

COMMENTARY

The need for reorientation toward cost-effective prediction:
Comments on ‘Evaluating the added predictive ability of a new
marker: From area under the ROC curve to reclassification and
beyond’ by M. J. Pencina *et al.*, *Statistics in Medicine*
(DOI: 10.1002/sim.2929)

Sander Greenland^{*,†}

Departments of Epidemiology and Statistics, University of California, Los Angeles, CA 90095-1772, U.S.A.

Diagnostic and prognostic risk scores from predictive models are often used to help decide whether further procedures or special attention should be given to a patient. It seems natural to evaluate their performance in screening terms such as sensitivity (correct classification probability among cases or events), specificity (correct classification probability among noncases or nonevents), and predictive values (correct classification probabilities among test positives and test negatives). However, these simple intuitive performance criteria are highly correlated, and vary with the choice of score cutpoint.

A popular summary over criteria and cutpoint choices is the AUC, the area under the receiver-operating characteristic (ROC) curve, which plots sensitivity against the false-positive rate (1–specificity). As Pencina *et al.* note [1], AUC equals the probability that the underlying risk score will assign higher probability of being a case to a case than to a noncase. Thus, the change in this probability on changing the score (e.g. by adding a predictor) is $DUC = AUC_{\text{new}} - AUC_{\text{old}}$, where the subscripts ‘new’ and ‘old’ denote quantities after and before the score change. Since the DUC is a univariate summary of two ROC curves, it must sacrifice information. This information can be recaptured with other summaries, such as the net reclassification improvement (NRI) and integrated discrimination improvement (IDI) proposed by Pencina *et al.* [1].

I will concentrate on the DUC and IDI, although most of my remarks also apply to NRI and similar measures. Letting IS and IP be the sensitivity and false-positive rate averaged over all possible cutpoints, we have

$$IDI = (IS_{\text{new}} - IS_{\text{old}}) - (IP_{\text{new}} - IP_{\text{old}}) = (IS_{\text{new}} - IP_{\text{new}}) - (IS_{\text{old}} - IP_{\text{old}})$$

*Correspondence to: Sander Greenland, Departments of Epidemiology and Statistics, University of California, Los Angeles, CA 90095-1772, U.S.A.

†E-mail: lesdomes@ucla.edu

COMMENTARY

which is the change in the average difference of the ordinate and abscissa of points on the ROC curve. Thus, while IDI does provide interesting information beyond DUC, it does not satisfy requests to 'take us beyond the ROC curve' [2]. More generally, a hallmark of DUC, NRI, IDI, and many other measures is that they involve only probabilities conditional on the outcome variable (i.e. the event being predicted). I believe such measures are clinically far less relevant than often thought, because predictive values, costs, and cutpoints must be considered together to make well-informed decisions.

Pencina *et al.* note the importance of improvement in predictive values, and recommend examining 'meaningful' cutoffs. I will argue that cost considerations provide the meaning. The limitations of outcome-conditional, cost-free, cutpoint-free measures like DUC and IDI can then be summarized as follows:

1. If accurate prediction or classification is our goal, then predictive values are the key functions through which sensitivity and specificity become relevant. Functions of the ROC alone lack a key parameter that determines the relative importance of sensitivity and specificity in practice: The background probability (prevalence or incidence) B of the outcome event. This lack is a weakness, not a strength.
2. Costs are crucial. At the very least, we have the cost of measuring the new variable, plus two free parameters representing the net misclassification costs. Hypotheses of whether a variable improves prediction are of minor importance. Clinically relevant null hypotheses are concerned with whether the decline (if any) in expected classification cost offsets the cost of adding the variable to our model.
3. Once we consider the classification costs, most cutpoints would be out of the running immediately. DUC, IDI, or other measures that are cutpoint free are in effect incorporating irrelevant information, which degrades performance in the cost-laden reality of practice.
4. The decision will not always be about whether to add a predictor or not; it may sometimes be best to have new predictors displace old ones. Furthermore, a costly measurement may be ordered or not in response to a previous prediction. Thus, to better simulate and stimulate intelligent clinical practice, the notion of a single model can be replaced by that of a sequential decision algorithm: High-cost variables can be added conditional on the predictions obtained from low-cost variables. Again, DUC and IDI are not good measures for this purpose.
5. Decisions to add or displace variables will depend on the model and fitting strategy deployed. This means that we should expand our risk-prediction strategies beyond maximum-likelihood fitting of logistic and Cox models to strategies that provide more flexibility and better out-of-sample performance.

I will address each of these problems in turn.

OUTCOME-CONDITIONAL MEASURES

Pencina *et al.* state that 'a key measure of clinical utility of a survival model is its ability to discriminate or separate those who will develop the event of interest from those who will not.' Letting Y be the indicator of the event of interest, this statement recognizes that the clinical goal is to separate $Y = 1$ from $Y = 0$, *in advance of knowing* Y . Measures of clinical utility should thus capture the ability to predict Y conditional on available predictors, and not the ability to predict

COMMENTARY

classification given the as-yet unknown Y . For example, if Y^* is a model-based classification indicator (the output of a classification rule), high utility corresponds to having positive and negative predictive values, $PPV = \Pr(Y = 1|Y^* = 1)$ and $NPV = \Pr(Y = 0|Y^* = 0)$, that are close to one. For risk scores based on a covariate vector Z , high utility corresponds to accurately estimating the regression $E(Y|Z = z) = \Pr(Y = 1|Z = z)$.

DUC and IDI are based only on the probabilities of being classified correctly among those who did and those who did not develop the event, the sensitivity $Se = \Pr(Y^* = 1|Y = 1)$ and specificity $Sp = \Pr(Y^* = 0|Y = 0)$. Use of performance measures based on $\Pr(Y^* = y|Y = y)$ to evaluate clinical utility could be seen as a variant of the prosecutor's fallacy [3], which decides guilt based on sensitivity rather than on PPV. With background risk $B = \Pr(Y = 1)$, recall that the probabilities are related *via* Bayes theorem:

$$PPV = Se \cdot B / [Se \cdot B + (1 - Sp)(1 - B)] \quad \text{and} \quad NPV = Sp(1 - B) / [(1 - Se)B + Sp(1 - B)]$$

Elementary relations like these show that with low B , DUC and IDI can be higher for variables that provide large but clinically unimportant sensitivity improvements for a given specificity, relative to variables that provide small but clinically important specificity improvements for a given sensitivity. These defects and others stem from the absence of B in DUC and IDI.

Statistically, focusing on predictive values is simpler than focusing on ROC parameters, since those values are nothing more than classification probabilities derived from dichotomized risk scores (risk-model predictions). Even if one has information only on score sensitivity and specificity, there is usually extensive information on background risks, allowing predictive values to be estimated from Bayes theorem. Below, I will argue further that all that matters for clinical decisions are the risk scores and predictive values from a relatively narrow range of cutpoints. These conclusions follow immediately from cost (loss-function) considerations.

COSTS

Loss functions are an integral part of the foundation of frequentist as well as Bayesian decision theory, yet it seems accepted statistical practice to neglect them in formal methodology (notable exceptions occur in the data mining and statistical learning literature, e.g. Hastie *et al.* [4]). That is understandable, given the difficult nonstatistical issues in estimating costs. This difficulty should not, however, make one neglect a sound heuristic.

Any decision rule entails an implicit loss function, and the loss functions implicit in rules that appear to neglect loss functions are usually clinically absurd. One property of the loss function implicit in IDI is that it treats equal improvements in average sensitivity and average specificity as equally beneficial. This is apparent by rewriting the defining formula as $IDI = (IS_{\text{new}} - IS_{\text{old}}) + (IT_{\text{new}} - IT_{\text{old}})$, where $IT = 1 - IP$ is the integrated specificity. Given a nonzero cost of false positives, this equal weighting becomes even more absurd as the background rate becomes smaller. That can be seen by noting that as B approaches zero with fixed costs, false positives become the dominant cost factor, NPV approaches 1, and any change in sensitivity approaches worthlessness; in this limit, all that matters is specificity. (Such extreme situations happen, as in the extensive screening that was conducted for certain possibly nonexistent syndromes putatively caused by silicone implants.) The IDI loss function also becomes even more absurd as the ratio of false-positive to false-negative costs moves toward extremes.

Even in nonextreme situations, however, it would only be a numerical accident if equal changes in sensitivity and specificity represented equal cost changes. Suppose then that we would like to make good decisions about the clinical value of a new predictor X more often than accidentally. Pencina *et al.* suggested weighting the IDI and NRI to accommodate differences in the importance of sensitivity and specificity. However, such weighting would make sense only if it were based on costs and baseline risks. If we are going through the trouble of weighting, we may as well use the components directly in a loss function.

As an example, suppose C_X is the cost of measuring X , C_{tp} the cost of true positive, C_{fp} the cost of false positive, C_{fn} the cost of false negative, and C_{tn} the cost of true negative. Then, the expected loss from using the existing set of variables (without X) is

$$E_{old} = B[PPV_{old}C_{tp} + (1 - PPV_{old})C_{fn}] + (1 - B)[(1 - NPV_{old})C_{fp} + NPV_{old}C_{tn}]$$

Similarly, the expected loss on adding X to the available variables is

$$E_{new} = B[PPV_{new}C_{tp} + (1 - PPV_{new})C_{fn}] + (1 - B)[(1 - NPV_{new})C_{fp} + NPV_{new}C_{tn}]$$

The expected cost of adding X is then

$$\begin{aligned} \Delta &= C_X + E_{new} - E_{old} \\ &= C_X + B(PPV_{new} - PPV_{old})(C_{tp} - C_{fn}) + (1 - B)(NPV_{new} - NPV_{old})(C_{tn} - C_{fp}) \end{aligned}$$

This is a function of the two predictive values and three cost parameters (C_X , $C_{tp} - C_{fn}$, $C_{tn} - C_{fp}$). The ideal selection criterion is now easy to state: We should want to obtain X if adding X entails $\Delta < 0$ (i.e. benefit). The relevant null hypothesis and test is thus one-sided: $\Delta \geq 0$, i.e. X is not worth its measurement cost.

The test criterion Δ involves cost parameters that can be far beyond the scope of statistical expertise, involving matters of valuation and quality of life. It is then natural and may often suffice to focus statistical efforts on maximizing the accuracy of the risk score with and without X , to provide an accurate basis for further evaluations. Nonetheless, by including costs as free parameters in a loss function, a statistician can (with the aid of contextual experts) perform a sensitivity analysis over a range of reasonable values, rather than rely on potentially absurd implicit defaults. Occasionally, it may even be deemed worthwhile to statistically estimate costs as well as risks from available data, to provide a complete health-service evaluation.

Preliminary testing as a heuristic to avoid cost estimation

Testing whether X improves score performance is a good example of testing a null that is probably false, even if nonsignificant: Mostly any measure seriously proposed for routine use will likely have some incremental predictive ability. The only question is whether X is worth the trouble and expense to collect and use. Nonetheless, tests of whether X improves relevant predictive measures (predictive values or predictive accuracy) can serve as a heuristic first screen before evaluation of Δ : If the improvement offered by X cannot be shown to be nonzero, it cannot be shown to reduce expected loss, and so we need not go on to consider costs.

Note that the 'improvement' in the preceding statistical question is in the fitted values of the predictive probabilities $\Pr(Y = 1|Z = z)$, not in the ROC summaries. Thus, a test for improvement would be given by a test of the X coefficients in the risk model. In accepting this or any preliminary-testing heuristic, however, we should recognize its logical gaps. For example, as usual there is no

COMMENTARY

justification for capping the Type-I error rate at $\alpha = 0.05$. Because the actual hypothesis of interest is whether X would pass cost evaluation, we should want the α -level to vary inversely with the cost of measuring X .

Pencina *et al.* emphasize that if X does pass the initial 'null' screen, we must evaluate the clinical importance of its contribution. Difficult as it may seem, clinical importance can be realistically evaluated only with measures that (like Δ) show how costs change with predictive gains. DUC, IDI, and other outcome-conditional measures are unnecessary for both the initial screening and for evaluating the importance of adding X , because they neither measure predictive gain nor account for costs. When attention is turned to formal cost analysis, PPV and NPV do not complement sensitivity and specificity, but rather are the only conduit through which sensitivity and specificity enter the evaluation. In effect, then, PPV and NPV screen off sensitivity and specificity from further consideration.

CUTPOINTS

Pencina *et al.* provide sensible advice when they recommend 'calculating PPV and NPV for a set of meaningful cut offs.' A sensible meaning for 'meaningful' will, however, necessitate cost considerations.

Note first that predictive values (and hence E_{new} and E_{old}) depend on the cutpoints chosen, or, more generally, on how one constructs the decision (classification) rules with and without X . To make Δ meaningful, the decision rules must be constructed to minimize E_{new} and E_{old} . Using conventional classifiers, this minimization should include cutpoint optimization as well as model fit: We should want the cutpoint (or more generally, the clinical decision rule) that minimizes expected loss, given the costs. This cutpoint will vary with costs, as well as with the model. Furthermore, if the average severity of cases detected varies with cutpoint, the costs of errors will vary with cutpoint.

Measures that average over all possible cutpoints (like DUC and IDI) will incorporate extensive information that would be seen as irrelevant if the expected loss were considered. Measures that further ignore background prevalence are most sensitive to the irrelevancies, especially when the costs of errors vary with cutpoint. These facts can be seen in screening for a rare condition when the costs of false positives are not negligible (either individually or in aggregate) relative to the cost of false negatives.

Consider screening for prostate cancer *via* the prostate-specific antigen (PSA) test. One question might be whether it is worth ordering the test (adding the variable to a null set) after a negative routine clinical examination. The PSA test supplies a continuous measure; hence, a model that makes maximal use of the information will supply a predictive probability that can range over the whole unit (0, 1) interval. The individual cost of a false negative ranges widely, from nothing for those men with indolent tumors (for whom death will come from something else before they would develop symptoms) to several years of life lost. The proportion of false negatives with these outcomes may further depend on the cutpoint chosen. These possibilities must be weighed against the cost of a false positive, such as an unpleasant, expensive, and invasive biopsy.

As in other screening controversies (e.g. mammography for young women), informed observers do not care about performance at clinically ridiculous cutpoints, e.g. one that would achieve a sensitivity of 0.999 at the cost of a specificity of 0.100. With such a cutpoint, the vast majority of test positives could be false, incurring enormous unnecessary pain, complication, and expense.

COMMENTARY

Conversely, a cutpoint achieving a false-positive rate of 0.001 at the cost of a sensitivity of 0.100 would miss the vast majority of cases, including the worst ones.

When a score is to be used as a screening tool, there is no reason to consider cutpoints that will grossly inflate the loss, which is what the use of AUC or IDI entails. To be sure, there will be great uncertainty about the best cutpoint, but this uncertainty does not extend to the entire range of the score, and can be incorporated into the evaluation process. By including cutpoints as free parameters in the decision rule, one can at least perform a sensitivity analysis over a reasonable range, rather than giving absurd cutpoints the same weight as reasonable ones.

PURE PREDICTIVE SCORES

There are many situations in which a risk score will be supplied directly to the clinician as an aid to informal clinical judgment. Such use is common with the Framingham score. For such purposes no cutpoint is required, so a rationale for cutpoint-based performance measures (whether outcome conditioned or predictive) is needed.

Under the usual idealization of rational decisions, the clinician and patient would weigh the expected loss from various decisions, given that the patient's score (estimated risk) is R^* . In practice, however, R^* may simply lead to classification of the patient into risk categories, essentially locating R^* between cutpoints. Available variables not in the score would informally enter the decision in ways unforeseen by the score developer. Evaluating performance over a broad range of cutpoints may then seem relevant. Nonetheless, as mentioned above, the category cutpoints at issue are for a predictive score, performance should be measured by predictive values, not outcome-conditional parameters, and performance at absurd extremes would not be relevant.

Examining measures as simple as the change in difference between observed and predicted cases within categories of predicted risk can be illuminating as long as care is taken to adjust for the optimistic bias of naive estimates prediction error (e.g. *via* cross validation or bootstrapping). Many other measures of predictive improvement can be constructed by taking differences of familiar error measures or more general expected loss functions for models [4, 5]. These measures deserve exploration in the risk-scoring arena, with special attention given to estimating out-of-sample performance.

MODELING ISSUES

As does much of the literature, I have thus far assumed a very narrow context in which the only choice is to include or exclude a measurement X as a discriminator or predictor in a given form, such as a logistic or Cox model. However, that is not how we should build predictive models in general. Typically, there may be many options for the set of predictors. Furthermore, many models and fitting methods are available, the choice of which can strongly influence prediction performance.

Subset selection and sequential scoring

Even if adding X confers no benefit, by entering X we may be able to eliminate other predictors with little or no degradation in predictive values. If some of these old predictors are costly and need not be routinely collected, the best strategy may be to add X and drop other costly predictors. The statistical question then becomes: What is the optimal subset to use from a now-expanded set of

COMMENTARY

variables comprising X and the old predictors? Some theoretical solutions to this problem follow from comparing the expected costs of using different variable subsets. Many clinical predictors like age, sex, blood pressures, and BMI are omnipresent and hence should appear in every subset, leaving only a few costly variables to consider for entry or removal.

We can come closer to simulating and aiding intelligent clinical practice *via* sequential scoring. An initial risk score can be computed from the 'free' (routinely collected) variables. That score can then be used to decide whether a costly nonroutine measurement X is warranted. All the above considerations apply, in that we would not consider X for such a sequential rule unless we can demonstrate that this use would reduce the loss. Now, however, the threshold for the use of X is lowered, in that it would be measured (and hence incur a cost) only in a subset of patients, presumably those who would most benefit from the information it adds. When X is measured, an updated score is computed, which may be used for further decisions.

Fitting and model form

The decision to reject X may be strongly influenced by the way current and new markers are used to make predictions. Recognizing this possibility dictates use of the best available risk-prediction methods. As has been said, 'new risk predictors should routinely be assessed against the best available multivariable models because these are regarded as the standard of care' [2]. Simple conventional models and fitting methods, such as linear logistic and Cox models with maximum likelihood (whether partial or conditional), are no longer the best available modeling methods for all settings. Staying with them exclusively is akin to staying exclusively with medications introduced over 30 years ago, ignoring all subsequent research and development.

For example, maximum-likelihood regression has long been known to provide out-of-sample predictions much less accurate than those obtainable from more modern methods, such as shrinkage estimation (ridge, penalized-likelihood, empirical-Bayes, and related methods) [4–6]. Furthermore, a decision to reject X based on logistic or Cox modeling may be seriously in error if X is highly informative in a way not adequately captured by the model. This type of error may be avoided by exploiting flexible predictive forms [4].

Use of modern risk models need not create difficulty for the end user of the score. Indeed, in many applications the user enters variables in a computer interface that supplies the risk prediction, rather than hand-calculating the score from a formula. This type of interface eliminates the need to restrict models to a simple form.

PORTABILITY ARGUMENTS

A common reason given for focusing on ROC parameters such as AUC is that they are somehow an intrinsic property of the measurement technique, whereas the predictive values involve background risks that will inevitably vary across populations. Thus, it makes sense to focus on ROC parameters to ensure portability or generalizability of scores. One might argue that portability concerns are moot if the focus is on a single population of interest. However, each population is a mix of heterogeneous subpopulations, and will vary over time. So the portability argument would be important even when considering one population, if it were correct.

It has long been known, however, that ROC parameters can depend on characteristics of the patient mix, clinicians, and laboratory, which vary across populations and time [7]. In real

COMMENTARY

applications, ROC parameters will vary over time as the patient mix and laboratory and clinician errors change, just as background risks and costs will vary. Thus, ROC portability (constancy) is unlikely to hold over time, let alone across populations. Furthermore, successful models or scores will endure with time and spread to populations beyond the one in which they were developed. Even if ROC parameters are more stable than other parameters, this stability cannot compensate for their logical defects in performance evaluation. At the very least, changes in background risks and costs will necessitate reconsideration of the model and cutpoints to suit each new time period and population of application.

CONCLUSIONS

Evaluation of prediction models (risk scores) is complex. Measures of ROC change such as DUC and IDI provide simple and interesting performance indices. Nonetheless, because they ignore crucial free parameters, such measures are not suitable as the primary basis for evaluation and comparison of models.

All evaluation methods imply a loss function, which is driven by predictor-conditional (as opposed to outcome-conditional) performance, cutpoint choices, and error costs. Well-informed decisions (e.g. about whether a predictor is worth adding) will consider the form of this function and its inputs. The one input that is clearly within the statistician's scope is the prediction model. For costly but effective predictors, decisions (such as whether to add a variable) may be best left to the end user, in which case the statistician can help by offering a sequential scoring system. Regardless, it is the statistician's responsibility to deploy the best available methods to develop the model, just as it is the clinician's responsibility to deploy the best available modalities for patient care.

REFERENCES

1. Pencina MJ, D'Agostino Sr RB, D'Agostino Jr 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* 2007; DOI: 10.1002/sim.2929.
2. Greenland P, O'Malley PG. When is a new prediction marker useful? A consideration of lipoprotein-associated phospholipase A2 and C-reactive protein for stroke risk. *Archives of Internal Medicine* 2005; **165**:2454–2456.
3. Balding DJ. Interpreting DNA evidence: can probability theory help? In *Statistical Science in the Courtroom*, Chapter 3, Gastwirth JL (ed.). Springer: New York, 2000; 51–70.
4. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer: New York, 2001.
5. Harrell FE. *Regression Modeling Strategies*. Springer: New York, 2001.
6. Carlin B, Louis TA. *Bayes and Empirical-Bayes Methods of Data Analysis* (2nd edn). Chapman & Hall: New York, 2000.
7. Begg CB. Biases in the assessment of diagnostic tests. *Statistics in Medicine* 1987; **6**:411–419.