EA = \frac{(TN+FP)*(TN+FN)+(FN+TP)*(FP+TP)}{(TP+FP+FN+TN)^2}$$

And Kappa is then calculated as

$$\kappa = \frac{ACC-EA}{1-EA}$$

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)**. One advantage is that ROC is prevalence-invariant, independent on whether what is being tested is common or not.

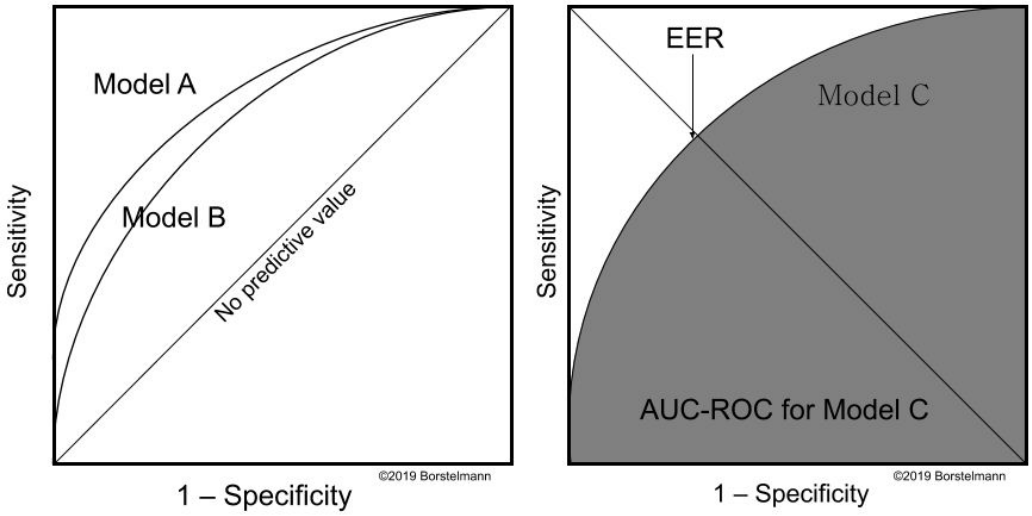


figure 3a and 3b - ROC and AUC-ROC

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**) 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.

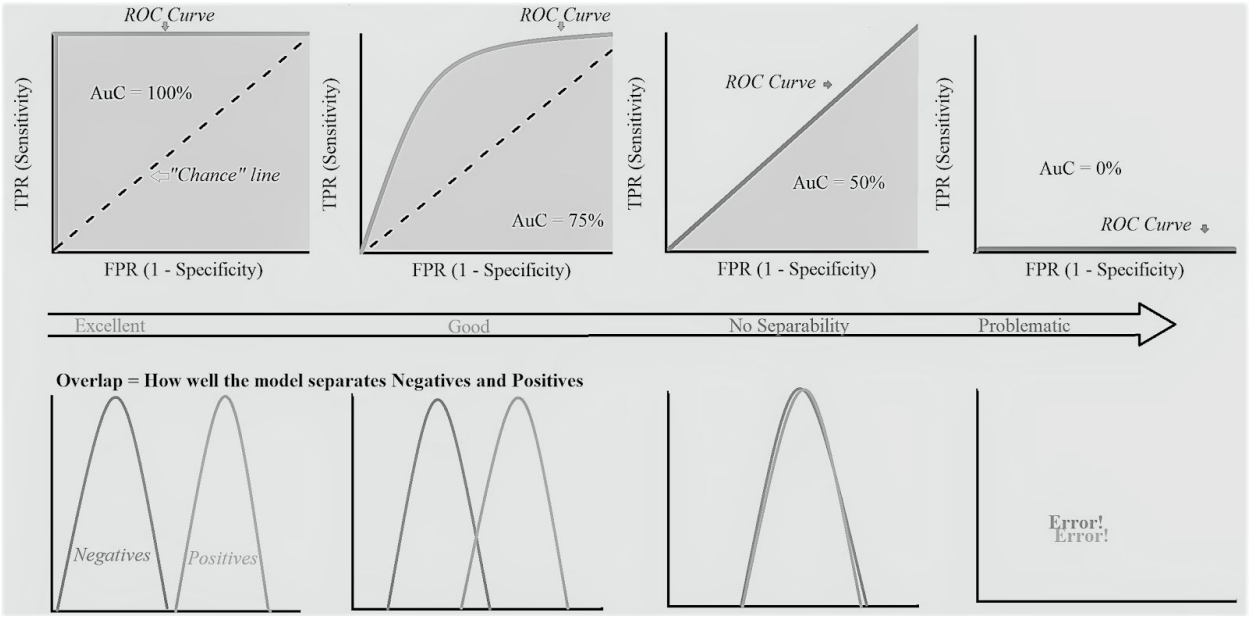


figure 4 – interpreting ROC curves[16]

Comparing ROC curves is a considered a quick way to a better test. But is it? An apples to apples comparison requires that the same underlying dataset be used between two ML model's raw ROC and AUC ROC for a valid comparison. If that is met, then even if two different models have equal AUC ROC's, they may not be equally good for the same purpose.[13] If the curves cross, this indicates that one test is superior to the other in some circumstances, like screening, but inferior in others such as definitive diagnosis. Since AUC ROC treats both sensitivity and specificity equally, a test or model with a lower score could potentially clinically outperform a higher score. ROC and AUC ROC are insensitive to class imbalances, and therefore can also suffer from a similar problem as accuracy.[15,21]

Once a ROC curve has been generated, the question can be raised where does one operate on the curve? The most extreme areas of the curve may, and frequently do, represent trivial cases where no one would wish to operate. The question is then raised, what sensitivity and specificity are to be chosen for operating parameters in a real world environment? In clinical practice, the radiologist chooses how much of an over/under caller they are, affecting their specificity. They are at a single operating threshold, whereas multiple threshold possibilities exist on the ROC. ROC and AUC ROC include performance over non-clinically relevant and possibly illogical thresholds.[14] This is why comparing a single radiologist against a full ROC and particularly an AUC ROC is a generally unfair comparison of man vs. machine!

The **Error Equal Rate** (**EER**) or Crossover Rate has also been suggested as one solution to the thresholding problem, and is used frequently in biometrics. It is simply the point on the ROC curve at a threshold where FP=FN; frequently the intersection of the curve with a diagonal line inverse to the null hypothesis line on the ROC curve. See figure 3b. An additional

choice can be made at the point where the difference between sensitivity & specificity is maximized. (valverde & lobo REF)

Different schema have been suggested for improving the AUC ROC, involving weighting. When little real-world experience exists with tests, in the early stage of test assessment, ROC comparison is reasonable. However, established tests in clinical use are subject to contexts of prevalence, and misclassification costs. A weighted formula for CT colonography screening where benefit of early disease detection outweighs the theoretical cost of a missed cancer was proposed as 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 established however.

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

PRECISION-RECALL CURVE, AUC-PR, 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 PR,** also termed **Average Precision (AP),** can also be calculated through an integral and allows for a single value summary comparison between models or tests.[17]

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 a better ROC, AUC ROC, PRC and AP than another can confidently be evaluated as better. Unlike ROC, 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]

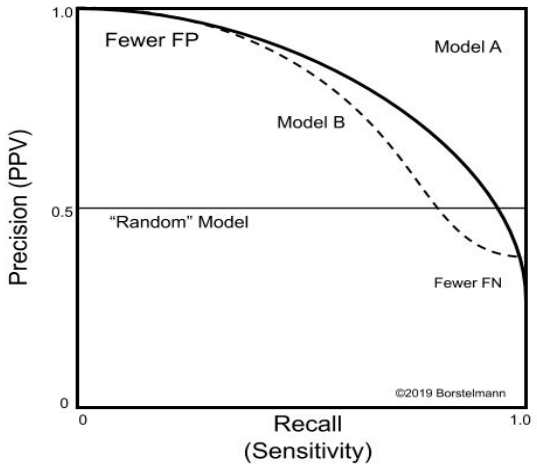


Figure 5 - PR curve

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 the 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:

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

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

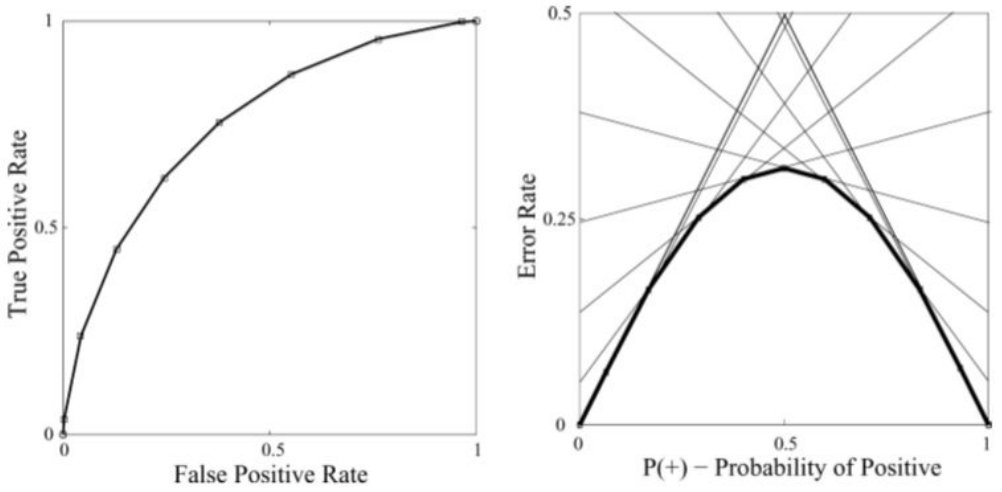


Fig. 4 (a) Ten ROC points and their ROC convex hull — (b) Corresponding set of cost lines and their lower envelope

Figure 6 - Cost Curves from Drummond et al. [19]

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, or more simply, lower. Classifiers (ML models) can be compared in this manner, and potentially optimized for various different criteria. Unlike ROC curves, 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 underperforms is in the setting of imbalanced data.[15] AUC CC could provide a potentially comparable scalar, but more experience would be necessary.

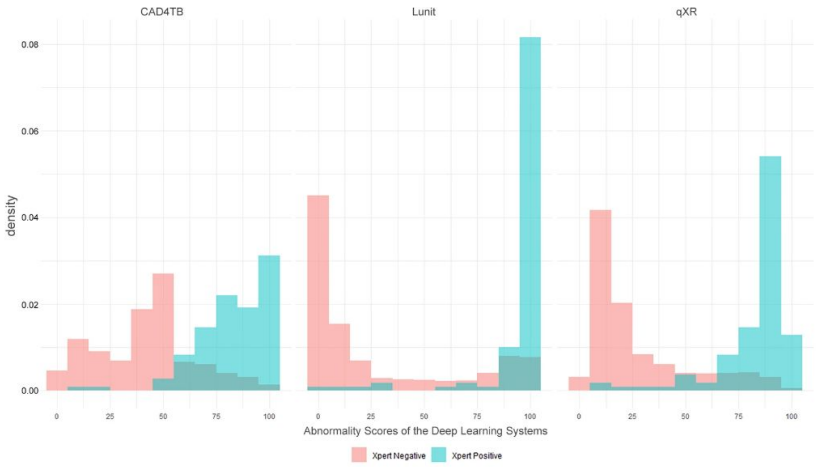
An Example

As this article was going to press, an interesting report was published comparing three AI/ML systems for the detection of tuberculosis[23]. The authors took three commercially available deep learning AI systems, CAD4TB v6 (Delft - Holland), qXR v2 (Qure.ai - India) and INSIGHT v4.7.2 (Lunit - South Korea) which output a probability score for TB, and applied them a previously obtained dataset from Nepal and Cameroon of patients with symptoms suggestive of tuberculosis, with chest x-rays and the highly specific XPERT RIA test available for all..

Frequency distributions of the output probabilities of the AI systems were given for confirmed positive and negative testing with skewness calculations. This is useful, as it shows the degree of separation of the output, reminiscent of the lower images on figure 4 above.

Figure 1

From: Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems



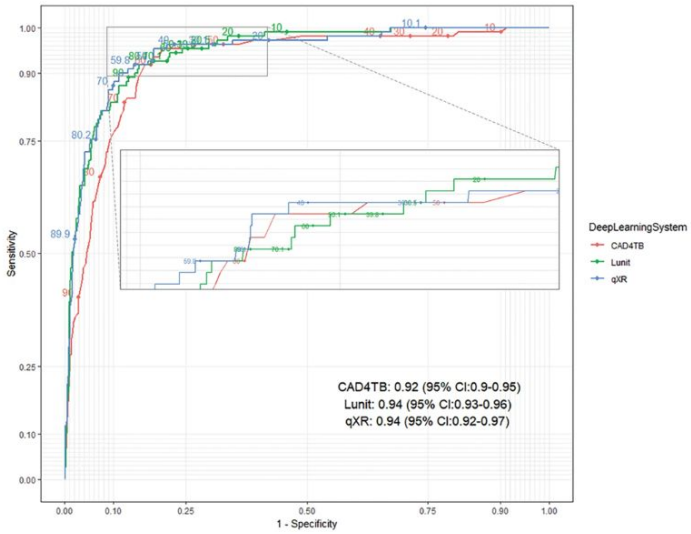
Frequency distribution of the abnormality scores of CAD4TB, Lunit, and qXR.

Figure 7 - Frequency Distribution from Zhi et al.

The authors provide multiple ROC curves and AUC ROC to compare the AI systems. AUC ROC calculated were: Lunit (0.94, 95% CI:0.93-0.96), qXR (0.94, 95% CI:0.92-0.97), CAD4TB (0.92, 95% CI:0.90-0.95). No statistical difference between the systems was seen. Sensitivities, specificities, and accuracies with 95% confidence intervals for human readers and the three systems were also provided. As one would expect from the high values, the extremely similar ROC curves make it difficult to ascertain superiority of any one algorithm over the other - see figure 8.

Figure 2

From: Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems



The ROC curves of CAD4TB (v6), Lunit (v4.7.2) and qXR (v2) using Xpert results as the reference (n = 1196).

Figure 8 - ROC curve from Zhi et al.

The authors extensively investigate thresholding for the optimization of inexpensive CXR screening vs. the expensive XPERT test, and found optimal cutoffs different in Nepal and Cameroon, a salient finding. Their recommendation was to recognize that threshold setting (in other words, setting where the AI will function on the ROC curve) is necessary even for commercial products, and must be done after evaluation on the local population.[23]

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. The recent example cited above is a good starting point. Too many investigators have insufficient experience with these metrics beyond sensitivity and specificity and perhaps the ROC curve.

It may be helpful to evaluate the frequency distribution of probability output scores of AI systems to ascertain the degree of separation by the classifier between normal and abnormal, as a way to evaluate the overlap between the two classes and how strongly the model separates positive and negative classes.

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, which gives the population number N in each class and is represented like this for a typical 2x2 confusion matrix: {TP:861}{FP:240}{TN:3002}{FN:743}. This may have utility in model evaluation, particularly on a formal basis by regulatory agencies, as such information would lend context to multiple measures like Accuracy, F1, and ROC..

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. We can create ensembles of multiclass One-vs-All approaches, and many of these measures can be extended as well. Finally, further experience with the Cost Curve method would be needed to decide if this method was superior to others mentioned.

Computer scientists are also trying to address the problem. Performance measures described do not assess the model's reliability. Prior attempts using Bayesian logic and bootstrapping fail after training the algorithm. The **Resampling Uncertainty Estimation** (**RUE)**, was recently unveiled in an effort to provide a true algorithmic audit.[22] RUE estimates the amount that a prediction would change if the model had been fit on different training data and generates a curve for which an associated AUC RUE can be generated. It is intriguing, but likely requires further development as it is highly computationally expensive.

Conclusion

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 sensitivity, specificity, 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, ROC, and a third statistic such as MCC.
2. Disclosure of the class imbalance ratio for any dataset or AI model should be strongly encouraged, as well as output probability distributions, 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 PR over ROC, and probably both should be routinely provided. A superior model will likely have both higher AUC ROC and AP.
4. Consideration to the development of the Drummond Cost Curve, a statistical method of calculating W in the Net Benefits function and new discriminative metrics such as the RUE should be given. Further basic research would help here.
5. A perfect AI/ML evaluation paper would include not only the confusion matrix, sensitivity and specificity, but the class imbalance ratio, ACC, F1, MCC, ROC, ROC AUC, PRC and AP statistics with 95% confidence intervals.
6. Considerations relating to external validation, spectrum bias, model drift, and edge cases are in no way minimized by the foregoing, and of course model performance should be evaluated on the same data.
7. On-site calibration and monitoring will need to be performed after installation to optimize for the population in service. AI does not seem to be a 'fire and forget' product.

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.
21. Fawcett, T., (2003). ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. HP Laboratories Publication.
22. Shulam, P., Saria, S. (2019). Can you trust this prediction? Auditing Pointwise Reliability After Learning. In *arXiv.org*. Retrieved from https://arxiv.org/abs/1901.00403
23. Qin, Z. Z., Sander, M. S., Rai, B., Titahong, C. N., Sudrungrot, S., Laah, S. N., … Creswell, J. (2019). Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Scientific Reports*, *9*(1), 15000. https://doi.org/10.1038/s41598-019-51503-3