A More Intuitive Interpretation of the Area Under the ROC Curve

A. Cecile J.W. Janssens, PhD

Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta GA, USA.

Corresponding author: Professor A. Cecile J. W. Janssens, Department of Epidemiology, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, Georgia 30322, USA. E-mail: cecile.janssens@emory.edu, Telephone: +1 404 727 6307, Fax: +1 404 727 8737

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3468v1 | CC BY 4.0 Open Access | rec: 15 Dec 2017, publ: 15 Dec 2017

Abstract

The area under the receiver operating characteristic (ROC) curve (AUC) is commonly used for assessing the discriminative ability of prediction models even though the measure is criticized for being clinically irrelevant and lacking an intuitive interpretation. Most of the criticism is traced back to the fact that the ROC curve was introduced as the discriminative ability of a binary classifier across all its possible thresholds. Yet, this is not the curve's only interpretation. Every tutorial explains how the coordinates of the ROC curve are obtained from the risk distributions of diseased and non-diseased individuals, but it has not become common sense that the ROC plot is another way of presenting these risk distributions. This alternative perspective on the ROC plot invalidates most limitations of the AUC and attributes others to the underlying risk distributions. Moreover, the separation of these distributions, represented by the area between the curves (ABC), can be a straightforward, alternative measure for the discriminative ability of prediction models.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3468v1 | CC BY 4.0 Open Access | rec: 15 Dec 2017, publ: 15 Dec 2017

In 1971, Lee Lusted applied signal-detection theory to the interpretation of chest roentgenograms and introduced the receiver operating characteristic (ROC) curve in medicine to present the percentage of true positive against the false positive diagnoses for different decision criteria applied by a radiologist.[1] A decade later, Hanley and McNeil proposed the area under this ROC curve (AUC) as a single metric of diagnostic accuracy for "rating methods or mathematical predictions based on patient characteristics."[2] The AUC is the most commonly used metric for assessing the ability of predictive and prognostic models to discriminate between individuals who will or will not develop the disease (here referred to as diseased and non-diseased individuals).

Despite its popularity, the AUC is frequently criticized, and its interpretation has been a challenge since its introduction in medicine.[2] The AUC value is generally described as the probability that predicted risks correctly identify a random pair of a diseased and non-diseased individual, but this probability is considered clinically irrelevant as doctors never have two random people in their office,[3,4] they are only interested in the clinically relevant thresholds of the ROC curve not in others,[5] and they often want to distinguish multiple risk categories for which they need more than a single threshold.[6] Also, the AUC is considered insensitive as the addition of substantial risk factors may have only minimal improvement of the AUC, especially when added to a baseline model with already good discrimination.[4,7-9] Much of this criticism about AUC can be traced back to the ROC curve, suggesting that a more intuitive interpretation of the ROC could change the appreciation of the AUC.

Every tutorial explains how the coordinates of the ROC curve are obtained from the risk distributions of diseased and non-diseased individuals. It can be shown that the ROC curve is merely an alternative way of presenting risk distributions that does not assume risk thresholds. This article shows how risk distributions transform into ROC curves and how, in turn, the ROC curves inform about the shapes and overlap of the underlying risk distributions. The interpretation and limitations of the AUC are re-evaluated from this alternative perspective, and a more straightforward measure of discriminative ability is put forward.

From risk distributions to ROC curve

In empirical studies that investigate the development or validation of prediction models, predicted risks can be presented as separate distributions for diseased and non-diseased individuals (**Figure 1a**). The separation between the distributions, indicated by the non-overlapping areas, gives the models their discriminative ability: the further the distributions are separated, the better the prediction model can differentiate between the two populations because more diseased individuals have higher risks than the non-diseased.

The risk distributions can also be presented as *cumulative* distributions, where the y-axis presents the proportion of individuals who have equal or lower predicted risks at each predicted risk (**Figure 1b**). The difference between the distributions of diseased and non-diseased reflects the same separation as the distributions in Figure 1a. The two non-overlapping areas are now one area, 'connected' at the same predicted risk that separated them in the previous figure.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3468v1 | CC BY 4.0 Open Access | rec: 15 Dec 2017, publ: 15 Dec 2017

Next, we can change the x-axis and replace predicted risks by the (cumulative) proportion of non-diseased individuals at each risk (**Figure 1c**). With the proportion on the x-axis, the distribution of non-diseased individuals is now a diagonal line as its x and y-axes are the same, and the distribution of diseased individuals is the curved line. The difference between the curves still reflects the difference between the risk distributions in Figure 1a. When the predicted risks are interpreted as thresholds, the proportions at each risk are the proportions of diseased and non-diseased individuals below the threshold, which equal 1-sensitivity and specificity. It then follows that the ROC plot can be obtained by flipping both axes (**Figure 1d**).

This transformation shows that the diagonal line is an integral part of the ROC plot, not just a reference line of random prediction. The diagonal line represents one of the two risk distributions and the difference between the diagonal line and the ROC curve indicates the separation between the distributions that gives prediction models their discriminative ability.

Reappraisal of AUC limitations

As an alternative presentation of risk distributions, it follows that the ROC plot does not assume risk thresholds. While the ROC curve can be used to determine the sensitivity and specificity at single risk threshold, this does not need to be its primary interpretation. The risk distributions and the separation between them are relevant for prediction models irrespective of the number of thresholds that is considered.

The AUC is commonly described as the probability that a random individual from the diseased population is more likely to have a higher predicted risk than an individual from the non-diseased population. This explanation still holds: the probability is higher when the risk distributions are further separated. These random individuals can be considered as pairs, which is how the AUC value is calculated from the Wilcoxon Rank Sum Test or Somers' d,[2,10] but the consideration of pairs is not essential or mandatory.

AUC has been criticized for being insensitive to detect improvements in the prediction that result from adding risk factors with stronger effects.[8,9,11,12] As the area under a curve that is a transformation of a risk distribution, it is evident that this insensitivity is not a limitation of the metric, but of the prediction algorithms. Improving prediction models requires adding predictors with very strong impact on disease risk to further separate the risk distributions, which is difficult especially when risk distributions are already separated and have higher 'baseline' AUC. When adding predictors does not improve the AUC, it means that the ROC curves of the baseline and updated models are the same; adding the predictors may have changed the predicted risks, but each sensitivity comes with the 'same' specificity and vice versa.

Finally, the criticism that the AUC lacks clinical relevance and omits the consideration of costs and harms in weighing false positives against false negatives[13,14] is valid but concerns the inappropriate use of the measure rather than its shortcomings. The AUC is nothing more than a measure related to the separation of the risk distributions of diseased and non-diseased individuals. The optimal threshold on the ROC curve may be irrelevant and suboptimal from a clinical perspective. The decision whether a prediction model is useful to guide medical

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3468v1 | CC BY 4.0 Open Access | rec: 15 Dec 2017, publ: 15 Dec 2017

decisions is not determined by its discriminative accuracy alone but requires additional evaluations such as the prevalence, predictive value, the decision impact of the test results, the implications of (false-)positive and negative results, and others.

The Area Between the Curves

Hanley and McNeil developed a single measure to express the discriminative accuracy that is presented by the ROC curve. The AUC is a valid quantification of the area under the curve, but it inherited the curve's shortcomings, limitations, and problems in the interpretation. When the ROC plot is interpreted as a transformation of risk distributions, these shortcomings and limitations no longer apply, but the interpretation of the metric remains challenging: the AUC is not the most obvious and intuitive measure of discriminative ability.

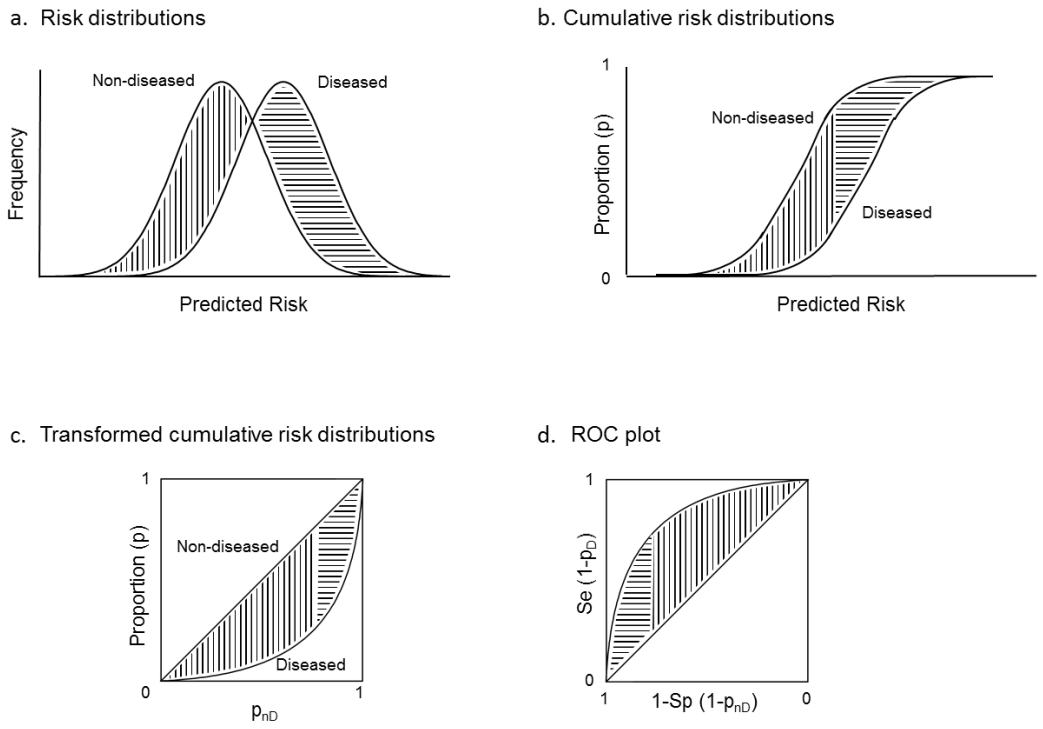
The present analysis showed that, when the ROC plot is interpreted as an alternative way of presenting risk distributions, the diagonal line is not merely a reference line representing a 'random', non-informative test, but it is the risk distribution of non-diseased individuals (**Figure 1**). The separation of the risk distributions is then indicated by the area between the curves (ABC): the higher the area between the curves, the more separation between the distributions and the higher the discriminative ability. The ABC value is equivalent to Somers' d,[10] a non-parametric rank correlation that is known from one of the formulas to obtain the AUC: AUC=(d+1)/2. This formula shows that the AUC value is calculated *from* the ABC.

The ABC has a linear relationship with the AUC, but its interpretation is more intuitive. First, not AUC but ABC *is* the degree of separation between the risk distributions of diseased and non-diseased individuals; the AUC is calculated from the ABC, not vice versa. Second, its scale ranges from 0 to 1, with clear interpretation: 0 means no separation between the distributions, 1 is total separation, and the values in-between indicate the proportion of separation. An ABC value of 0.40 means that 40% of each risk distribution does not overlap with the other or that the overlap of the distributions is 60%. And finally, the ABC has no other straightforward interpretation that it can easily be confused with, in contrast to the AUC that is frequently interpreted as a measure of accuracy or compared to the predictive ability of tossing a coin (which is only correct when applied to random pairs of a diseased and non-diseased individuals).[15,16]

The ABC expresses the essence of discriminative ability, the degree of separation between the risk distributions of diseased and non-diseased individuals. These risk distributions and the separation between them are relevant for all prediction models irrespective of the number of thresholds that is considered or their intended use. A more intuitive measure may help to understand this relevance.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3468v1 | CC BY 4.0 Open Access | rec: 15 Dec 2017, publ: 15 Dec 2017

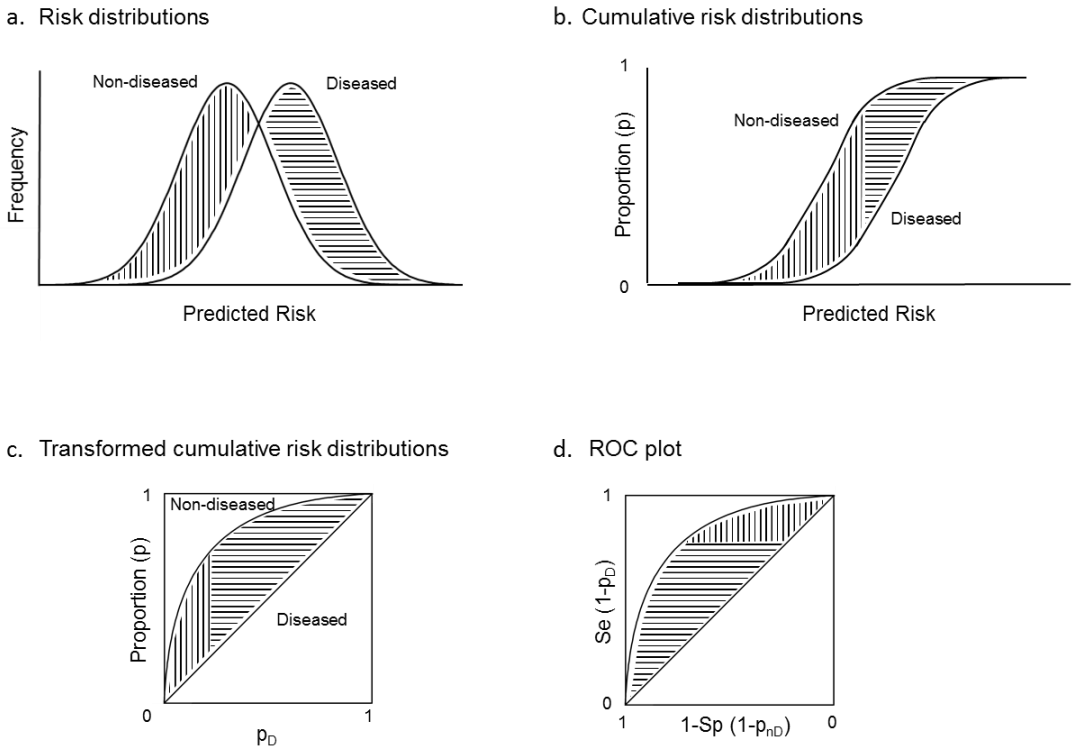
Figure 1. From risk distributions to the ROC curve



a. Risk distributions of diseased and non-diseased individuals. Separation of the distributions creates two nonoverlapping (pattern) and one overlapping (white) areas. b. Cumulative risk distributions. The two nonoverlapping areas are now one area, connected at the same predicted risk that separated them in Figure 1a. c. Transformed cumulative risk distributions. The x-axis presents the proportion of non-diseased individuals (p_{nD}) at each predicted risk instead of the predicted risk. The proportion p equals p_D for diseased and p_{nD} for non-diseased individuals. d. ROC plot. This plot is obtained by reversing both the x-axis and y-axis of Figure 1c. The same ROC plot is obtained when the x-axis in Figure 1c had shown the proportion of diseased individuals (Supplementary Figure 1). Sensitivity (Se) is the percentage of diseased individuals who have predicted risks higher than the threshold. Specificity (Sp) is the percentage of non-diseased who have predicted risks lower than the threshold.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3468v1 | CC BY 4.0 Open Access | rec: 15 Dec 2017, publ: 15 Dec 2017

Supplementary Figure 1 From risk distributions to the ROC curve: alternative derivation



a. Risk distributions of diseased and non-diseased individuals. Separation of the distributions creates two non-overlapping (pattern) and one overlapping (white) areas. b. Cumulative risk distributions. The two non-overlapping areas are now one area, connected at the same predicted risk that separated them in Figure a. c. Transformed cumulative risk distributions. The x-axis presents the proportion of diseased individuals (p_D) at each predicted risk instead of the predicted risk itself. Proportion p equals p_D for diseased and p_{nD} for non-diseased individuals. d. ROC plot. This plot is obtained by reversing both the x-axis and y-axis and swapping the x-axis and y-axis of Figure c. Sensitivity (Se) is the percentage of diseased individuals who have predicted risks higher than the threshold. Specificity (Sp) is the percentage of non-diseased who have predicted risks lower than the threshold.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3468v1 | CC BY 4.0 Open Access | rec: 15 Dec 2017, publ: 15 Dec 2017

References

1. Lusted LB. Decision-making studies in patient management. N Engl J Med 1971;284:416-24.
2. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.
3. Parikh CR, Thiessen-Philbrook H. Key Concepts and Limitations of Statistical Methods for Evaluating Biomarkers of Kidney Disease. Journal of the American Society of Nephrology 2014;25:1621-9.
4. Pepe MS, Janes HE. Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. J Natl Cancer Inst 2008;100:978-9.
5. Halligan S, Altman DG, Mallett S. 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 2015;25:932-9.
6. Flach P. ROC Analysis. In: Sammut C, Webb G, eds. Encyclopedia of Machine Learning: Springer US; 2010:869-75.
7. Pencina MJ, D' Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. Statistics in Medicine 2008;27:157-72.
8. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115:928-35.
9. Ware JH. The Limitations of Risk Factors as Prognostic Tools. New England Journal of Medicine 2006;355:2615-7.
10. Somers RH. A new asymmetric measure of association for ordinal variables. American Sociological Review 1962;27:799-811.
11. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008;27:157-72; discussion 207-12.
12. Pepe MS. Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic, Prognostic, or Screening Marker. American Journal of Epidemiology 2004;159:882-90.
13. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning 2009;77:103-23.
14. Samawi HM, Yin J, Rochani H, Panchal V. Notes on the overlap measure as an alternative to the Youden index: How are they related? Statistics in Medicine 2017;36:4230-40.
15. Takwoingi Y, Riley RD, Deeks JJ. Meta-analysis of diagnostic accuracy studies in mental health. Evid Based Ment Health 2015;18:103-9.
16. Farinholt P, Park M, Guo Y, Bruera E, Hui D. A Comparison of the Accuracy of Clinician Prediction of Survival versus the Palliative Prognostic Index. J Pain Symptom Manage 2017.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3468v1 | CC BY 4.0 Open Access | rec: 15 Dec 2017, publ: 15 Dec 2017