Use of ROC curve analysis for prediction gives fallacious results: Use predictivity-based indices

Indrayan A, Malhotra RK[1], Pawar M

Department of Clinical Research, Max Healthcare, New Delhi, [1]Dr BRA Institute-Rotary Cancer Hospital, All India Institute of Medical Sciences, New Delhi, India

Address for correspondence:
Dr. Indrayan A,
E-mail: a.indrayan@gmail.com

Received : 25-09-2023
Review completed : 18-02-2024
Accepted : 26-02-2024
Published : 18-04-2024

ABSTRACT

The area under the ROC curve is frequently used for assessing the predictive efficacy of a model, and the Youden index is commonly used to provide the optimal cut-off. Both are misleading tools for predictions. A ROC curve is drawn for the sensitivity of a quantitative test against its (1 – specificity) at different values of the test. Both sensitivity and specificity are retrospective in nature as these are indicators of correct classification of already known conditions. They are not indicators of future events and are not valid for predictions. Predictivity intimately depends on the prevalence which may be ignored by sensitivity and specificity. We explain this fallacy in detail and illustrate with several examples that the actual predictivity could differ greatly from the ROC curve-based predictivity reported by many authors. The predictive efficacy of a test or a model is best assessed by the percentage correctly predicted in a prospective framework. We propose predictivity-based ROC curves as tools for providing predictivities at varying prevalence in different populations. For optimal cut-off for prediction, in place of the Youden index, we propose a P-index where the sum of positive and negative predictivities is maximum after subtracting 1. To conclude, for correctly assessing adequacy of a prediction models, predictivity-based ROC curves should be used instead of the usual sensitivity-specificity-based ROC curves and the P-index should replace the Youden index.

KEY WORDS: Area under ROC curve, C-index, P-index, prediction models, predictivity, predictivity-based ROC curve

Introduction

Prediction of future events based on past occurrences is considered a legitimate scientific activity, although this assumes that the future will generally follow the past trend. Many prediction models appear in the literature every day. Steyerberg[1] reported from the Web of Science that the number of publications on prediction models has risen from less than 9000 in 1993 to more than 50,000 in 2017.

A model is a simplified version of a process and helps to understand the components of the process. Statistical models are often used for the prediction of an event. Model-based prediction of diagnosis and prognosis lately received a great boost with the emergence of CoViD-19 and the urgency to provide tools for clinical decisions. In the early phase of the pandemic, Wynants *et al.*[2] reviewed 232 models, including 118 diagnostic models for detecting CoViD-19 and 107 prognostic models for predicting severity and mortality. So many models have appeared in a short time and have been proposed as prediction tools.

The most used indicator of the validity of a prediction model for binary outcomes based on quantitative characteristics is C-index or C-statistic. This is measured by the area under the receiving operating characteristic (AUROC) curve, which is drawn for

Access this article online	
Quick Response Code:	Website: www.jpgmonline.com
	DOI: 10.4103/jpgm.jpgm_753_23
	PubMed ID: 38668827

How to cite this article: Indrayan A, Malhotra RK, Pawar M. Use of ROC curve analysis for prediction gives fallacious results: Use predictivity-based indices. J Postgrad Med 2024;70:91-6.

sensitivity (true positive rate) against (1 – specificity) (false positive rate). For example, in the case of mortality models, the ROC curve would be drawn for

P (model estimated chance of death ≥ *l* in those who die)

vs.

P (model estimated chance of death ≥ *l* in those who survive)

for different values of *l* between 0 and 1. These are the sensitivity and (1 – specificity), respectively, in this case. The deaths and survivals are known, and the model estimates the chance of each person belonging to one group or the other. Sensitivity and specificity are indicators of classification performance or of discrimination of the *known* outcomes and not of prediction of the unknown. C-index has a retrospective nature because the outcome is already known and is a summary measure of the average classification accuracy across the spectrum of values. This continues to be so even when the model is internally or externally validated because these validations also mostly use AUROC or C-index. The same is true for the Youden index, which is used to derive the optimal cut-off with the largest sum of sensitivity and specificity. These can give misleading predictions, and it is no wonder that most models fail to perform well in actual setups. Predictivity intimately depends on the prevalence, and this is mostly ignored by sensitivity and specificity, particularly where the cumulative probability function of the test in the sample is not similar to that in the population,[3] and consequently also by the ROC curves and the Youden index. We explain this fallacy with several examples and propose a new method of assessing and reporting predictivity and a new index for optimal cut-off.

Common Misuse of AUROC Curve (C-Index) for Reporting Predictivity

ROC curves have been wrongly used for two kinds of inferences in our context. One, to assess the overall predictivity of the model by the AUROC, and two, to find not only the sensitivity-specificity but also PPV-NPV at the optimal cut-off where the sum of sensitivity and specificity is the highest. Both these assessments can provide misleading results in applications for predictivity, and the following examples illustrate that such misleading assessments are common in the literature.

In a study of the usefulness of a combination of mammography and scientimammography in suspected breast cancer cases, Buscombe *et al*.[4] concluded, based on the AUROC curve, that the combination is more accurate in *predicting* the disease than either modality alone. Wei and Lu[5] used the AUROC curve to report that the models built on all single-nucleotide polymorphisms gained more accuracy for risk *predictions* than the ones built on common variants alone. Liu *et al*.[6]used the ROC curve for assessing the value of the size of the antepartum foramen ovale in the *prediction* of the puerperal arterial septal defect. Gazi *et al*.[7]used ROC curve analysis to conclude the high average *prediction* accuracy of hypothetical proteins from Shigella flexneri in diarrheal disease. Johnston *et al*.[8] developed an online tool for condition-independent dimer *prediction* and assessed its accuracy through ROC analysis. Mei *et al*.[9] developed a CoViD-19 mortality risk *prediction* algorithm and reported its good performance based on high AUROC. Wang *et al*.[10] used a high value of AUROC to conclude that a combination of laboratory parameters could be an effective *predictor* of mortality in CoViD-19 patients. The popular systematic review of models for diagnosis and prognosis of CoViD-19 by Wynants *et al*.[2] discussed the *predictivity* of different models primarily based on the C-index, although they considered several other parameters as well. For other recent studies using such an index for prediction, see Hou *et al*.,[11] Lin *et al*.,[12] and Hsu *et al*.[13] A large-scale screening study on 323,344 participants by Pan *et al*.[14] for lung cancer inferred a high degree of *predictive accuracy*, mostly based on AUROCs, although they earlier rightly called this discriminative capability. They also used average net benefit for assessing clinical utility. Misuse of AUROC or C-index for prediction is so widespread that it is considered a norm, and its fallacy tends to be overlooked.

The other inference generally drawn from a ROC curve analysis is regarding the optimal cut-off that gives the maximum Youden index (sensitivity + specificity – 1). Based on the highest sum of sensitivity and specificity, Adjahoto *et al*.[15] concluded that a uterine height of less than 32 cm is optimal to suspect fetal hypotrophy. They not only provided sensitivity and specificity at this cut-off but also PPV and NPV at the same cut-off in support of their conclusion. Olivares-Morales *et al*.[16] built ROC curves and obtained optimal animal threshold prevalence for the qualitative prediction of human oral bioavailability from animal data. This optimal value too was based on the highest Youden index. Both these papers ignored that PPV and NPV heavily depend on prevalence and can be very different from sensitivity and specificity.

Predictivity Intimately Depends on Prevalence

At the risk of stating the well-known, it is necessary to highlight the dependence of predictivities on the prevalence of the condition under study. We know from Bayes Rule that

$$\text{Positive predictivity, } P(+) = \frac{Se * p}{Se * p + (1 - Sp) * (1 - p)}, \text{ and}$$

$$\text{Negative predictivity, } P(-) = \frac{Sp * (1 - p)}{Sp * (1 - p) + (1 - Se) * p},$$

where Se = Sensitivity, Sp = Specificity, and p = Prevalence.

Sensitivity and specificity are generally stable indicators; however, predictivities quickly change according to the local prevalence. For fixed sensitivity and specificity, the PPV increases and NPV decreases as the prevalence increases from 1% to 50% [Figure 1]. The rate of change depends on the values of sensitivity and specificity. The increase in PPV with increasing prevalence is sharp when sensitivity is high, and the decrease in NPV is sharp when specificity is low. In most practical situations, $(Se + Sp) \geq 1$ and, in this case, it is easy to show that PPV $\geq p$ and NPV $\geq 1 - p$. This is an interesting result, possibly never mentioned earlier, and provides a lower bound for PPV and

NPV for most practical situations. As p is mostly small, it is not surprising that NPV would be generally high irrespective of the levels of sensitivity and specificity. In addition, it is interesting to note that $P(+) = p$ and $P(-) = 1 - p$ when $(Se + Sp) = 1$. Sensitivity follows a decreasing and specificity an increasing trend with the value of the test (in cases where higher values of the test indicate a positive outcome); however, PPV and NPV do not have this feature. Thus, they require extra precaution.

Because of such intimate dependence of predictivities on prevalence, the reporting of PPV and NPV for any test would be valid only for the prevalence in that study and will not apply to other setups where the prevalence is different. For example, Mei *et al.*[9] reported sensitivity = 74.2%, specificity = 97.2%, PPV = 71.7%, and NPV = 97.5% for their model for the CoViD mortality risk threshold of 30% [Table 1]. The mortality in their cases was 9.5%. If the mortality in another setup is 5%, these PPV and NPV translate to 58.2% and 98.6%, respectively, with the Bayes Rule stated earlier, and if the mortality is 15%, these become 82.4% and 95.5%, respectively. Note how PPV and NPV quickly change depending on the prevalence.

There are many examples in the literature that report PPV and NPV at the threshold obtained from the ROC curve analysis. Some of these were cited earlier in this communication. These thresholds maximize $(Se + Sp)$ and not PPV or NPV or their sum. Adjahoto *et al.*[15] reported PPV = 18.9% and NPV = 95.4% at the sensitivity-specificity-based optimal cut-off of 32 cm of uterine height for the prediction of fetal hypertrophy [Table 1]. The prevalence of hypertrophy in their cases was 9.9%. If the prevalence of hypertrophy is 20% in another population and the sensitivity (70.1%)–specificity (67.2%) remain the same, the NPV reduces to 90.0% and PPV goes up to 34.8%.

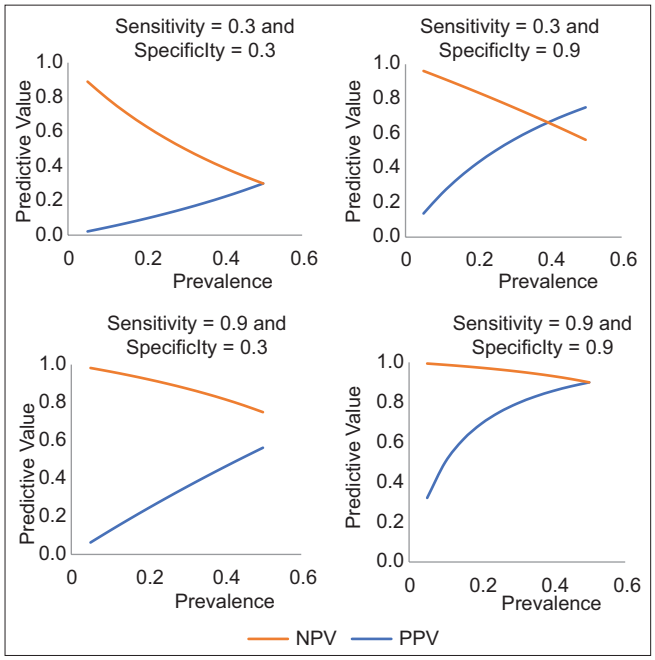

Figure 1: Trend of PPV and NPV with increasing prevalence at different levels of sensitivity and specificity

Budhiraja *et al.*[17] reported an optimal cut-off of 12.4 mg/L of CRP values for the highest $(Se + Sp)$ with sensitivity = 73.1% and specificity = 57.9% for classifying mortality in CoViD-19 cases admitted to a group of hospitals. The mortality in their cases with available CRP values was 7.8%. With this prevalence, the Bayes rule gives PPV = 12.8% and NPV = 96.2% at this optimal. Thus, all the reports on PPV and NPV must relate to the prevalence, and this seems to be missing in the literature. In addition, when PPV and NPV are reported at the Youden index-based optimal threshold, the implication is that this threshold is optimal for prediction as well. This is not correct and ignores the fact that PPV and NPV can be low at this threshold. The optimal for the highest (PPV + NPV) could be very different. This is illustrated in Table 2, which is based on the data given by Hou *et al.*[11] on SOFA score for 30-day mortality in cases of MIMIC-III with sepsis-3 and the data of Budhiraja *et al.*[17] on CRP values for CoViD mortality to which we had access. The cut-off for the maximum of $(Se + Sp)$ is the SOFA score 6.5 with the Youden index 0.291. The PPV and NPV at this cut-off are 33.7% and 87.2%, respectively. Similar findings were obtained for the data of Budhiraja *et al.*[17] on CRP values for CoViD mortality [Table 2]. Note that PPV and NPV are very different from sensitivity-specificity.

The best cut-off for predictivity is the point where (PPV + NPV − 1) is the highest. Call this predictivity index (P-index) on the pattern of the Youden index. This measures the overall efficacy of a level of a parameter for the prediction of positive and negative outcomes. For an illustration of the changes in the best cut-off and P-index, consider the study on the SOFA score just cited. The threshold for maximum (PPV + NPV) is the score 20.5 and not 6.5. The P-index for the threshold 20.5 in this case is a healthy 0.806 [Table 2]. Although 20.5 is extremely high for the SOFA score, it still has a negative predictivity of 80.6% for the data in this study. For CRP values in CoViD, the results are even more surprising. The best cut-off for the highest P-index is 334.6 mcg/dL. This is also high but has an overall predictivity of 0.450 against only 0.090 for the ROC curve-based best cut-off of 12.4 mcg/dL. Obviously, the ROC curve-based cut-off is not optimal for the prediction of the outcome. The data and full calculations are available with the Corresponding Author for anybody to review.

Table 1: Comparison of PPV and NPV with sensitivity-specificity in some studies at ROC-based best cut-off

Reference article	Variable investigated	Outcome studied	Best cut-off*	Sensitivity at best cut-off	Specificity at best cut-off	Youden Index	Prevalence	PPV at this cut-off	NPV at this cut-off
Mei *et al.*[9]	Regression model	Mortality risk	30%	74.2%	97.2%	0.714	9.5%	71.7%	97.5%
Adjahoto *et al.*[15]	Uterine height	Fetal hypertrophy	32 cm	70.1%	67.2%	0.373	9.9%	18.9%	95.4%

*Best cut-off for the highest sensitivity+specificity

Table 2: Comparison of ROC-based results and predictivity-based results in two studies

	ROC curve-based results (Incorrect for prediction)	Predictivity-based results (Correct for prediction)
SOFA score for mortality in MIMIC-III with sepsis (Hou *et al.*[11])		
Optimal cut-off	6.5	20.5
Sensitivity at this cut-off	55.5%	0.3%
Specificity at this cut-off	73.6%	100.0%
Youden index	0.291	0.003
Prevalence	0.195	0.195
PPV at this cut-off	33.7%	100.0%
NPV at this cut-off	87.2%	80.6%
P-index	0.209	0.806
CRP values for mortality in CoViD (Budhiraja *et al.*[17])		
Optimal cut-off	12.4	334.6
Sensitivity at this cut-off	73.1%	1.8%
Specificity at this cut-off	57.9%	99.9%
Youden index	0.302	0.017
Prevalence	0.078	0.078
PPV at this cut-off	12.8%	52.7%
NPV at this cut-off	96.2%	92.3%
P-index	0.090	0.450

Assessing Model Predictivity

Having established that the use of ROC curve analysis is fallacious for assessing the predictivity of a test or of a model, or for finding the optimal cut-off, what alternatives are available for assessing predictivity? Steyerberg[1] has discussed clinical prediction models in detail that could help. Pepe *et al.*[18] tried to integrate predictiveness with classification performance but that is for logistic models and not for the ROC curve analysis. Vickers and Elkin[19] proposed decision curve analysis for evaluating prediction models that requires a plot of "net benefit" to the patient against "threshold probability" at which a patient would opt for a treatment and identifies the range of threshold probabilities for which a model is of value.

While discussing the use and misuse of the ROC curves in risk prediction, Cook[20] discussed how C-statistic compromises the important contribution of individual biomarkers that can "naively" eliminate important risk factors from the prediction score, but he did not mention the fallacy arising from retrospective nature of this statistic. Vetter *et al.*[21] discussed the problem of predictivities' dependence on prevalence but ended up suggesting likelihood ratios, ignoring that these too are partially based on sensitivity and specificity. Hong and Choi[22] proposed the use of a total operating characteristic (TOC) curve where PPV and NPV at any threshold can be obtained as the angles of two right-angled triangles on the TOC curve. This is too complex, whereas we proposed a simple method to obtain the predictivity of a model.

Predictivity has a future perspective, and it is best estimated by a prospective study that begins from the antecedents and observes the outcomes. That would provide the estimated risk of outcome for different levels of the antecedents (sometimes termed predictors). Our concern in this communication is restricted to ROC-related issues, but several other aspects such as choice of predictors, correct measurements, appropriate methodology, and internal and external validations are all important as provided under Prediction model Risk Of Bias Assessment Tool (PROBAST) (Wolff *et al.*[23]). Wynants *et al.*[2]discussed most of these in the context of prediction models for diagnosis and prognosis of CoViD-19 but ignored the fact that C-index based on ROC curve analysis can lead to wrong results. This error is quite common.

As of now, we do not have a predictivity equivalent of the AUROC curve that can measure the overall predictive efficacy of a test or a model across different values, but we can easily obtain the percentage of cases correctly predicted for their outcome by a logistic model in a prospective setup, and both PPV and NPV can be obtained. This directly comes in the logistic output from most statistical software. Some studies already do this. For example, Budhiraja *et al.*[17] reported for a combination of eight inflammatory markers that their logistic model was able to correctly predict outcomes in 91.7% of cases of CoViD – 98.4% survivals but only 23.2% deaths. The study was based on the hospital records, but the format was statistically prospective from the level of the markers to the outcome. These numbers are valid estimates of the overall predictivity, PPV, and NPV. This example also highlights the differential performance of the model for predicting positive and negative outcomes. In the study by Budhiraja *et al.*,[17] more than 90% of survivors dominated the overall predictivity and ignored the poor predictivity (23.2%) for mortality by their model. A good model predicts both types of outcomes with equal dexterity.

A good alternative is, what we call, a predictivity based on the ROC curve[24] of the type shown in Figure 2 between

prevalence and predictivity for selected values of sensitivity and specificity. It is a straight line in the diagonal when $(Se + Sp) = 1$ because then PPV = (1 – NPV) = p but takes other shapes for other values of sensitivity and specificity. These curves help to read the values of PPV and (1 – NPV) for different rates of prevalence for given sensitivity and specificity when drawn in large size with an appropriate grid. Both PPV and NPV increase with prevalence but at a differential rate. The prediction models should report such predictivity-based ROC curves for the sensitivity-specificity observed in their study so that any user can read the values of PPV and NPV for the prevalence in his/her study. Sensitivity-specificity generally remains the same across populations but predictivities change according to the prevalence.

The procedure mentioned above will describe the overall performance of the model at different prevalences but will not give the optimal cut-off for maximum (PPV + NPV) in the study. Call this cut-off the predictivity (P) index on the pattern of the Youden index as proposed earlier. To obtain the P-index, use prevalence and Bayes Rule to convert sensitivity and specificity to PPV and NPV at each value of the test and pick up the threshold where (PPV + NPV) is the maximum. For a prospective study, PPV and NPV at different values of the test can also be directly obtained by counting the cases with positive outcomes with values more than or equal to each value, and by counting the cases with negative outcomes with values less than each value. In the course of time, when the method is widely adopted, statistical software will provide this information. Our experience with several datasets suggests that the value with the highest (PPV + NPV) would be an extreme value where PPV would be quite high and NPV very high. This happens because NPV has a lower bound $(1 - p)$ in practically all situations as explained earlier, and P is mostly small. The examples of SOFA score and CR*P* values in Table 2 illustrate this with a high value of 20.5 of SOFA score and a high value of 334.6 mcg/dL for CR*P* values as the cut-off for best prediction. Many would be surprised by these high values, but that is what is obtained for best predictivity.

In addition to the prospective studies, the above mentioned procedure would work for a cross-sectional study as well, provided the study is based on a random sample of all the subjects and the representation of the antecedents and outcomes is proportionate to their frequencies in the target population. However, the ROC curves are commonly used for logistic regression in a case-control setup, and the predictivities are elusive in this retrospective format. If prevalence is known, and the Bayes Rule can be used to report PPV and NPV in these

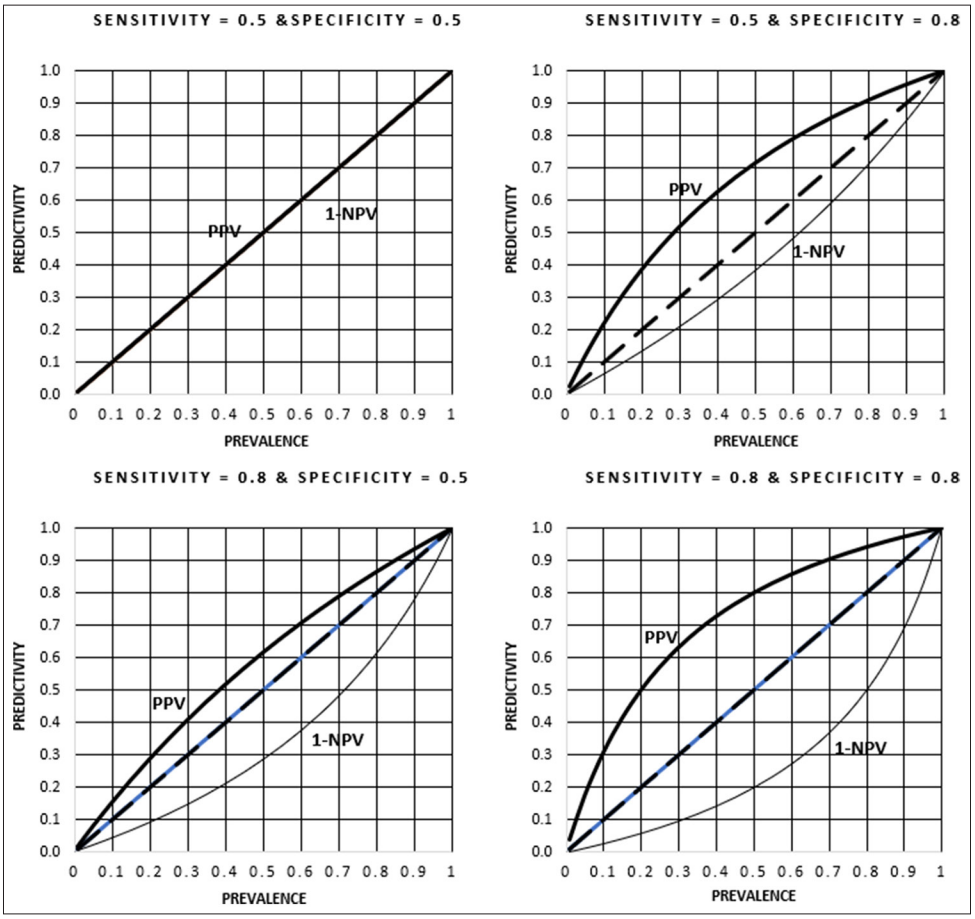

Figure 2: Predictivity-based ROC curves for some specific values of sensitivity and specificity

studies at the value with the highest Youden index, they must be accompanied by an exclusive statement that these are not necessarily the same as with the highest P-index.

Limitations

Sensitivity-specificity-based indices (ROC curve, C-statistics, and Youden index) have been extensively studied, and their statistical properties are known. C-statistic has interpretation in terms of rank correlation and Wilcoxon rank sum statistic, and it measures the discriminating ability of the logistic models. These indicators are mostly invariant to prevalence and have a global nature. However, they are good only for discriminating the known outcomes, and we need to realize that they are misleading indices of predictivity. Predictivities are local and based on the prevalence in that area. In this communication, we pointed out the fallacies commonly occurring from the use of ROC curves for assessing predictivity and proposed that predictivity should be assessed by PPV and NPV and the related indices. A limitation is that statistical features of the indices based on PPV and NPV, such as the best cut-off and P-index, are yet to be studied. They will be worked out once it is realized that PPV and NPV are the right indices for assessing predictivity and not sensitivity-specificity-based ROC curves. The predictivity indices are based on disease characteristics (prevalence) rather than test characteristics (sensitivity and specificity).

Conclusion

The area under the ROC curve and C-index for binary outcomes are based on already-known outcomes, and they provide misleading results regarding the predictivity of future outcomes because predictivity depends heavily on prevalence. Their use for assessing predictivity must be discontinued. Optimal cut-offs provided by the maximum Youden index also are not appropriate for predictivity for the same reason. Predictivity of a test or a model should be obtained in terms of the percentage of correctly predicted outcomes – overall as well as separately for positive and negative outcomes – preferably through a prospective study that moves from the antecedents to the outcomes so that the required future perspective for prediction is available. Prediction models should provide predictivity-based ROC curves between PPV and (1 – NPV) at different prevalences for proper inference in other setups where the prevalence is different. Optimal cut-offs for the prediction of outcomes too should be based on P-index, which is based on the highest PPV and NPV and not the sensitivity-specificity-based Youden index.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Steyerberg EW. Clinical Predictor Models: A Practical Approach. 2nd ed. Nature, Switzerland: Springer; 2019. p. 4.
2. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, *et al*. Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. BMJ 2020;369:m1328. Update in: BMJ 2021;372:n236. Erratum in: BMJ 2020;369:m2204.
3. Brenner H, Gefeller O. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Stat Med 1997;16:981-91.
4. Buscombe JR, Cwikla JB, Holloway B, Hilson AJ. Prediction of the usefulness of combined mammography and scintimammography in suspected primary breast cancer using ROC curves. J Nucl Med 2001;42:3-8.
5. Wei C, Lu Q. Collapsing ROC approach for risk prediction research on both common and rare variants. BMC Proc 2011;5(Suppl 9):S42.
6. Liu L, He YH, Li ZA, Zhang Y, Gu XY, Han JC, *et al*. Diagnostic value of an ROC curve of the size of the antepartum foramen ovale in the prediction of puerperal atrial septal defect. Exp Ther Med 2013;5:1501-5.
7. Gazi MA, Mahmud S, Fahim SM, Kibria MG, Palit P, Islam MR, *et al*. Functional prediction of hypothetical proteins from Shigella flexneri and validation of the predicted models by using ROC curve analysis. Genomics Inform 2018;16:e26.
8. Johnston AD, Lu J, Ru KL, Korbie D, Trau M. Primer ROC: Accurate condition-independent dimer prediction using ROC analysis. Sci Rep 2019;9:209.
9. Mei J, Hu W, Chen Q, Li C, Chen Z, Fan Y, *et al*. Development and external validation of a COVID-19 mortality risk prediction algorithm: A multicentre retrospective cohort study. BMJ Open 2020;10:e044028.
10. Wang Q, Cheng J, Shang J, Wang Y, Wan J, Yan YQ, *et al*. Clinical value of laboratory indicators for predicting disease progression and death in patients with COVID-19: A retrospective cohort study. BMJ Open 2021;11:e043790.
11. Hou N, Li M, He L, Xie B, Wang L, Zhang R, *et al*. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: A machine learning approach using XGboost. J Transl Med 2020;18:462.
12. Lin Q, Wang Y, Luo Y, Tang G, Li S, Zhang Y, *et al*. The effect of host immunity on predicting the mortality of Carbapenem-resistant organism infection. Front Cell Infect Microbiol 2020;10:480.
13. Hsu YT, He YT, Ting CK, Tsou MY, Tang GJ, Pu C. Administrative and claims data help predict patient mortality in intensive care units by logistic regression: A nationwide database study. Biomed Res Int 2020;2020:9076739.
14. Pan Z, Zhang R, Shen S, Lin Y, Zhang L, Wang X, *et al*. OWL: An optimized and independently validated machine learning prediction model for lung cancer screening based on the UK Biobank, PLCO, and NLST populations. eBioMedicine 2023;88:104443.
15. Adjahoto EO, Gogovor Y, Hodonou KA. Utilisation de la courbe ROC (receiver operating characteristic) dans la prédiction de l'hypotrophie foetale par la mesure de la hauteur utérine [Use of the ROC (receiver operating characteristic) curve in the prediction of fetal hypotrophy with uterine height measurement]. J Gynecol Obstet Biol Reprod (Paris) 1999;28:472-5. French.
16. Olivares-Morales A, Hatley OJ, Turner D, Galetin A, Aarons L, Rostami-Hodjegan A. The use of ROC analysis for the qualitative prediction of human oral bioavailability from animal data. Pharm Res 2014;31:720-30.
17. Budhiraja S, Indrayan A, Das P, Dewan A, Singh O, Nangia V, *et al*. Relative importance of various inflammatory markers and their critical thresholds for COVID-19 mortality. medRxiv 2021. Available from: https://www.medrxiv.org/content/10.1101/2021.12.24.21268371v1.full-text. [Last accessed: 2024 April 05].
18. Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, *et al*. Integrating the predictiveness of a marker with its performance as a classifier. Am J Epidemiol 2008;167:362-8.
19. Vickers AJ, Elkin EB. Decision curve analysis: A novel method for evaluating prediction models. Med Decis Making 2006;26:565-74.
20. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115:928-35.
21. Vetter TR, Schober P, Mascha EJ. Diagnostic testing and decision-making: Beauty is not just in the eye of the beholder. Anesth Analg 2018;127:1085-91.
22. Hong CS, Choi SY. Positive and negative predictive values by the TOC curve. Commun Stat Appl Methods 2020;27:211-24.
23. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, *et al.*; PROBAST Group. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019;170:51-8.
24. Indrayan A, Malhotra RK. Medical Biostatistics. 4th ed. Boca Raton, Florida: CRC Press; 2018. p. 274.