3-19-2013

The Net Reclassification Index (NRI): a Misleading Measure of Prediction Improvement with Miscalibrated or Overfit Models

Margaret Pepe
University of Washington, Fred Hutch Cancer Research Center, mspepe@u.washington.edu

Jin Fang
Fred Hutch Cancer Research Center, jfan@fhcrc.org

Ziding Feng
University of Washington & Fred Hutchinson Cancer Research Center, zfeng@fhcrc.org

Thomas Gerds
University of Copenhagen, tagteam@sund.ku.dk

Jorgen Hilden
University of Copenhagen, j.hilden@biostat.ku.dk

DRAFT UNDER REVIEW

1. Introduction

The Net Reclassification Index (NRI) was introduced in 2008 (Pencina *and others*, 2008) as a new statistic to measure the improvement in prediction performance gained by adding a marker, Y, to a set of baseline predictors, X, for predicting a binary outcome, D. The statistic has gained huge popularity in the applied biomedical literature. On March 13, 2013 through a search with Google Scholar we found 840 papers (44 since January 2012) that contained the acronym 'NRI' and referenced Pencina *and others* (2008). The measure has been extended from its original formulation (Pencina *and others*, 2010; Li *and others*, 2012). In this note we demonstrate a fundamental problem with use of the NRI in practice. We refer to work by Hilden and Gerds (2013) that provides insight into the source of the problems.

2. Illustration with Simulated Data

Consider a study that fits the baseline model $\text{risk}(X) = P(D = 1|X)$ and the expanded model $\text{risk}(X, Y) = P(D = 1|X, Y)$ using a training dataset. The fitted models that we denote by $\hat{\text{risk}}(X)$ and $\hat{\text{risk}}(X, Y)$ are then evaluated and compared in a test dataset. The continuous NRI statistic (Pencina *and others*, 2010) is calculated as

$$\text{NRI} = 2\{P[\hat{\text{risk}}(X, Y) > \hat{\text{risk}}(X)|D = 1] - P[\hat{\text{risk}}(X, Y) > \hat{\text{risk}}(X)|D = 0]\} \quad (2.1)$$

the proportion of cases in the test dataset for whom $\hat{\text{risk}}(X, Y) > \hat{\text{risk}}(X)$ minus the corresponding proportion of controls, multiplied by 2.

We generated data from a very simple simulation model described in the Supplementary Materials where X and Y are univariate and the logistic regression models hold:

$$\text{logit}P(D=1|X) = \alpha_0 + \alpha_1 X \tag{2.2}$$

$$\text{logit}P(D=1|X,Y) = \beta_0 + \beta_1 X + \beta_2 Y. \tag{2.3}$$

We used a small training set and fit logistic regression models of the correct forms in (2.2) and (2.3). Using a large test dataset we calculated the continuous NRI statistic for the training set derived models:

$$\text{logit } \hat{\text{risk}}(X) = \hat{\alpha}_0 + \hat{\alpha}_1 X$$

$$\text{logit } \hat{\text{risk}}(X,Y) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 Y.$$

The data were generated under the null scenario where Y does not add predictive information, i.e., $\beta_2 = 0$. The results in Table 1 indicate, however, that the NRI statistic is positive on average. That is, the NRI statistic calculated on the test dataset tends to indicate the erroneous result that Y contributes predictive information when in fact it does not.

We also calculated more traditional measures of performance improvement using their empirical estimates in the test dataset: $\Delta\text{AUC} = \text{AUC}(\hat{\text{risk}}(X,Y)) - \text{AUC}(\hat{\text{risk}}(X))$; $\Delta\text{ROC}(f) = \text{ROC}(f, \hat{\text{risk}}(X,Y)) - \text{ROC}(f, \hat{\text{risk}}(X))$; $\Delta S\text{NB}(t) = S\text{NB}(t, \hat{\text{risk}}(X,Y)) - S\text{NB}(t, \hat{\text{risk}}(X))$ and $\Delta\text{Brier} = \text{Brier}(\hat{\text{risk}}(X)) - \text{Brier}(\hat{\text{risk}}(X,Y))$ where

$$\text{AUC}(\text{risk}) = P(\text{risk}_i \geqslant \text{risk}_j | D_i = 1, D_j = 0)$$

$$\text{ROC}(f, \text{risk}) = P(\text{risk}_i \geqslant \tau(f) | D_i = 1) \text{ where } \tau(f) : P(\text{risk}_j \geqslant \tau(f) | D_j = 0) = f$$

$$S\text{NB}(t, \text{risk}) = P(\text{risk} > t | D = 1) - \frac{P(D=0)}{P(D=1)} \frac{t}{1-t} P(\text{risk} > t | D = 0)$$

$$\text{Brier}(\text{risk}) = E(D - \text{risk})^2.$$

The AUC is the area under the receiver operating characteristic (ROC) curve. The $\mathrm{ROC}(f, \mathrm{risk})$ measure is the proportion of cases classified as high risk when the high risk threshold is chosen as that exceeded by no more than a proportion f of controls. The standardized net benefit, $S\mathrm{NB}(t)$, is a weighted average of the true and false positive rates associated with use of the risk threshold t to classify subjects as high risk. This is a measure known by various names in the literature, including the decision curve (Vickers and Elkin, 2006) and the relative utility (Baker *and others*, 2009). The Brier score is a classic sum of squares measure. In simulation studies we set $f = 0.2$ for $\Delta\mathrm{ROC}(f)$ and $t = P[D = 1]$, the average risk, for $\Delta S\mathrm{NB}(t)$.

In contrast to the NRI statistic, we found that changes in the two ROC based measures, the standardized net benefit and the Brier score were negative on average in the test datasets, in all simulation scenarios (Table 1). Negative values for measures of performance improvement in the test dataset are appropriate because, given that Y is not predictive we expect that the fitted model $\hat{\mathrm{risk}}(X, Y)$ is further from the true risk, $P(D = 1|X)$, than is $\hat{\mathrm{risk}}(X)$. In particular, the model giving rise to $\hat{\mathrm{risk}}(X)$ requires estimating only 2 parameters and takes advantage of setting β_2 at its true value, $\beta_2 = 0$. In contrast, by fitting the three parameter model (2.3) that enters Y as a predictor, we incorporate noise and variability into $\hat{\mathrm{risk}}(X, Y)$. The ΔBrier score, $\Delta\mathrm{ROC}(f)$, ΔAUC and $\Delta S\mathrm{NB}(t)$ quantify the disimprovement in the performance of $\hat{\mathrm{risk}}(X, Y)$ relative to $\hat{\mathrm{risk}}(X)$ in different ways. In contrast, the NRI statistic tends to mislead us into thinking that the expanded model is an improvement over the baseline model.

3. Illustration with a Theoretical Example

Hilden and Gerds (2013) constructed some artificial examples of miscalibrated risk models and showed in simulation studies that the NRI statistic can be misleading. We now consider a simplified version of one of their examples and prove a theoretical large sample result. The example provides some insight into the simulation study results. Specifically, let Y be a constant, $Y = 0$

say, and consider a model $\text{risk}^*(X, Y)$ that is a distorted version of the true baseline risk function $\text{risk}(X)$ but that contains no additional predictive information:

$$\text{logit risk}^*(X, Y) = \text{logit risk}(X) + \varepsilon \text{ if risk}(X) > \rho \tag{3.4}$$

$$\text{logit risk}^*(X, Y) = \text{logit risk}(X) - \varepsilon \text{ if risk}(X) < \rho \tag{3.5}$$

where $\rho = P(D = 1)$. Result 1 below shows that the $NRI > 0$ for comparing the model $risk^*(X, Y)$ with the baseline model $\text{risk}(X)$. Here the training and test datasets are considered to be very large so there is no sampling variability, but the expanded model $\text{risk}^*(X, Y)$ is clearly miscalibrated while the baseline model is not.

Result 1

Assume that the baseline model is not the null model,

$$P(\text{risk}(X) \neq \rho) > 0.$$

Then $NRI > 0$ for the model (3.4)-(3.5).

Proof

Since the baseline model is well calibrated and

$$P(D = 1|\text{risk}(X) > \rho) = E(\text{risk}(X)|\text{risk}(X) > \rho),$$

we have

$$P(D = 1|\text{risk}(X) > \rho) = \rho + \delta \text{ for some } \delta > 0.$$

$$
\begin{aligned}
\text{NRI} =& 2\left\{P(\text{risk}^*(X,Y) > \text{risk}(X)|D=1) - P(\text{risk}^*(X,Y) > \text{risk}(X)|D=0)\right\} \\
=& 2\left\{\frac{P(D=1|\text{risk}(X)>\rho)P(\text{risk}(X)>\rho)}{P(D=1)} - \frac{P(D=0|\text{risk}(X)>\rho)P(\text{risk}(X)>\rho)}{P(D=0)}\right\} \\
=& 2P(\text{risk}(X)>\rho)\left\{\frac{\rho+\delta}{\rho} - \frac{1-\rho-\delta}{1-\rho}\right\} \\
=& 2P(\text{risk}(X)>\rho)\left\{\frac{\delta}{\rho} + \frac{\delta}{1-\rho}\right\} \\
=& \frac{2\delta P(\text{risk}(X)>\rho)}{\rho(1-\rho)} > 0
\end{aligned}
$$

■

We see that even in an infinitely large test dataset, the NRI associated with the expanded model in (3.4)-(3.5) is positive despite the fact that the expanded model contains no more predictive information than the baseline model. The integrated discrimination improvement (IDI) statistic was also proposed by Pencina *and others* (2008) and is quite widely used (Kerr *and others*, 2011). Hilden and Gerds (2013) proved that the IDI>0 for a different example of an uninformed expanded model.

4. Further Results

The expanded model in Result 1 is an extreme form of a miscalibrated model. Similarly, in the simulated data example, the expanded model derived from the small training dataset is likely to be miscalibrated in the test dataset. This is due to the phenomenon of overfitting. Miscalibration in the test dataset due to overfitting in the training set is likely to be exacerbated by inclusion of multiple novel markers in the expanded model. We see in Table 2 that the effects on NRI are more pronounced in the presence of multiple novel markers that are not predictive.

We next considered a scenario where a marker Y *does* add predictive information. The true expanded model in Table 3 is

$$\text{model}(X,Y) : \text{logit}P(D=1|X,Y) = \beta_0 + \beta_1 X + \beta_2 Y.$$

We fit this model and a model with superfluous interaction term to the training data

$$\text{model}(X, Y, XY) : \text{logit}P(D = 1|X, Y) = \gamma_0 + \gamma_1 X + \gamma_2 Y + \gamma_3 XY.$$

The test set NRIs comparing each of these fitted models with the fitted baseline model are summarized in Table 3. For comparison we display the NRI calculated using the true risk model parameter values. In some scenarios the NRI derived from the overfit model with interaction is substantially larger than the true NRI. For example, when $\text{AUC}_X = 0.9$ and $\text{AUC}_Y = 0.7$, the average NRI is 39.92% compared with the true NRI of 28.41%.

Considering the fact that the models fit to training data should be observed to perform *worse* than the true risk models, their tendency to appear better than the true risk models is particularly disconcerting. We see from Table 3 that the ROC based measures, the Brier Score and the net benefit all indicate that the performances of both of the expanded models fit to training data are worse than the performance of the true risk model. Moreover, as expected, the overfit model, model(X, Y, XY), is generally shown to have worse performance than the model without interaction. The NRI statistic however, only rarely conforms with this pattern. In five of the six scenarios considered, the NRI statistic for the overfit model(X, Y, XY) was larger than that for the model(X, Y). We conclude that miscalibration of the expanded model is problematic not only when the new marker is uninformative but also when the new marker is informative. In particular, overfitting can lead to inappropriately large values for the NRI in the test dataset.

5. Insights

Although we cannot fully explain why the NRI statistic tends to be large when the model for risk(X, Y) is overfit to training data, we can share a few relevant observations.

5.1 *NRI is not a Proper Measure of Performance Improvement*

Hilden and Gerds (2013) attribute the problem with the NRI statistic to the possibility that it is not based on a 'proper scoring rule.' See Gneiting and Raftery (2007) for an in-depth discussion of proper scoring rules.

In our context we need to expand on the definition of propriety. Let a population prediction performance improvement measure (PIM) comparing $r^*(X, Y)$, a function of (X, Y), to the true baseline risk function $r(X) = P(D = 1|X)$, be denoted by S:

$$\text{PIM} = S(r^*(X, Y), r(X), F(D, X, Y))$$

where F is the population distribution of (D, X, Y).

Definition

The PIM is *proper* if for all F and $r^*(X, Y)$:

$$S(r(X, Y), r(X), F(D, X, Y)) \geqslant S(r^*(X, Y), r(X), F(D, X, Y)). \tag{5.6}$$

In other words, a prediction improvement measure is proper if it is maximized at the true risk function of (X, Y), $r(X, Y) = P(D = 1|X, Y)$. If the inequality in (5.6) is strict, then $r(X, Y)$ is the unique function that maximizes S and the PIM is said to be *strictly proper*.

Propriety is generally considered a desirable attribute (Hilden and Gerds, 2013; Gneiting and Raftery, 2007). An unquestionably appealing attribute of a proper PIM is that improvement in performance cannot be due simply to disimprovement in calibration of the baseline model. Result 1 proves with a counter example that the NRI is not proper because $NRI > 0$ with use of the function $\text{risk}^*(X, Y)$ while $NRI = 0$ with use of the true risk function $\text{risk}(X, Y)$ that in this example is the same as $\text{risk}(X)$. On the other hand, it is well known from the theory of least squares that the change in the Brier score is proper, a fact that follows from the equality $E(D|X, Y) = \text{risk}(X, Y)$. In addition, the ΔAUC and $\Delta\text{ROC}(f)$ measures are proper since the ROC curve for (X, Y) is maximized at all points by the risk function (McIntosh and Pepe, 2002). Interestingly, these are not strictly proper measures because ROC curves are also maximized by any monotone increasing function of the risk. We show in supplementary materials that the change in the standardized net benefit, $\Delta S\text{NB}(t)$, is proper. Being proper measures of prediction improvement appears to translate into more sensible comparisons of risk models in our simulation studies. In particular, distortion of the baseline model by adding unnecessary predictors to the model does not increase the estimated values of the proper performance measures but can increase the NRI.

5.2 *Manifestations of Overfitting*

When risk models include superfluous predictor variables, predictions are apt to predict more poorly than predictions derived from models without them. In Figure 1 we demonstrate this for one simulated dataset corresponding to the scenario in the second-to-last row of Table 1. Observe that the predictions from the baseline model, $\hat{\text{risk}}(X)$, are seen to be closer to the true risk, $\text{risk}(X)$, than are the more variable predictions based on $\hat{\text{risk}}(X, Y)$, where Y is an uninformative variable that is therefore superfluous. The NRI statistic does not acknowledge the poorer predictions while the other performance improvement measures do (Figure 1 caption).

We compared the estimated odds ratios for X in the overfit models described in Table 2 with that in the fitted baseline model. Results shown in Table 4 indicate that odds ratios are biased too large in the overfit models. This presumably provides some rationale for use of shrinkage techniques to address problems with overfitting (Hastie *and others* (2001), section 10.12.1). Interestingly, when the odds ratio for X is larger in an overfit model than in the baseline model, the NRI statistic is generally positive (Figure 2). We now provide some intuition for this observation. Note that the NRI statistic compares $\hat{\text{risk}}(X, Y)$ with $\hat{\text{risk}}(X)$ for each individual. Assuming that the intercept terms center X at 0 but otherwise can be ignored, the NRI statistic adds positive contributions when

$$\hat{\beta}_1 X + \hat{\beta}_2 Y > \hat{\alpha}_1 X \text{ and } D = 1 \tag{5.7}$$

and when

$$\hat{\beta}_1 X + \hat{\beta}_2 Y < \hat{\alpha}_1 X \text{ and } D = 0. \tag{5.8}$$

But since X is large (positive) in cases and small (negative) in controls, the inequalities (5.7) and (5.8) tend to hold because of the tendency for $\hat{\beta}_1 > \hat{\alpha}_1$. Note that Y is centered at 0 and $\hat{\beta}_2$ is likely to be small in the simulations because $\beta_2 = 0$. In the simulation scenario corresponding to Figure 2 (and the second-to-last rows of Tables 1 and 4) we found that $\hat{\beta}_1 > \hat{\alpha}_1$ in 66.4% of simulated datasets leading to NRI>0 in 66.9% of datasets and an average NRI of 6.56% (Table 1). In supplementary materials (figure A.1-A.4) we see that for the same scenario, the ΔROC(0.2), ΔSNB(ρ), ΔAUC and ΔBrier statistics were generally negative regardless of the comparative values of $\hat{\alpha}_1$ and $\hat{\beta}_1$.

6. Discussion

The simulation and theoretical results provided here and in Hilden and Gerds (2013) demonstrate that the NRI statistic can be biased large by use of miscalibrated risk models. Of particular concern is the fact that models overfit to training data can appear to improve prediction performance when evaluated with test data even when they do not actually improve prediction performance. Even small amounts of overfitting, by adding a single unnecessary predictor, lead to biased test set evaluations in our simulation studies. Following the same logic, this sort of bias is likely also to be manifested in cross-validated estimates of the NRI although we have not specifically investigated the phenomenon here.

In practice, one should not use the NRI statistic, or other prediction improvement performance measures such as ΔAUC, ΔROC(f), ΔSNB(t), or ΔBrier for that matter, to determine *if* there is predictive information in a novel marker Y. We have argued previously that testing the null hypothesis about the regression coefficient associated with Y in the risk model risk(X, Y) is equivalent and more powerful to tests based on estimates of prediction improvement performance measures (Pepe *and others*, 2013).

On the other hand, for *quantifying* the improvement in prediction performance one must choose summary statistics. There has been much debate in the literature about which measures are most appropriate (Pepe and Janes, 2013). Arguments have centered on the interpretations and clinical relevance of various measures. The results in this paper and in Hilden and Gerds (2013) add another dimension to the debate. Regardless of the intuitive appeal that the NRI statistic may garner, its potential for being inflated by miscalibrated or over-fit models is a very serious concern.

Our results underscore the need to check model calibration as a crucial part of the exercise of evaluating risk prediction models. In additional simulation studies (results not shown) we found that after recalibrating the training set models to the test dataset, problems with inflated NRIs were much reduced. However, guaranteeing well fitting risk models in practical applications is not always possible. For this reason, other statistical measures of prediction improvement that cannot be made large by miscalibration may be preferred for practical application. We especially encourage use of the change in the standardized net benefit statistic and its components, the changes in true and false positive rates, calculated at a relevant risk threshold, because, not only is it a proper prediction improvement statistic, but it also quantifies prediction performance in a clinically meaningful way (Pepe and Janes, 2013; Vickers and Elkin, 2006; Baker *and others*, 2009).

Acknowledgments

This work was supported in part by NIH grants R01 GM054438, U24 CA086368, and R01 CA152089.

Conflict of Interest: None declared.

Supplementary Material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

References

Baker, Stuart G., Cook, Nancy R., Vickers, Andrew and Kramer, Barnett S. (2009). Using relative utility curves to evaluate risk prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **172**(4), 729–748.

Gneiting, Tilmann and Raftery, Adrian E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378.

Hastie, Trevor, Tibshirani, Robert and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag.

Hilden, Jorgen and Gerds, Thomas A. (2013). Evaluating the impact of novel biomarkers: Do not rely on IDI and NRI. *Statistics in Medicine* (in press).

Kerr, Kathleen F., McClelland, Robyn L., Brown, Elizabeth R. and Lumley, Thomas. (2011). Evaluating the incremental value of new biomarkers with integrated discrimination improvement. *American Journal of Epidemiology* **174**(3), 364–374.

Li, Jialiang, Jiang, Binyan and Fine, Jason P. (2012, Nov). Multicategory reclassification statistics for assessing improvements in diagnostic accuracy. *Biostatistics* (epub ahead of print).

McIntosh, Martin W. and Pepe, Margaret Sullivan. (2002, Sep). Combining several screening tests: optimality of the risk score. *Biometrics* **58**(3), 657–664.

Pencina, M.J., D'Agostino, R.B., D'Agostino, R.B. and Vasan, R.S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine* **27**(2), 157–172. PMID: 17569110 PMCID: N/A; precedes mandate.

Pencina, Michael J., D'Agostino, Ralph B. and Vasan, Ramachandran S. (2010, Dec).

Statistical methods for assessment of added usefulness of new biomarkers. *Clin Chem Lab Med* **48**(12), 1703–1711.

Pepe, M.S. and Janes, H. (2013). Methods for Evaluating Prediction Performance of Biomarkers and Tests . In *Risk Assessment and Evaluation of Predictions*, Springer.

Pepe, MS, Kerr, KF, Longton, G and Wang, Z. (2013). Testing for improvement in prediction model performance. *Statistics in Medicine*, Epub ahead of print. doi: 10.1002/sim.5727.

Vickers, A.J. and Elkin, E.B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* **26**(6), 565.

Tables and Figures

Table 1. Measures of improvement in prediction when risk models, $\text{logit}P(D = 1|X) = \alpha_0 + \alpha_1 X$ and $\text{logit}P(D = 1|X, Y) = \beta_0 + \beta_1 X + \beta_2 Y$, are fit to a training dataset and applied to a test dataset. The novel marker Y does not improve prediction: the true models are linear logistic (2.2) and (2.3) with coefficients: $\beta_2 = 0$, $\beta_1 = \alpha_1$ and $\beta_0 = \alpha_0$. Performance measures are averaged over 1000 simulations. Data generation is described in Supplementary Materials

One Uninformative Marker

Simulation Scenario				Performance Increment ×100 Average (standard error)				
$\rho = P(D = 1)$	AUC_X[1]	N-training[2]	N-test[3]	NRI	ΔROC(0.2)	ΔAUC	ΔBrier	ΔSNB(ρ)
0.1	0.6	250	25,000	0.27(0.09)	−1.70(0.09)	−1.28(0.08)	−0.044(0.002)	−1.85(0.12)
0.1	0.7	250	25,000	1.38(0.16)	−1.37(0.07)	−0.86(0.05)	−0.049(0.002)	−1.31(0.07)
0.1	0.8	250	25,000	3.22(0.28)	−0.90(0.05)	−0.48(0.02)	−0.058(0.003)	−0.80(0.04)
0.1	0.9	250	25,000	7.72(0.52)	−0.52(0.03)	−0.25(0.01)	−0.066(0.003)	−0.57(0.03)
0.5	0.6	50	5,000	0.57(0.15)	−1.67(0.12)	−1.19(0.11)	−0.479(0.023)	−1.69(0.15)
0.5	0.7	50	5,000	2.78(0.28)	−2.59(0.12)	−1.69(0.08)	−0.540(0.024)	−2.49(0.13)
0.5	0.8	50	5,000	6.56(0.47)	−1.83(0.09)	−1.00(0.05)	−0.492(0.022)	−1.62(0.08)
0.5	0.9	50	5,000	17.09(0.91)	−1.11(0.05)	−0.56(0.03)	−0.433(0.021)	−1.17(0.06)

[1]Area under the ROC curve for the baseline model (X)

[2]Size of training dataset

[3]Size of test dataset

Table 2. Measures of improvement in prediction $\times 100$ when risk models are fit to a training dataset of 50 observations and assessed on a test dataset of 5000 observations where $P(D=1)=0.5$. The linear logistic regression models fit to the training data are (i) baseline $\text{logit}P(D=1|X)=\alpha_0+\alpha_1 X$; (ii) model(X,Y_1) : $\text{logit}P(D=1|X,Y_1)=\beta_0+\beta_1 X+\beta_2 Y_1$; (iii) model$(X,Y_1,Y_2)$: $\text{logit}P(D=1|X,Y_1,Y_2)=\gamma_0+\gamma_1 X+\gamma_2 Y_1+\gamma_3 Y_2$. Data generation is described in Supplementary Materials. Shown are averages over 1000 simulations. Neither Y_1 nor Y_2 are informative — the true values of β_2, γ_2, and γ_3 are zero.

Two Uninformative Markers

	NRI		ΔROC(0.2)		ΔAUC		Δ Brier		ΔSNB(ρ)	
Model	(X,Y_1)	(X,Y_1,Y_2)	(X,Y_1)	(X,Y_1,Y_2)	(X,Y_1)	(X,Y_1,Y_2)	(X,Y_1)	(X,Y_1,Y_2)	(X,Y_1)	(X,Y_1,Y_2)
AUC_X										
0.60	0.61	0.78	−1.81	−2.91	−1.30	−2.18	−0.55	−1.12	−1.84	−3.06
0.70	2.08	3.63	−2.63	−4.55	−1.66	−2.95	−0.52	−1.08	−2.44	−4.36
0.80	6.18	10.60	−1.75	−3.50	−0.95	−1.89	−0.47	−0.96	−1.61	−3.14
0.90	17.83	28.00	−1.33	−2.57	−0.65	−1.27	−0.51	−1.03	−1.36	−2.63

Table 3. Measures of improvement in prediction when risk models are fit to a training dataset of N-training=50 observations and assessed on a test dataset of N-test=5000 observations where $P(D=1)=0.5$. The linear logistic regression models fit to the training datasets are (i) baseline $\text{logit}P(D=1|X)=\alpha_0+\alpha_1 X$; (ii) model(X,Y) : $\text{logit}P(D=1|X,Y)=\beta_0+\beta_1 X+\beta_2 Y$; (iii) model$(X,Y,XY)$: $\text{logit}P(D=1|X,Y)=\gamma_0+\gamma_1 X+\gamma_2 Y+\gamma_3 XY$. The marker Y is informative and the correct model is model(X,Y) while the model(X,Y,XY) is overfit. Data generation is described in Supplementary Materials. Shown are the true values of the performance measures (True) that use the true risk, as well as averages of estimated performance using training and test data over 1,000 simulations. All measures shown as %.

One Informative Marker

		NRI			ΔROC(0.2)			ΔAUC			ΔBrier			$\Delta SNB(\rho)$		
AUC_X	AUC_Y	True	Model (X,Y)	Model (X,Y,XY)	True	Model (X,Y)	Model (X,Y,XY)	True	Model (X,Y)	Model (X,Y,XY)	True	Model (X,Y)	Model (X,Y,XY)	True	Model (X,Y)	Model (X,Y,XY)
0.7	0.7	28.32	23.37	22.74	3.25	0.99	−0.89	1.98	0.64	−0.69	0.61	0.16	−0.41	3.21	0.98	−0.40
0.8	0.7	28.39	23.37	27.03	1.97	0.14	−1.13	1.03	0.06	−0.80	0.47	0.01	−0.45	1.84	0.15	−0.90
0.8	0.8	57.84	55.65	55.82	7.73	6.09	4.99	3.94	3.12	2.25	1.89	1.48	1.00	7.10	5.39	4.59
0.9	0.7	28.41	27.16	39.92	0.88	− 0.31	−1.15	0.43	−0.16	−0.77	0.29	−0.16	−0.59	0.92	−0.33	−1.25
0.9	0.8	57.87	57.13	63.05	3.40	2.31	1.61	1.69	1.14	0.53	1.19	0.74	0.30	3.79	2.49	1.71
0.9	0.9	89.58	87.02	88.18	7.35	6.46	5.59	3.74	3.23	2.54	2.81	2.40	1.95	8.85	7.50	6.74

Table 4. Average estimated odds ratios for X in models fit to training data generated using the same settings as in Table 2. Both Y_1 and Y_2 are uninformative markers.

AUC_X	True $\exp(\alpha_1)$	Baseline Model $\exp(\hat{\alpha}_1)$	Model (X, Y_1) $\exp(\hat{\beta}_1)$	Model (X, Y_1, Y_2) $\exp(\hat{\gamma}_1)$
0.6	1.43	1.56	1.58	1.60
0.7	2.10	2.46	2.55	2.63
0.8	3.29	4.02	4.24	4.49
0.9	6.13	6.78*	7.37*	8.23*

*medians displayed when distribution is highly skewed

Fig. 1. Risk estimates calculated, using models fit to training data, on a test data set of 5000 observations. Data were simulated from the scenario shown in the second to last row of Table 1. Values of performance improvement statistics are: NRI= 33.95%, ΔROC(0.2) = −1.65%, ΔAUC= −0.37%, ΔBrier= −0.54% and ΔSNB(ρ) = −1.77%. The thick curve is $\hat{\text{risk}}(X)$ while the thin curve is a lowess curve through the points $\hat{\text{risk}}(X, Y)$ that are displaed with light circles.

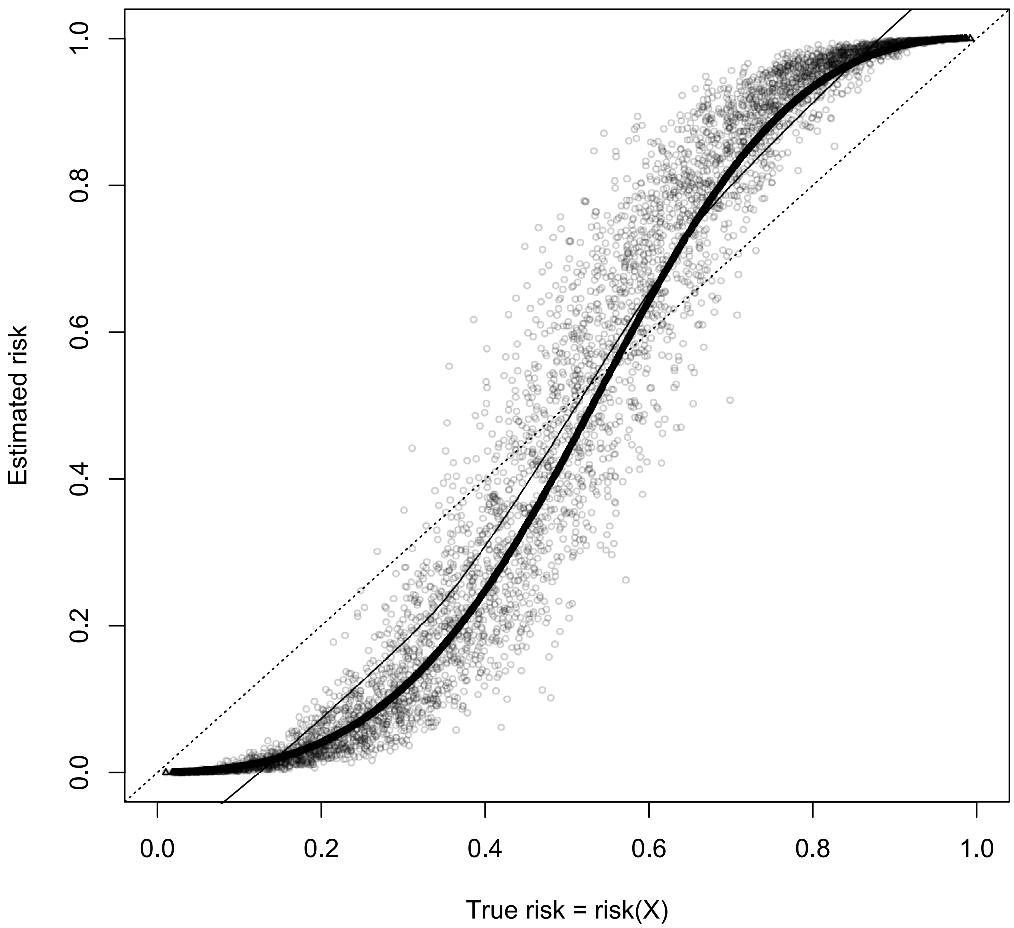
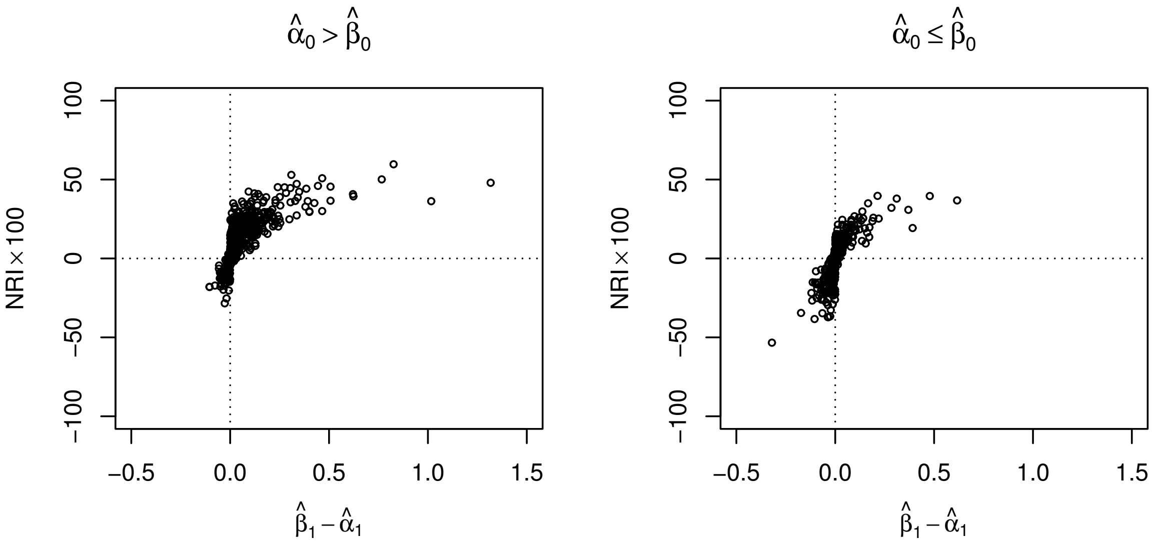


Fig. 2. Scatterplots showing the relationship between the NRI statistic ($\times 100$) and $\hat{\beta}_1 - \hat{\alpha}_1$ in 1000 simulated datasets generated according to the scenario shown in the second to last row of Table 1. The coefficients are calculated by fitting the models $\text{logit}P(D=1|X) = \alpha_0 + \alpha_1 X$ and $\text{logit}P(D=1|X,Y) = \beta_0 + \beta_1 X + \beta_2 Y$ to the training data. The NRI is calculated using the test dataset.



SUPPLEMENTARY MATERIALS

A.1 Data Generation

The simulations reported in Table 1 are based on data generated as

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X D \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

where D is random binary variable with specified $P(D=1)$. We chose $\mu_X = \sqrt{2}\ \Phi^{-1}(\text{AUC}_X)$ so that the AUC for the baseline model was fixed by design.

Data for Table 2 were simulated as

$$\begin{pmatrix} X \\ Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X D \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right).$$

Data for Table 3 were generated from the distribution

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_X D \\ \mu_Y D \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

where $\mu_X = \sqrt{2}\Phi^{-1}(\text{AUC}_X)$ and $\mu_Y = \sqrt{2}\Phi^{-1}(\text{AUC}_Y)$.

A.2 Net Benefit is a Proper Scoring Statistic

Let X denote the available predictors and let $r^*(X)$ and $r(X)$ be two functions of X. Assume that the true model is $P(D=1|X) = r(X)$ and the aim is to assess the decision rule based on the model $r^*(X)$ defined by $r^*(X) \geqslant t$. The standardized net benefit statistic at threshold t associated with the function $r^*(X)$ is

$$\text{SNB}(t, r^*(X)) = P(r^*(X) \geqslant t|D=1) - \frac{1-\rho}{\rho}\frac{t}{1-t}P(r^*(X) \geqslant t|D=0)$$

where $\rho = P(D = 1)$. In terms of expectation over the marginal distribution of X, we write

$$\begin{aligned}
\rho S\text{NB}(t, r^*(X)) &= E\left\{I(r^*(X) \geqslant t, D = 1) - \frac{t}{1-t} I(r^*(X) \geqslant t, D = 0)\right\} \\
&= E\left\{P(D = 1|X)I(r^*(X) \geqslant t) - \frac{t}{1-t}[1 - P(D = 1|X)]I(r^*(X) \geqslant t)\right\} \\
&= E\left\{I(r^*(X) \geqslant t)[r(X) - \frac{t}{1-t}(1 - r(X))]\right\} \\
&= E\left\{I(r^*(X) \geqslant t)[\frac{r(X) - t}{1-t}]\right\}.
\end{aligned}$$

Now consider the difference between $\rho S\text{NB}(t, r(X))$ and $\rho S\text{NB}(t, r^*(X))$: $\rho(S\text{NB}(t, r(X)) - S\text{NB}(t, r^*(X))) = \frac{1}{1-t} E\{(r(X) - t)(I(r(X) \geqslant t) - I(r^*(X)) \geqslant t))\}$

The entity inside the expectation, namely

$$(r(X) - t)\{I(r(X) \geqslant t) - I(r^*(X) \geqslant t)\} \tag{A.1}$$

is non-negative with probability 1. To see this, consider the various cases possible: (i) $r(X) = t$; (ii) $r(X) > t, r^*(X) \geqslant t$; (iii) $r(X) > t, r^*(X) < t$ (iv) $r(X) < t, r^*(X) < t$ and (v) $r(X) < t, r^*(X) \geqslant t$ and observe that A.1 is $\geqslant 0$ in each case. Therefore the expectation is nonnegative. In other words,

$$S\text{NB}(t, r(X))) \geqslant S\text{NB}(t, r^*(X))$$

for every function $r^*(X)$. That is, $S\text{NB}(t, r^*(X))$ is maximized by $r^*(X) = r(X)$, the true risk function.

Figure A.1. Scatterplots showing the relationship between the ΔROC(0.2) statistic ($\times 100$) and $\hat{\beta}_1 - \hat{\alpha}_1$ in 1000 simulated datasets generated according to the scenario shown in the second to last row of Table 1. The coefficients are calculated by fitting the models logit$P(D=1|X) = \alpha_0 + \alpha_1 X$ and logit$P(D=1|X,Y) = \beta_0 + \beta_1 X + \beta_2 Y$ to the training data. The ΔROC(0.2) is calculated using the test dataset.

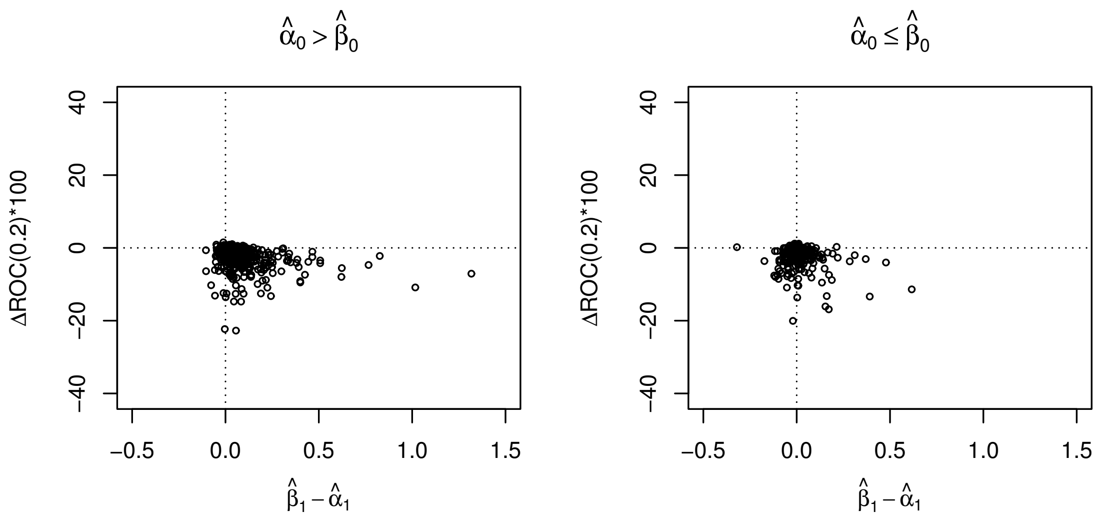


Figure A.2. Scatterplots showing the relationship between the $\Delta S\text{NB}(\rho)$ statistic ($\times 100$) and $\hat{\beta}_1 - \hat{\alpha}_1$ in 1000 simulated datasets generated according to the scenario shown in the second to last row of Table 1. The coefficients are calculated by fitting the models $\text{logit}P(D = 1|X) = \alpha_0 + \alpha_1 X$ and $\text{logit}P(D = 1|X, Y) = \beta_0 + \beta_1 X + \beta_2 Y$ to the training data. The $\Delta S\text{NB}(\rho)$ is calculated using the test dataset.

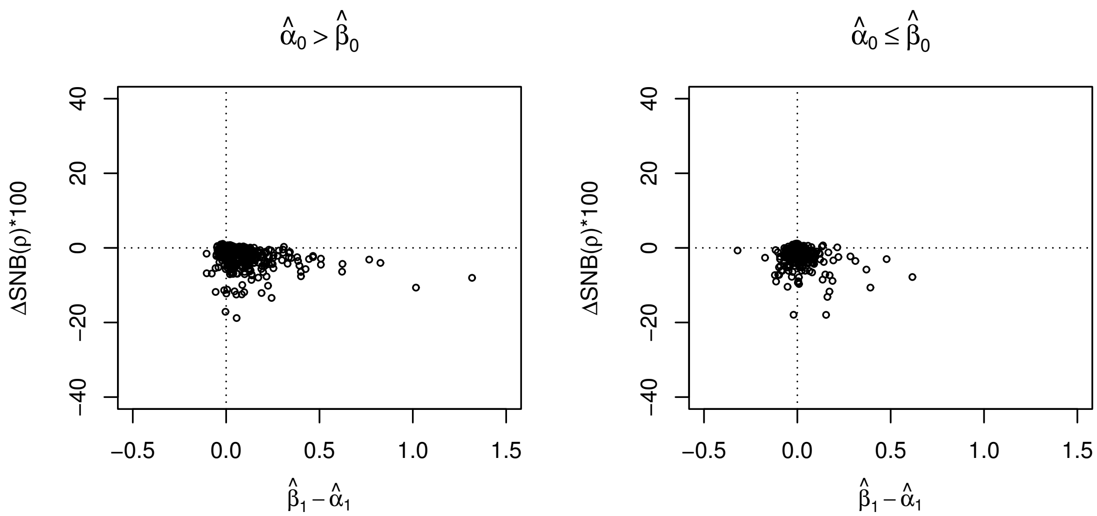


Figure A.3. Scatterplots showing the relationship between the ΔAUC statistic ($\times 100$) and $\hat{\beta}_1 - \hat{\alpha}_1$ in 1000 simulated datasets generated according to the scenario shown in the second to last row of Table 1. The coefficients are calculated by fitting the models logit$P(D = 1|X) = \alpha_0 + \alpha_1 X$ and logit$P(D = 1|X, Y) = \beta_0 + \beta_1 X + \beta_2 Y$ to the training data. The ΔAUC is calculated using the test dataset.

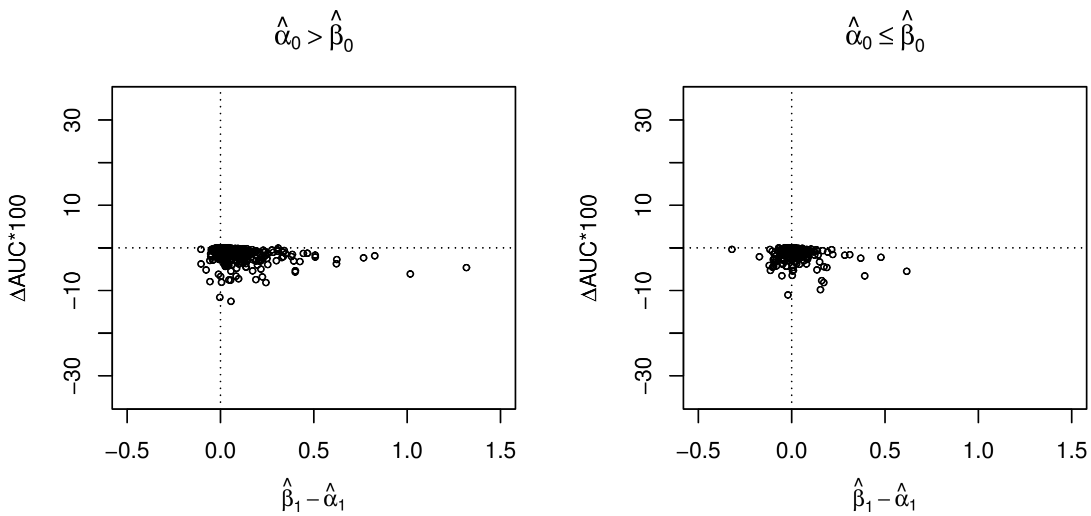


Figure A.4. Scatterplots showing the relationship between the ΔBrier statistic (×100) and $\hat{\beta}_1 - \hat{\alpha}_1$ in 1000 simulated datasets generated according to the scenario shown in the second to last row of Table 1. The coefficients are calculated by fitting the models $\text{logit}P(D=1|X) = \alpha_0 + \alpha_1 X$ and $\text{logit}P(D=1|X,Y) = \beta_0 + \beta_1 X + \beta_2 Y$ to the training data. The ΔBrier is calculated using the test dataset.