A 2-category NRI and Net Benefit

For a single risk model, let B to be the benefit of identifying an event as high risk and C as the cost of identifying a nonevent as high risk. Define the Net Benefit (3) of a risk model as

$$NB = B \cdot P(event)P(high|event) - C \cdot P(nonevent)P(high|nonevent). \quad (1)$$

Now, suppose we have "old" and "new" risk models, where the new model adds an additional marker to the old model. It is natural to quantify the incremental value of the new marker as ΔNB, the change in the Net Benefit by using the new marker for prediction. Let $high_n$ and $high_o$ denote that a subject is in the high risk category according to the new and old risk models, respectively. Then

$$\begin{aligned} \Delta NB &= B \cdot P(event)[P(high_n|event) - P(high_o|event)] \\ &\quad -C \cdot P(nonevent)[P(high_n|nonevent) - P(high_o|nonevent)]. \end{aligned} \quad (2)$$

For any individual, considering the old and new risk models there are four cases: the individual can be classified low risk by both models, high risk by both models, low and then high, or high and then low. Let ll, hh, lh, hl denote these four cases, where the first position is for the old risk model and the second position is for the new risk model. Then we can write the first line of (2) as

$$\begin{aligned} & B \cdot P(event)[P(hh|event) + P(lh|event) - P(hh|event) - P(hl|event)] \\ = \; & B \cdot P(event)[P(lh|event) - P(hl|event)] \\ = \; & B \cdot P(event)[P(up|event) - P(down|event)] \end{aligned} \quad (3)$$

Similarly, the second line of (2) can be written

$$-C \cdot P(nonevent)[P(up|nonevent) - P(down|nonevent)]. \quad (4)$$

Therefore,

$$\begin{aligned} \Delta NB &= B \cdot P(event)[P(up|event) - P(down|event)] \\ &\quad -C \cdot P(nonevent)[P(up|nonevent) - P(down|nonevent)] \\ &= B \cdot P(event)[\frac{P(event|up)P(up)}{P(event)} - \frac{P(event|down)P(down)}{P(event)}] \\ &\quad -C \cdot P(nonevent)[\frac{P(nonevent|up)P(up)}{P(nonevent)} - \frac{P(nonevent|down)P(down)}{P(nonevent)}] \\ &= B[P(event|up)P(up) - P(event|down)P(down)] \\ &\quad -C[P(nonevent|up)P(up) - P(nonevent|down)P(down)] \end{aligned} \quad (5)$$

Thus the $wNRI$ is exactly the change in the Net Benefit for the old and new risk models.

B 3-category NRI and Net Benefit

First, we generalize the definition of the 3-category NRI by considering the different ways individuals can move between risk categories. Second, we define Net Benefit for a risk model when there are three categories and derive ΔNB for the prediction increment. Last, we derive $wNRI$ for the 3-category NRI similar to the derivation of the $wNRI$ for two-categories in Pencina et al. (4). We we show that $wNRI$ for three categories is the same as ΔNB, just as they are equal for two categories.

B.1 Generalized NRI for 3 categories

The definition of the NRI is

$$NRI = P(up|event) - P(down|event) + P(down|nonevent) - P(up|nonevent). \tag{6}$$

For three categories, "up" can mean three things: move from low to medium, from medium to high, or from low to high. Let l, m, and h represent the low, medium, and high categories. For 3 categories we can write the NRI as

$$\begin{aligned}
& P(lm|event) + P(lh|event) + P(mh|event) \\
- \; & P(ml|event) - P(hl|event) - P(hm|event) \\
+ \; & P(ml|nonev) + P(hl|nonev) + P(hm|nonev) \\
- \; & P(lm|nonev) + P(lh|nonev) + P(mh|nonev) \qquad (7) \\
= \; & [P(event|lm)P(lm) + P(event|lh)P(lh) + P(event|mh)P(mh)]/P(event) \\
- \; & [P(event|ml)P(ml) + P(event|hl)P(hl) + P(event|hm)P(hm)]/P(event) \\
+ \; & [P(nonev|ml)P(ml) + P(nonev|hl)P(hl) + P(nonev|hm)P(hm)]/P(nonev) \\
- \; & [P(nonev|lm)P(lm) + P(nonev|lh)P(lh) + P(nonev|mh)P(mh)]/P(nonev) \qquad (8)
\end{aligned}$$

This is a linear combination of $P(event|*)P(*)$ and $P(nonev|*)P(*)$ where $*$ represents movement between risk categories.

B.2 Net Benefit and Three categories

Let B_h and B_m be the benefits for assigning a case to the high and medium risk categories, respectively. Let C_h and C_m be the costs for assigning a control to the high and medium risk categories, respectively. Then the Net Benefit or a risk model is

$$\begin{aligned}
NB &= B_h P(h|event)P(event) + B_m P(m|event)P(event) \\
&- C_h P(h|nonev)P(nonev) - C_m P(m|nonev)P(nonev).
\end{aligned}$$

Use the subscript n and o for the new and old risk models, respectively. Then

$$\begin{aligned}
\Delta NB &= B_h P(event)[P(h_n|event) - P(h_o|event)] & (9) \\
&+ B_m P(event)[P(m_n|event) - P(m_o|event)] & (10) \\
&- C_h P(nonev)[P(h_n|nonev) - P(h_o|nonev)] & (11) \\
&- C_m P(nonev)[P(m_n|nonev) - P(m_o|nonev)] & (12)
\end{aligned}$$

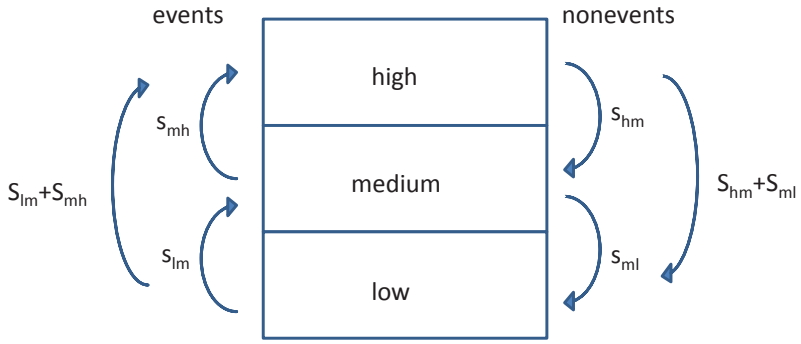

Figure 1: Parameters for the derivation of wNRI for 3 risk categories.

Now, $P(h_n) = P(lh) + P(mh) + P(hh)$ and $P(h_o) = P(hl) + P(hm) + P(hh)$, so $P(h_n) - P(h_o) = P(lh) + P(mh) - P(hl) - P(hm)$. The same holds when conditioning on event status and the same reasoning can be applied to $P(m_n) - P(m_o)$. Applying this reasoning and Bayes' rule gives the following expression for the change in Net Benefit for using the new risk model instead of the old risk model:

$$\begin{aligned}
\Delta NB &= B_h[P(event|lh)P(lh) + P(event|mh)P(mh) - P(event|hl)P(hl) - P(event|hm)P(hm)] \\
&+ B_m[P(event|lm)P(lm) + P(event|hm)P(hm) - P(event|ml)P(ml) - P(event|mh)P(mh)] \\
&- C_h[P(nonev|lh)P(lh) + P(nonev|mh)P(mh) - P(nonev|hl)P(hl) - P(nonev|hm)P(hm)] \\
&- C_m[P(nonev|lm)P(lm) + P(nonev|hm)P(hm) - P(nonev|ml)P(ml) - P(nonev|mh)P(mh)].
\end{aligned}$$

B.3 wNRI derived for three categories

Following Pencina et al. (4), let s_{lm} be the savings for re-classifying an event from low risk to medium risk and s_{mh} be the savings for re-classifying an event from medium risk to high risk. The savings from re-classifying an event from low risk to high risk is then $s_{lm} + s_{mh}$. Similarly, for nonevents we use parameters s_{hm} and s_{ml}. The total savings using the new risk model instead of the old risk model is

$$\begin{aligned}
n_{mh}[P(event|mh)s_{mh} - P(nonev|mh)s_{hm}] + \\
n_{lm}[P(event|lm)s_{lm} - P(nonev|lm)s_{ml}] + \\
n_{lh}[P(event|lh)(s_{mh} + s_{lm}) - P(nonev|lh)(s_{hm} + s_{ml})] + \\
n_{hm}[-P(event|hm)s_{mh} + P(nonev|hm)s_{hm}] + \\
n_{ml}[-P(event|ml)s_{lm} + P(nonev|ml)s_{ml}] + \\
n_{hl}[-P(event|hl)(s_{mh} + s_{lm}) + n_{lh}P(nonev|lh)(s_{hm} + s_{ml})]
\end{aligned}$$

Divide through by n so that $n_{mh}/n = P(mh)$ and so forth. Then the expected savings for use of the new risk model:

$$\begin{aligned}
& s_{mh}[P(event|mh)P(mh) + P(event|lh)P(lh) - P(event|hm)P(hm) - P(event|hl)P(hl)] \\
+ \; & s_{lm}[P(event|lm)P(lm) + P(event|lh)P(lh) - P(event|ml)P(ml) - P(event|hl)P(hl)] \\
+ \; & s_{hm}[P(nonev|hm)P(hm) + P(nonev|hl)P(hl) - P(nonev|mh)P(mh) - P(nonev|lh)P(lh)] \\
+ \; & s_{ml}[P(nonev|ml)P(ml) + P(nonev|hl)P(hl) - P(nonev|lm)P(lm) - P(nonev|lh)P(lh)]
\end{aligned}$$

Compare this expected savings with expression (8) for the generalized definition of the 3-category NRI. The expected savings can be viewed as a differently-weighted linear combination of $P(event|*)P(*)$ and $P(nonev|*)P(*)$ where $*$ represents movement between risk categories.

Now return to the expression for ΔNB and reparametrize: replace B_m with s_{lm} and B_h with $s_{lm} + s_{mh}$. Then from the first two lines we get:

$$\begin{aligned}
& s_{lm}[P(event|lh)P(lh) + P(event|lm)P(lm) - P(event|hl)P(hl) - P(event|ml)P(ml)] \\
+ \; & s_{mh}[P(event|lh)P(lh) + P(event|mh)P(mh) - P(event|hl)P(hl) - P(event|hm)P(hm)] \\
= \; & s_{lm}P(event)[P(lh|event) + P(lm|event) - P(hl|event) - P(ml|event)] \\
+ \; & s_{mh}P(event)[P(lh|event) + P(mh|event) - P(hl|event) - P(hm|event)]
\end{aligned}$$

For the second two lines replace C_m with s_{ml} and C_h with $s_{hm} + s_{ml}$. The last two lines of ΔNB are:

$$\begin{aligned}
& s_{ml}[P(nonev|hl)P(hl) + P(nonev|ml)P(ml) - P(nonev|lh)P(lh) - P(nonev|lm)P(lm)] \\
+ \; & s_{hm}[P(nonev|hl)P(hl) + P(nonev|hm)P(hm) - P(nonev|lh)P(lh) - P(nonev|mh)P(mh)] \\
= \; & s_{ml}P(nonev)[P(hl|nonev) + P(ml|nonev) - P(lh|nonev) - P(ln|nonev)] \\
+ \; & s_{hm}P(nonev)[P(ml|nonev) + P(mh|nonev) - P(lm|nonev) + P(hm|nonev)]
\end{aligned}$$

C Simulation Study: Methods

Our primary simulation model is Binormal Equal Correlation data (5). Let ρ denote disease prevalence. The old marker X and the new marker Y are bivariate Normal in both events and nonevents.

$$\begin{pmatrix} X \\ Y \end{pmatrix} \Big| D = 0 \sim N_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \Big)$$

$$\begin{pmatrix} X \\ Y \end{pmatrix} \Big| D = 1 \sim N_2\Big(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \Big)$$

A feature of this model is that the logistic model holds for both $P(D = 1|X = x, Y = y)$ and $P(D = 1|X = x)$.

$$\begin{aligned} logitP(D = 1|X = x) &= \log\frac{\rho}{1-\rho} - \frac{\mu_x^2}{2} + \mu_x x \\ logitP(D = 1|X = x, Y = y) &= \frac{\mu_X - r\mu_Y}{1-r^2}x + \frac{\mu_Y - r\mu_X}{1-r^2}y + \log\frac{\rho}{1-\rho} - \frac{\mu_X^2 + \mu_Y^2 - 2r\mu_X\mu_Y}{2(1-r^2)}. \end{aligned}$$

Therefore, when we apply logistic regression with data simulated from this model the risk model is correctly specified.

Note that μ_X and μ_Y summarize the marginal predictive abilities of X and Y respectively. r is the conditional correlation between the markers – conditional on disease status. Throughout this paper X represents the established marker(s) and Y represents the new predictor. The incremental value of Y depends not just on μ_Y but also on r and μ_X. In general the incremental value of Y is not a monotone function of μ_Y when $r \neq 0$ (2).

A convenient feature of this model is that there is a simple formula for $NRI^{>0}$:

$$NRI_e^{>0} = NRI_{ne}^{>0} = \frac{1}{2}NRI^{>0} = 2\Phi\Big(\frac{\sqrt{M_{X,Y}^2 - M_X^2}}{2}\Big) - 1.$$

where $M_{X,Y}^2$ is the squared Mahalanobis distance between events and nonevents in the distribution of (X, Y) and M_X^2 is the squared Mahalanobis distance between events and nonevents in the distribution of X. Φ is the distribution function of a standard Normal random variable. Any choice of simulation parameters, μ_X, μ_Y, and r exactly determine $NRI^{>0}$. When we consider the two-category NRI we use consider $NRI^{0.1}$. We calculated true values for $NRI^{0.1}$ by simulating datasets of size 5,000,000 and fitting the logistic models to get very precise estimates of the proportion of subjects with predicted risks above and below the high-risk threshold.

D Confidence intervals for NRI

Investigators seek to understand the nature of the improvement in risk prediction offered by a marker. To that end, it is of interest to estimate summaries of the prediction increment, and to quantify the uncertainty of those estimates using confidence intervals. For example, researchers routinely provide estimates and confidence intervals for the change in the area under the ROC curve, ΔAUC.

Many researchers are familiar with constructing confidence intervals for a parameter using the point estimate for the statistic and an estimate of its standard error: a 95% confidence interval for a parameter θ is formed as $\hat{\theta} \pm 1.96 \cdot \widehat{SE}(\hat{\theta})$. There are three requirements for a confidence interval constructed in this way to have the proper coverage: the estimate must be (1) consistent, which means that it estimates the true value in large samples; (2) have a Normal sampling distribution; and (3) $\widehat{SE}$ must be a consistent estimate of the standard error of the estimate.

Pencina et al. (4) provide a formula for estimating V_1, the variance of $\widehat{NRI}$. It is natural to construct a 95% confidence interval for the NRI using $\widehat{NRI} \pm 1.96 \cdot \sqrt{\widehat{V_1}}$. However, a confidence interval constructed in this way is valid only if conditions (1), (2), and (3) in the previous paragraph are true (or approximately true).

Pepe et al. (6) noted that $\hat{V}_1$ does not account for the variability of the fitted model. That is, when a risk model is fit to a dataset, there is uncertainty in coefficients of the model. This uncertainty should be incorporated into inferences about summaries of prediction performance or the increment of prediction. $\hat{V}_1$ ignores this uncertainty. Appendix E further elucidates problems with $\hat{V}_1$ as an estimate of the variance of $(\widehat{NRI^{>0}})$.

We conducted a simulation study to investigate whether confidence intervals have the correct coverage. We considered confidence intervals constructed as described above. We also evaluated confidence intervals constructed using $\widehat{NRI} \pm 1.96 \cdot \widehat{SE_B}(\widehat{NRI})$, where $\widehat{SE_B}(\widehat{NRI})$ is a bootstrap estimate of the standard error. Bootstrap estimates are obtained as follows. Re-sample rows of the original dataset with replacement to construct a "bootstrap dataset" of the same size as the original dataset. For a bootstrap dataset, re-fit the "old" and "new" risk models and calculate the NRI summary measures. Repeat this procedure a large number of times (e.g., 1000). This produces a distribution of values for the summary measure called the bootstrap distribution. The standard deviation of the bootstrap distribution is $\widehat{SE_B}$. Note that the bootstrap procedure incorporates the variability of the fitted model coefficients into estimating $SE(\widehat{NRI})$ because the risk model is re-fit on each bootstrap dataset.

Appendix C describes the simulation study. Table 1 gives the results for confidence intervals constructed using $\hat{V}_1$ and various bootstrap methods. Values in Table 1 should be compared to a target value of 0.05. Confidence intervals constructed using the formula for $\hat{V}_1$ have non-coverage proportions substantially above or below the target value. Non-coverage proportions substantially below 5% indicate conservative inference – confidence intervals are wider than they should be. Non-coverage proportions above 5% indicate anti-conservative inference. With anti-conservative inference, confidence intervals are too narrow and one is falsely confident of the precision of results. The worst performance was making confidence intervals for $NRI^{>0}_{ne}$ and $NRI^{0.1}_{ne}$, with non-coverage proportions 2-5 times as large as the target value.

Confidence intervals constructed using $\widehat{SE_B}$ show a clear tendency to give conservative results. While conservative inference is not desirable, anti-conservative inference is not acceptable, particularly at the levels we see in the tables for the formula for $\hat{V}_1$.

The other bootstrap methods for constructing confidence intervals did not work as as well as

$\widehat{NRI} \pm 1.96 \cdot \widehat{SE_B}(\widehat{NRI})$. We therefore recommend constructing confidence intervals by using a bootstrap estimate of the standard error of the statistic. Note that this method relies on approximate Normality for $\widehat{NRI}$. This is true asymptotically, but may not be a good assumption in small samples or for weak biomarkers, especially for the 2-category NRI (7).

Table 1 gives results of our simulation study evaluating seven methods of forming confidence intervals. Data were simulated as described in Appendix C with $\mu_X = 0.74, r = 0$, and three values for μ_Y. We considered seven methods for constructing confidence intervals.

1. $\widehat{NRI} \pm 1.96 \cdot \sqrt{\widehat{V}_1}$
2. $\widehat{NRI} \pm 1.96 \cdot \widehat{SE_B}(\widehat{NRI})$. This is the same as 1 but uses resampling-subjects bootstrapping to estimate the standard error.
3. Unadjusted. Uses resampling-subjects bootstrap but keeps the fitted models fixed.
4. Normal. This is similar to 2 but attempts to bias-correct the bootstrap estimate of the standard error.
5. Basic
6. Percentile. Take the .025 and .975 quantiles of the bootstrap distribution of the statistic.
7. Bias-corrected and accelerated intervals.

The last four methods are described at
www.unc.edu/courses/2007spring/enst/562/001/docs/lectures/lecture28.htm.

E The Variance of $\widehat{NRI}$

We simulated data as described in Supplement C. For all simulations we set the prevalence at 10% ($\rho = 0.1$) and conditional independence ($r = 0$). We considered various values for the marginal strength of the new marker Y, as indicated in the horizontal axis in the figures. We also considered small, medium, and large samples sizes (300, 1000, and 10000). For each simulated dataset, we fit the logistic model, computed $NRI^{>0}$, and computed $\hat{V}_1$. Across the 4000 simulations, we also computed the empirical variance of $\widehat{NRI}^{>0}$. This resulted in a single empirical estimate of variance($\widehat{NRI}^{>0}$) to compare to 4000 values of $\hat{V}_1$.

Figure 2 shows some of the problems with using $\hat{V}_1$ to estimate the variance of $NRI^{>0}$. If the incremental value of a marker is away from the null, $\hat{V}_1$ tends to underestimate the variance of $NRI^{>0}$. Near the null, $\hat{V}_1$ tends to overestimate the variance of $NRI^{>0}$. This may be because of boundary effects as described in Demler et al. (1) for ΔAUC.

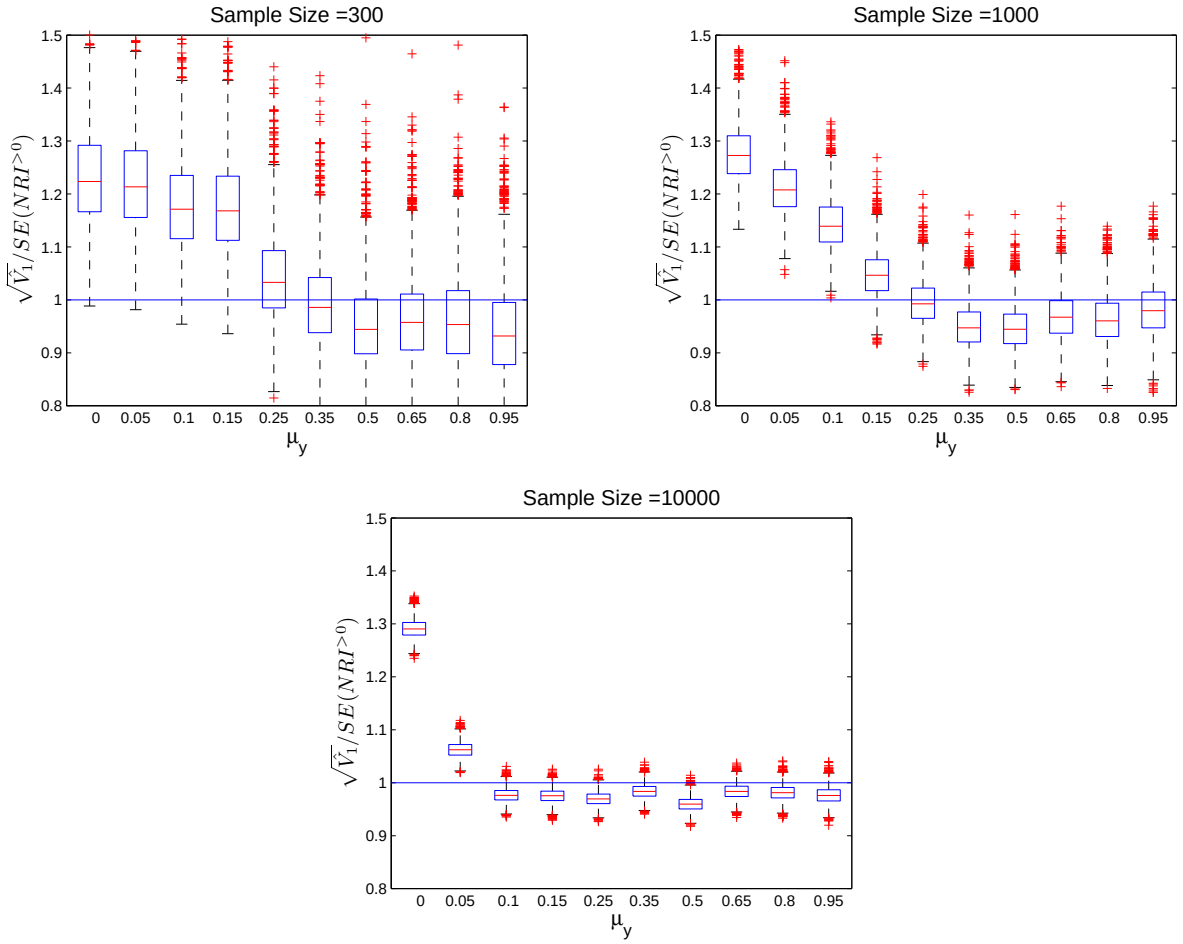

Figure 2: $\hat{V}_1$ as an estimate of the variance of $\widehat{NRI}^{>0}$. Results here are based on 4000 simulations for each μ_Y with $\rho = 0.1$ and $r = 0$. The sample size of the simulated datasets is given over each set of boxplots. The boxplots show the ratio of $\sqrt{\hat{V}_1}$ divided by the empirical standard deviation across the 4000 simulations. $\hat{V}_1$ tends to overestimate the variance when the incremental value of the marker is small and the sample size is small. For markers of modest incremental value and medium to larger sample sizes, $\hat{V}_1$ tends to underestimate the standard error of $NRI^{>0}$.

References

[1] Olga V. Demler, Michael J. Pencina, and Ralph?B. D'Agostino. Misuse of DeLong test to compare aucs for nested models. *Statistics in Medicine*, 31(23):2577–2587, 2012. ISSN 1097-0258. doi: 10.1002/sim.5328. URL `http://dx.doi.org/10.1002/sim.5328`.

[2] Kathleen F. Kerr, Aasthaa Bansal, and Margaret S. Pepe. Further insight into the incremental value of new markers: the interpretation of performance measures and the importance of clinical context. *Amerian Journal of Epidemiology*, X:in press, 2012.

[3] C.S. Peirce. The numerical measure of the success of prediction. *Science*, 4:453–454, 1884.

[4] Michael J. Pencina, Ralph B. D'Agostino Sr, and Ewout W. Steyerberg. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine*, 30:11–21, 2011.

[5] Michael J. Pencina, Ralph B. D'Agostino, and Olga V. Demler. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. *Statistics in Medicine*, 31(2):101–113, 2012.

[6] M.S. Pepe, Z. Feng, and J.W. Gu. 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. *Statistics in Medicine*, 27:173–181, 2008.

[7] Zheyu Wang. Asymptotic and finite sample behavior of net reclassification indices. Technical report, Department of Biostatistics, University of Washington, 2012. URL `http://biostats.bepress.com/uwbiostat/`.

	weak new marker ($\mu_Y = 0.17$)			
	$NRI_e^{>0}$	$NRI_{ne}^{>0}$	$NRI_e^{0.1}$	$NRI_{ne}^{0.1}$
formula	0.009	0.135	0.134	0.091
$\widehat{SE_B}$	0.012	0.035	0.004	0.004
Unadjusted	0.006	0.134	0.206	0.101
Normal	0.074	0.141	0.096	0.059
Basic	0.098	0.162	0.087	0.066
Percentile	0.009	0.024	0.001	0.002
BCA	0.066	0.132	0.142	0.097

	medium new marker ($\mu_Y = 0.34$)			
	$NRI_e^{>0}$	$NRI_{ne}^{>0}$	$NRI_e^{0.1}$	$NRI_{ne}^{0.1}$
formula	0.011	0.179	0.061	0.113
$\widehat{SE_B}$	0.035	0.067	0.011	0.011
Unadjusted	0.007	0.183	0.067	0.114
Normal	0.072	0.084	0.091	0.052
Basic	0.079	0.099	0.09	0.055
Percentile	0.016	0.040	0.001	0.009
BCA	0.065	0.065	0.124	0.087

	stronger new marker ($\mu_Y = 0.74$)			
	$NRI_e^{>0}$	$NRI_{ne}^{>0}$	$NRI_e^{0.1}$	$NRI_{ne}^{0.1}$
formula	0.008	0.178	0.044	0.266
$\widehat{SE_B}$	0.042	0.043	0.022	0.049
Unadjusted	0.006	0.179	0.046	0.268
Normal	0.068	0.051	0.061	0.064
Basic	0.073	0.056	0.071	0.079
Percentile	0.026	0.040	0.009	0.037
BCA	0.060	0.0423	0.074	0.067

Table 1: Non-coverage proportions for different types of confidence intervals. The method we recommend is in the row labeled $\widehat{SE_B}$ (it is called simply "bootstrap" in Table 3 in the article). Unadjusted, Normal, Basic, Percentile, and BCA are various types of bootstrap confidence intervals and are described in Appendix D.