8-16-2013

Net Reclassification Indices for Evaluating Risk Prediction Instruments: A Critical Review

Kathleen F. Kerr
University of Washington, katiek@u.washington.edu

Zheyu Wang
University of Washington - Seattle Campus, wangzy@u.washington.edu

Holly Janes
Fred Hutchinson Cancer Research Center, hjanes@scharp.org

Robyn McClelland
University of Washington, rmcclell@u.washington.edu

Bruce M. Psaty
University of Washington, psaty@u.washington.edu

See next page for additional authors

Authors
Kathleen F. Kerr, Zheyu Wang, Holly Janes, Robyn McClelland, Bruce M. Psaty, and Margaret S. Pepe

1. Introduction

Risk prediction is an important component of effective systems of medical care and public health. Examples of models for risk prediction in current use are the Framingham model[1] in cardiovascular disease and the Gail model[2] in breast cancer. Accurate risk prediction enables clinicians to match the intensity of treatment to the level of risk.[3] For many conditions, clinicians have a limited ability to accurately identify high risk patients, and research efforts continue to be devoted to improving risk prediction models. In cardiovascular disease, many epidemiological publications have evaluated whether new predictors can improve the risk predictions from the Framingham model, which includes the established risk factors age, sex, systolic blood pressure, lipids and smoking. The goal of such investigations is to evaluate new biomarkers for the predictive capacity they offer above and beyond established predictors. The improvement in risk prediction is called the incremental value or prediction increment of the biomarker.

In 2008 Pencina and colleagues[4] introduced a new measure of incremental value called the Net Reclassification Index or NRI. They expanded the definition of the NRI in 2011.[5] Variants of the NRI have recently become very popular in some areas of medical research, especially cardiovascular epidemiology. There are approximately 800 papers that contain "NRI" and cite the original[4] paper. It is important to understand what such a popular statistic measures and how it behaves.

Although NRI statistics have become popular, there are common mistakes in interpretation. Further, since there are now multiple NRIs to choose from, investigators may be unsure which, if any, to use. In addition, statistical methods pertaining to these indices are not yet well-developed. The goals of this review are (i) to clarify the interpretation of NRI statistics; (ii) to relate NRI statistics to more traditional measures; (iii) to provide guidance on choice of NRI statistics; (iv) to highlight problems with current methods for calculating confidence intervals and p-values with NRI statistics; and (v) to recommend methods for NRI confidence intervals.

1.1 NRI and other measures of the prediction increment

This section provides basic definitions and introduces the data on cardiovascular disease risk that we will use for illustration. Section 2 describes issues with the interpretation and application of both categorical and category-free NRI statistics. Section 3 describes statistical issues in applying NRI statistics. Section 4 applies the findings from Sections 2 and 3 to the MESA data. Section 5 summarizes our review and recommendations.

The context of this article is risk prediction. The specific goal is to improve risk prediction by adding a new predictor to an existing set of predictors. A traditional way to evaluate the prediction increment of a new biomarker is to consider the improvement in the area under the ROC curves for the expanded risk model compared to the risk model without the new predictor. In other words, one can consider the improvement in AUC (ΔAUC). However, promising new

markers have failed to produce meaningfully large increases in AUC. [4] There have been explicit calls to find ways to evaluate new marker other than ΔAUC.[6] Responding to these calls, Pencina and colleagues[4] proposed new metrics, IDI and NRI, for quantifying the prediction increment of a new marker. The NRI statistic has become extremely popular, and is the topic of this review.

The NRI, as originally proposed, seeks to quantify the effect of a new marker in moving predictions across clinically meaningful boundaries. In the definition of NRI, the risk prediction model that uses the established predictors is called the "old" model. The model that adds the new marker to the established predictors is the "new" model. "Events" are cases — individuals who have or will have the disease or outcome in the absence of intervention. "Nonevents" are controls. The formula defining NRI statistics is [4]

$$NRI = P(up|event) - P(down|event) + P(down|nonevent) - P(up|nonevent) \quad (1)$$

"Up" means that the new risk model places an individual into a higher risk category than the old model. Similarly, "down" means the new model places an individual into a lower risk category. For example, $NRI^{0.2}$ means a two-category NRI with cut-off at 0.20 defining low and high risk. $NRI^{0.1,0.2}$ is a three-category NRI with cut-offs at 0.10 and 0.20 defining low, medium, and high risk categories. Any set of risk thresholds can be used to define an NRI statistic.

The definition of the NRI in expression (1), which was originally based on discrete pre-defined risk categories, generalizes to any upward or downward movement in predicted risks.[5] The "category-free NRI" (also called "continuous NRI") interprets (1) this way. We use $NRI^{>0}$ to denote the category-free NRI.

The idea behind the NRI is that a valuable new biomarker will tend to increase predicted risks or risk categories for events; and decrease predicted risks or risk categories for nonevents. P(up|event) and P(down|nonevent) form the positive components of the NRI in expression (1). On the other hand, events that move down and nonevents that move up are mistakes introduced by the new marker — these are the negative components of (1).

An NRI statistic is the sum of the "event NRI" and the "nonevent NRI":

$$NRI_e = P(up|event) - P(down|event) \quad (2)$$
$$NRI_{ne} = P(down|nonevent) - P(up|nonevent) \quad (3)$$

For example, $NRI^{0.2} = NRI_e^{0.2} + NRI_{ne}^{0.2}$ and $NRI^{>0} = NRI_e^{>0} + NRI_{ne}^{>0}$.

For the two-category setting, Pencina et al.[5] generalized the NRI to consider the savings s_1 from identifying an event as high risk and s_2 from identifying a nonevent as low risk. s_1 is meant to capture the adverse events that are avoided by labeling a person destined to have an event as

high risk. s_2 should capture all the savings (adverse events, money) from allowing a nonevent to avoid unnecessary treatment. The "weighted NRI," wNRI, is the average savings per person.

$$wNRI = s_1\big(P(event|up)P(up) - P(event|down)P(down)\big) + s_2(P(nonevent|down)P(down) - P(nonevent|up)P(up)) \quad (4)$$

In this review we refer to two other measures of the prediction increment, ΔAUC (mentioned above) and ΔNB. The metric ΔNB refers to the change in Net Benefit associated with the use of the new marker.[7] For example, if the risk model is used to classify individuals as "high risk" or "low risk" and high risk entails an intervention, the Net Benefit is

$$NB = B \cdot P(event)P(high|event) - C \cdot P(nonevent)P(high|nonevent) \quad (5)$$

where *B* is the average benefit of the intervention among those who otherwise would have an event and *C* is the cost of intervention (including side effects) to nonevents. For old and new risk models, the change in Net Benefit, Δ*NB,* is a measure of the prediction increment of the new marker.

1.2 Example: Coronary Artery Calcification and Predicting Coronary Events

Polonsky et al.[8] examined the prediction increment of the coronary artery calcium score (CACS) for predicting coronary heart disease (CHD) among 5878 participants in the Multi-Ethnic Study of Atherosclerosis (MESA). Median follow-up was 5.8 years and 209 CHD events were observed. The cohort was 54% female, and the mean age was 62 years with a standard deviation of 10 years. The "old" risk model included the risk factors from the Framingham risk model and race; the "new" model added CACS. We will use these data to illustrate metrics and methods. We estimate risks using Cox models; for simplicity we otherwise ignore censoring in the data, following Polonsky et al.[8] We refer readers to the original paper[8] for more details.

2. Interpreting *NRI*

2.1 *NRI* is not a proportion

A common mistake is to interpret the *NRI* as a proportion.[9] For example, it is incorrect to interpret the *NRI* as "the proportion of patients reclassified to a more appropriate risk category."[10] That is *P*(*up* and *event*)+*P*(*down* and *nonevent*). The *NRI* combines four proportions but is not a proportion itself.[9] In particular, the maximum value of the *NRI* is 2.

NRI_e and NRI_{ne} are easier to interpret than NRI because they are differences in proportions. NRI_e is the net proportion of events assigned a higher risk or risk category. NRI_{ne} is the net

proportion of nonevents assigned a lower risk or risk category. The word “net” here is crucial for correct interpretation.

2.2 Issues with combining event and nonevent *NRI*s

Given the interpretations of NRI_e and NRI_{ne}, it is not clear why one would want to take a simple sum (or unweighted average) to produce the NRI. One logical alternative is to weight by the prevalence of events. This weighting extends the interpretations of NRI_e and NRI_{ne} to the whole population. We define the “population-weighted NRI” as $\rho NRI_e + (1-\rho)NRI_{ne}$, where ρ is the prevalence of the condition. The population-weighted NRI can be interpreted as the net change in the proportion of subjects assigned a more appropriate risk or risk category under the new model.

The MESA data illustrate another problem with the unweighted sum of NRI_e and NRI_{ne}. Using 5-year risks, $NRI^{0.1}$=0.164. Looking at the components we see that $\mathrm{NRI}_{\mathrm{e}}^{0.1}$=0.191 but the nonevent NRI is negative, $\mathrm{NRI}_{\mathrm{ne}}^{0.1}$= –0.027. Among nonevents, CACS introduces many more errors than corrections at the 10% risk threshold. Since there are many more nonevents than events (a common situation), the new risk model introduces many more errors than corrections overall. The positive value for $NRI^{0.1}$ masks the population-level results. Estimating the prevalence with 3.6%, the population-weighted $NRI^{0.1}$ is –0.020. That is, the net proportion of subjects assigned to a more appropriate risk category using the 0.1 threshold is –0.02.

The population-weighted NRI illustrates one problem with the NRI. However, we do not advocate its use because there is no compelling advantage in collapsing NRI_e and NRI_{ne} into a single number. NRI_e and NRI_{ne} tell us how the new risk model (potentially) improves prediction for events and, separately, for nonevents. The two types of improvements have different implications. Important information is lost when these two summaries are combined. [11]

2.3 Large and small values for $NRI^{>0}$ are undefined

Ideally, a measure of incremental value has an interpretation in terms of the clinical or public health benefit of incorporating the new marker. Pencina et al.[12] remark that “further research is needed to determine meaningful or sufficient degree of improvement” in $NRI^{>0}$”. $NRI^{>0}$ has no interpretation that translates to the clinical benefit of the new marker.[13] If it did, then the magnitude of the index would be directly applicable and a marker's sufficiency for improving prediction would be apparent. Other metrics, including ΔAUC, share the problem of lacking a clinically meaningful interpretation. However, an additional problem with $NRI^{>0}$ is that its scale is unfamiliar.

Pencina et al.[12] give a mathematical example of a new marker described as having “strong effect size.” Supplement C describes the structure of the data considered by Pencina et al.[12] Here and throughout this review, *X* represents the established predictor or set of predictors, and

Y is the candidate new predictor. In the example,[12] the new marker *Y* yields $NRI^{>0}$ =0.622. Is 0.622 large? Consider Figures 1 and 2. In all four examples in the figures, *Y* has the same distribution, and the odds ratio for *Y* given the baseline marker *X* is constant. The four examples differ only in the strength of the old risk model, i.e., the predictive capacity of *X*. At one extreme, the old risk model is useless, with *AUC*=0.5. At the other extreme, the old risk model is excellent with AUC=0.99. The figures suggest that the prediction increment for *Y* diminishes as the strength of the old model increases. Yet $NRI^{>0}$=0.622 in all four cases. Clearly there are important aspects of prediction not captured by $NRI^{>0}$.[12]

2.4 $NRI^{>0}$ does not contrast the performance of the new risk model with the performance of the old risk model

Most measures of incremental value are constructed by summarizing the performance of the old risk model, summarizing the performance of the new risk model, and comparing the two summaries. ΔAUC and ΔNB are two examples. $NRI^{>0}$ is fundamentally different. It is not a difference of two performance measures for the two risk models. Instead, for each individual it compares the old and new risk values. However, within-individual changes in risk do not necessarily translate into improved performance on a population level. For example, Figure 2 (bottom row) shows examples where there are lots of changes in individual predicted risks ($NRI^{>0}$=0.622), but the distribution of predicted risks in the population remains almost exactly the same.

When assessing a new biomarker, ultimately we want to know whether clinicians should continue using the old risk model or switch to the new, expanded risk model. To answer this question we need to assess the performances of each of the risk models and compare them. $NRI^{>0}$ measures the difference between the old and new risk models within individuals without teaching us about the performances of the models.

2.5 $NRI^{>0}$ incorporates irrelevant information

$NRI^{>0}$, like ΔAUC, does not rely on risk thresholds. Greenland[14] points out that “cutpoint free” indices incorporate irrelevant information, diminishing their potential for clinical relevance. For example, AUC summarizes the entire ROC curve, including parts of the curve describing sensitivity for unacceptably poor specificity. There are two ways in which $NRI^{>0}$ incorporates irrelevant information. First, $NRI^{>0}$ does not account for the size of changes in a predicted risk. Infinitesimally small changes “count” even though they are clinically irrelevant. Second, $NRI^{>0}$ does not account for an individual's position on the risk distribution. An event at the high end of the risk distribution who moves to an even higher risk reflects positively on $NRI^{>0}$. Such movement likely has little effect on treatment decisions. A new marker is beneficial if it improves treatment decisions, which often means the marker can discriminate between events and nonevents in the middle of the risk distribution.

For the MESA data, $\mathrm{NRI}_{\mathrm{e}}^{>0}$=0.378 and $\mathrm{NRI}_{\mathrm{ne}}^{>0}$=0.319. 20.6% of events have a new 5-year risk within 1% of the old risk. Among non-events the proportion is 52.8%. Therefore, a sizeable proportion of changes summarized by $\mathrm{NRI}_{\mathrm{e}}^{>0}$ and especially by $\mathrm{NRI}_{\mathrm{ne}}^{>0}$ are small, likely inconsequential changes.

2.6 NRI$^{>0}$ can make uninformative new markers appear predictive

Hilden and Gerds[15] and Pepe and colleagues[16] report a problematic feature of NRI$^{>0}$. Suppose an old risk model (risk(X)) and a new risk model (risk(X,Y)) are fit to a training dataset. Suppose further that the new marker Y is completely uninformative. To avoid the optimistic bias caused by using the same data to fit and evaluate model performance, a standard strategy is to use an independent dataset to assess the models' performances. However, *NRI*$^{>0}$ tends to be positive for uninformative Y, even when NRI$^{>0}$ is computed on a large, independent validation dataset that was not used to fit the models.[16] This problem is likely to arise in settings where the risk models are not well calibrated, a common phenomenon in practice. In contrast to NRI$^{>0}$, more standard measures such as ΔAUC do not suffer this problem. These results show that NRI$^{>0}$ can mislead researchers to believe that an uninformative marker improves prediction.

2.7 For 3+ categories *NRI* weights reclassifications indiscriminately

The purpose of risk categorization is to guide appropriate treatment decisions. For cardiovascular disease, suppose low risk indicates no intervention, medium risk indicates lifestyle changes, and high risk indicates both lifestyle changes and pharmaceutical intervention. When categories correspond to treatment decisions, the nature of reclassification matters, not just the direction. For example, an event whose risk category changes from high risk to low risk is a more serious error than an event moving from high risk to medium risk.

When there are three or more risk categories, one should consider all the ways a new biomarker can move individuals among risk categories. For three risk categories there are three ways of moving "up": low risk to medium risk; medium to high; and low to high. The 3-category $\mathrm{NRI}_{\mathrm{e}}$ gives each of these equal weight; in particular, moving up two risk categories counts the same as moving up one. Supplement B describes how an appropriate weighting could be incorporated into a statistic. Weighting the different types of reclassification is extremely challenging, but that challenge does not justify using equal weights. As an alternative to assigning weights and providing a single numerical summary, one can instead examine the different types of reclassification in a reclassification table (e.g., Table 2).

Polonsky et al.[8] considered 3-category NRIs with thresholds at 0.03 and 0.1 defining low, medium, and high 5-year risk. NRI$^{0.03,\,0.1}$=0.25. The value is driven by events ($\mathrm{NRI}_{\mathrm{e}}^{0.03,0.1}$=0.225 and $\mathrm{NRI}_{\mathrm{ne}}^{0.03,0.1}$=0.023), even though most of the population are nonevents. NRI$^{0.03,\,0.1}$=0.25 is a very coarse summary and almost impossible to interpret (see 2.2). Table 1 shows that the new risk model tends to place nonevents in the low and high risk categories, placing fewer

nonevents in the medium risk category than the old risk model. If the harm of moving a nonevent from medium risk to high risk is greater than the benefit of moving a nonevent from medium risk to low risk, then the harms of the new risk model outweigh the benefits among nonevents. The single numerical summary, $\mathrm{NRI}_{\mathrm{ne}}^{0.03,0.1}$=0.023, does not reflect this.

Table 2 shows the reclassifications of nonevents and, separately, events between the old and new risk models in the MESA data. Such tables are interesting and potentially instructive. However, it is easiest and most informative to simply look at how a risk model assigns nonevents and events to risk categories. This information appears on the margins of Table 2, and more succinctly in Table 1. NRI statistics do not capture this important information.

2.8 2-category *NRI*s: new names for existing measures

When there are two risk categories, low and high, NRI_e is the change in the proportion of events assigned to the high risk category, i.e., the change in the True Positive Rate (ΔTPR). NRI_{ne} is the change in the proportion of nonevents designated low risk. In other words, NRI_{ne} = –ΔFPR, where ΔFPR is the change in the False Positive Rate. For 2 risk categories, the population-weighted NRI (Section 2.2) is the change in the misclassification rate.

Furthermore, wNRI is the same as the change in Net Benefit between the old and new risk models (Supplement A or Van Calster et al[17]). In other words, wNRI=ΔNB.

3. Data Analysis with *NRI*

Common practice is as follows. Investigators have a dataset that includes established risk factors (*X*) for a condition of interest and a potentially useful new marker (*Y*). They fit two regression models: an "old" model that uses only X, and a "new" model that uses both X and Y. The risk models are typically logistic regression models, or Cox models if data are censored. The prediction increment of *Y* is then assessed, typically using the same data that were used to fit the models.

3.1 *NRI* should not be used for testing

A researcher may consider testing the null hypothesis H_0 : NRI=0. Pencina et al.[4] provide a z-statistic for NRI-based testing. However, the test based on this z-statistic has never been validated. Section 3.2 and Supplements D and E discuss problems with the variance formula that this z-statistic is based on.[18]

Interestingly for the category-free NRI, $\mathrm{NRI}^{>0}$, hypothesis testing is unnecessary. Pepe et al.[19] show that rejecting the null hypothesis H_0 : $\mathrm{NRI}^{>0}$=0 is implied by rejecting the null hypothesis about the novel marker being a risk factor. In other words, once a test on the coefficient of the new marker is performed, it is redundant to perform a test based on $\mathrm{NRI}^{>0}$.

For the two-category $\mathrm{NRI}_{\mathrm{e}}^{\mathrm{t}}$ or $\mathrm{NRI}_{\mathrm{ne}}^{\mathrm{t}}$ where t is the risk threshold, one cannot reject $H_0 : \mathrm{NRI}_{\mathrm{e}}^{\mathrm{t}} = 0$ and $H_0 : \mathrm{NRI}_{\mathrm{ne}}^{\mathrm{t}} = 0$ on the basis of Y being a risk factor. Good tests are not yet established for these null hypotheses.

We favor inference about the nature and size of the prediction increment rather than testing a null hypothesis of no improvement. Such inference is challenging. At the early stages of model development it might be unclear how a risk model will be used, yet understanding how a risk model will be used is important for appropriately evaluating incremental value. Setting aside these larger considerations, the next section considers methods for constructing confidence intervals for NRI statistics.

3.2 *NRI* Confidence Intervals

We conducted a simulation study to evaluate methods for constructed NRI confidence intervals. Based on Section 2, we only considered category-free and 2-category event and non-event NRI statistics. Results indicate that the most reliable confidence intervals use a bootstrap estimate of the variance of the statistic. Such confidence intervals outperformed confidence intervals constructed using the estimator $\widehat{V_1}$ proposed by Pencina et al.[4] and other types of bootstrap confidence intervals. Supplements C and D describe the simulation study and its results in detail.

4. *NRI* inference in the MESA data

In the MESA data, we used 5-year risk thresholds 0.03 and 0.1 following Polonsky et al.[8] Table 3 compares confidence intervals for category-free and various 2-category event and nonevent NRIs. Confidence intervals computed with bootstrapping are usually, but not always, wider than confidence intervals computed using $\widehat{V_1}$. For the 2-category NRIs with threshold 0.03 for 5-year risk, the changes in the true and false positive rates are modest, with an estimated 5.5% reduction in the false positive rate and 2.9% increase in the true positive rate. For the 0.1 risk threshold, adding *CACS* to risk prediction increases the true positive rate substantially (19.1%), but also increases the false positive rate by 2.7%.

Although the reclassification table (Table 2) and summary statistics (Table 3) are interesting, we find the risk distributions (Table 1) most useful. Table 1 shows that adding *CACS* to prediction increases the proportion of events labeled as high risk. Unfortunately, it also increases the proportion of nonevents labeled as high risk. Since nonevents vastly outnumber events, Table 1 identifies an important problem with adding *CACS* to the risk model.

5. Discussion and Summary

The recent literature on measures of incremental value developed as follows. Dissatisfaction with ΔAUC led to proposals for measures based on risk categories and reclassification.[20] The category-based NRI soon followed to address issues with those new measures.[4] A preference to avoid arbitrary or weakly-justified risk thresholds led to the proposal for $NRI^{>0}$.[5] Unfortunately, $NRI^{>0}$ has many of the same problems as ΔAUC. Neither measure is clinically meaningful, both measures are broad summaries of changes in risk models, and both measures incorporate irrelevant information. In these respects, things have come full circle. It is difficult to understand whether a value of $NRI^{>0}$ is large or small, and this is only partly due to lack of experience with the index. Furthermore, without proper attention to model fit, $NRI^{>0}$ can mislead researchers to believe that an uninformative marker improves prediction.[15-16] We are skeptical that $NRI^{>0}$ will help investigators develop biomarkers or improve risk models, and are concerned about the potential for $NRI^{>0}$ to mislead.

The NRI statistics that are most useful are re-named versions of existing measures. Specifically, (1) event and nonevent 2-category NRIs are the changes in the true and false positive rates; and (2) the weighted 2-category NRI is the change in Net Benefit. In both cases, we prefer the established, descriptive terminology.

We recommend the bootstrap for estimating the variance of NRI estimates and constructing confidence intervals. However, methodology that works well for markers with small prediction increment is needed.[21]

The issues described in Sections 2.3 and 2.4 for $NRI^{>0}$ also apply to NRIs for 3+ categories. However, the overriding issues with NRIs for 3+ categories is that they do not discriminate between different types of reclassification — all upward movements in risk categories count the same, as do all downward movements. We thus recommend against NRI statistics for 3 or more categories. As in the 2-category case, if the benefits and costs of different types of classification can be specified, these can be used as weights in a weighted NRI, which would be the same as the change in net benefit. This is a challenging approach and, to the best of our knowledge, has not yet been done in practice. A practical alternative is to examine how the old and new risk models place events and nonevents into the risk categories (e.g. Table 1). A reclassification table (e.g. Table 2) may also be interesting as it informs about classification achieved with the new marker within strata defined by the baseline risk model. Depending on the application, select 2-category summary statistics may be appropriate, particularly for risk thresholds that indicate expensive or invasive treatment.

$NRI^{>0}$ should not be used in hypothesis testing. Better tests are available and validated for the regression setting. However, we emphasize the limited value of hypothesis testing in assessing biomarkers. We recommend that investigators focus on describing the operating characteristics of risk models. Ideally, then, the prediction increment of a new marker is described in terms of how it improves risk model operating characteristics.

References

1. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97(18):1837-1847.

2. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. Journal of the National Cancer Institute. 1989;81(24):1879-86.

3. 27th Bethesda Conference. Matching the intensity of risk factor management with the hazard for coronary disease events. Journal of the American College of Cardiology. 1996;27(5):957-1047.

4. 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. 2008;27:157-172.

5. Pencina MJ, D'Agostino Sr RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Statistics in Medicine. 2011;30:11-21.

6. 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(21):2454-2456.

7. Peirce CS. The Numerical Measure of the Success of Prediction. Science. 1884;4:453-454.

8. Polonsky T, McClelland R, Jorgensen N, Bild D, Burke G, Guerci A, Greenland P. Coronary artery calcium score and risk classification for coronary heart disease prediction. Journal of the American Medical Association. 2010;303(16):1610-1616.

9. Leening MJG, Steyerberg EW. Fibrosis and mortality in patients with dilated cardiomyopathy. Journal of the American Medical Association. 2013;309(24):2547-2549.

10. Pickering JW, Endre ZH. New metrics for assessing diagnostic potential of candidate biomarkers. Clinical Journal of the American Society of Nephrology. 2012;7:1-10.

11. Pepe MS. Problems with risk reclassification methods for evaluating prediction models. American Journal of Epidemiology. 2011;173(11):1327-1335.

12. Pencina MJ, D'Agostino Sr RB, Pencina K, Janssens A, Greenland P. Interpreting incremental value of markers added to risk prediction models. American Journal of Epidemiology. 2012;176:473-481.

13. Kerr KF, Bansal A, Pepe MS. Further insight into the incremental value of new markers: the interpretation of performance measures and the importance of clinical context. American Journal of Epidemiology. 2012;176:482-487.

14. Greenland S. 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 MJ Pencina et al. Statistics in Medicine. 2008;27(2):199-206.

15. Hilden J, Gerds TA. Evaluating the impact of novel biomarkers: Do not rely on integrated discrimination improvement and net reclassification index. Statistics in Medicine. 2013.

16. Pepe M, Fang J, Feng Z, Gerds T, Hilden J. The Net Reclassification Index (NRI): a Misleading Measure of Prediction Improvement with Miscalibrated or Overfit Models. UW Department of Biostatistics Working Paper Series. 2013; Paper 392.

17. Van Calster B, Vickers AJ, Pencina MJ, Baker SG, Timmerman D, Steyerberg EW. Evaluation of Markers and Risk Prediction Models: Overview of Relationships between NRI and Decision-Analytic Measures. Medical Decision Making 2013 33:490.

18. Pepe MS, Feng Z, Gu JW. Comments on `Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by Pencina et al. Statistics in Medicine. 2008;27:173-181.

19. Pepe MS, Kerr KF, Longton G, Wang Z. Testing for improvement in prediction model performance. Statistics in Medicine. 2013;32(9):1467-1482.

20. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007;115(7):928-935.

21. Uno H, Tian L, Cai T, Kohane IS, Wei LJ. A unified inference procedure for a class of measures to assess improvement in risk prediction systems with survival data. Statistics in Medicine. 2013;32(14):2430-2442.

Acknowledgements

The MESA study was supported by contracts N01-HC-95159 through N01-HC-95169 from the National Heart, Lung, and Blood Institute. This work was also supported by grant NIH GM054438 to MSP, grant NIH R01 CA152089 to HJ, and a subcontract to the University of Washington from NIH grant HL085757-07 to KFK.

Figures

Figure 1. In each plot the solid black line is the ROC curve for the “old” marker and the dotted blue line is the ROC curve for the “new” risk model that incorporates the new marker. The new marker has identical distribution in all four cases. $NRI^{>0}$=0.622 in all cases, despite the fact that the prediction increment of the new marker decreases as the strength of the old model increases.

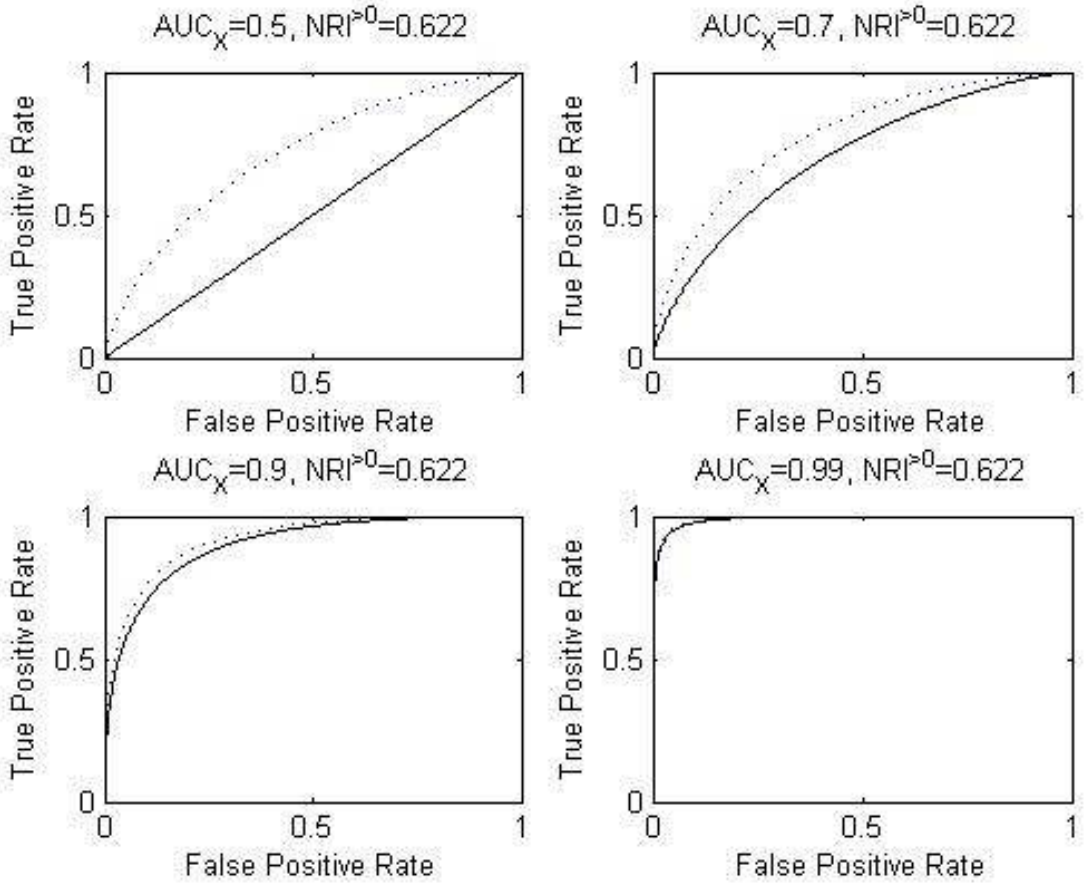
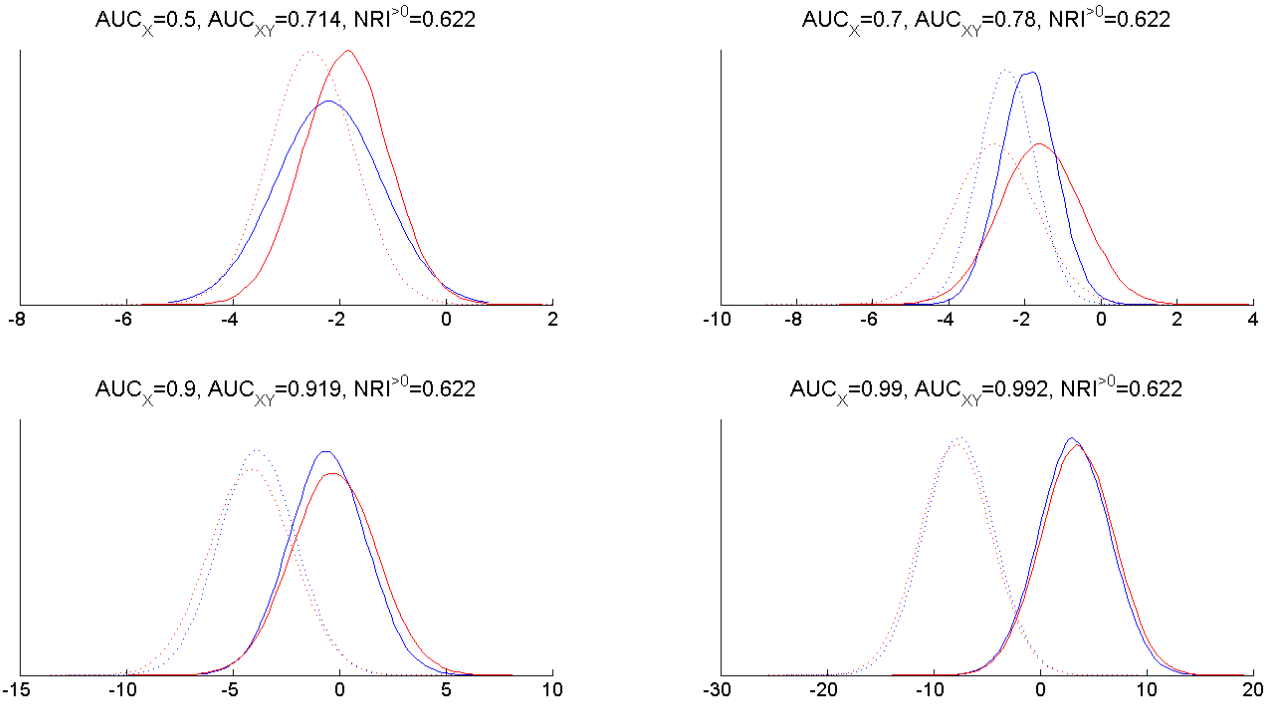


Figure 2: The same data as Figure 1 are shown here in terms of the distributions of risks for old and new risk models. Risks are shown on the log odds scale. Blue curves are the risks using the established predictors *X* and red curves are risks using *X* as well as the new marker *Y*. Dotted lines are nonevents and solid lines are events.



Tables

Table 1. Proportions of subjects in low, medium, and high risk categories in the MESA data, presented separately for events (those with coronary heart disease) and nonevent (those without coronary heart disease).

Risk	Old risk model		New risk model (model with CACS)	
Category	nonevent	event	nonevent	event
0-3%	67.1%	27.3%	70.6%	24.4%
3–10%	30.6%	55.0%	22.3%	38.8%
>10%	4.4%	17.7%	7.1%	36.8%
Total	5669	209	5669	209
	100%	100%	100%	100%

Table 2: Reclassification table for nonevents and events in the MESA data. Each interior cell contains the number of individuals in the corresponding risk categories under the old and new risk models. The percentages in interior cells are among nonevents or events. The rows and columns labeled “Total” show the distributions of nonevents and events into the three risk categories under the old and new risk models – the same data are found in Table 1

Nonevents

Old Model	Model with CACS 0-3%	3-10%	>10%	Total
0–3%	58%	7%	1%	
	3276	408	5	65%
3–10%	12%	14%	4%	
	697	791	244	31%
>10%	1%	1%	3%	
	30	63	155	4%
Total	71%	22%	7%	5669

Events

Old Model	Model with CACS 0-3%	3-10%	>10%	Total
0–3%	16%	11%	0%	
	34	22	1	27%
3–10%	7%	25%	23%	
	15	52	48	55%
>10%	1%	3%	13%	
	2	7	28	18%
Total	24%	39%	37%	209

Table 3: Confidence Intervals for select event and nonevent *NRI*s in the MESA data. Intervals based on bootstrap estimates of the standard error, which we recommend, tend to be wider than intervals based on the formula for the variance of the estimated NRI statistic.. Recall that for a threshold *t* delineating high risk, NRI_e^t = Δ*TPR* and NRI_{ne}^t = –Δ*FPR*.

	$NRI_e^{>0} = 0.378$	$NRI_{ne}^{>0} = 0.319$
formula	(0.252,0.503)	(0.294,0.344)
bootstrap	(0.275,0.481)	(0.257,0.382)
	$NRI_e^{0.03} = 0.029$	$NRI_{ne}^{0.03} = 0.055$
formula	(–0.030,0.088)	(0.044,0.067)
bootstrap	(–0.039,0.097)	(0.026,0.084)
	$NRI_e^{0.1} = 0.191$	$NRI_{ne}^{0.1} = 0.027$
formula	(0.125,0.258)	(–0.034,–0.021)
bootstrap	(0.097,0.286)	(–0.039,–0.016)

SUPPLEMENTARY ONLINE MATERIAL
Net Reclassification Indices for Evaluating Risk Prediction Instruments: A Critical Review

Kathleen F. Kerr, Zheyu Wang, Holly Janes,
Robyn L. McClelland, Bruce M. Psaty, Margaret S. Pepe

July 10, 2013

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

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

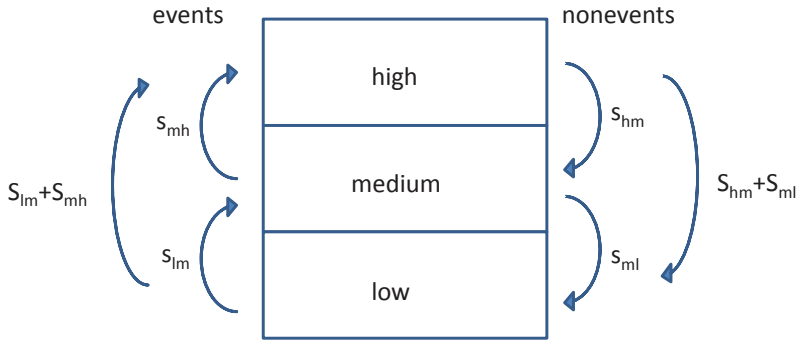


Figure 1: Parameters for the derivation of wNRI for 3 risk categories.

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.

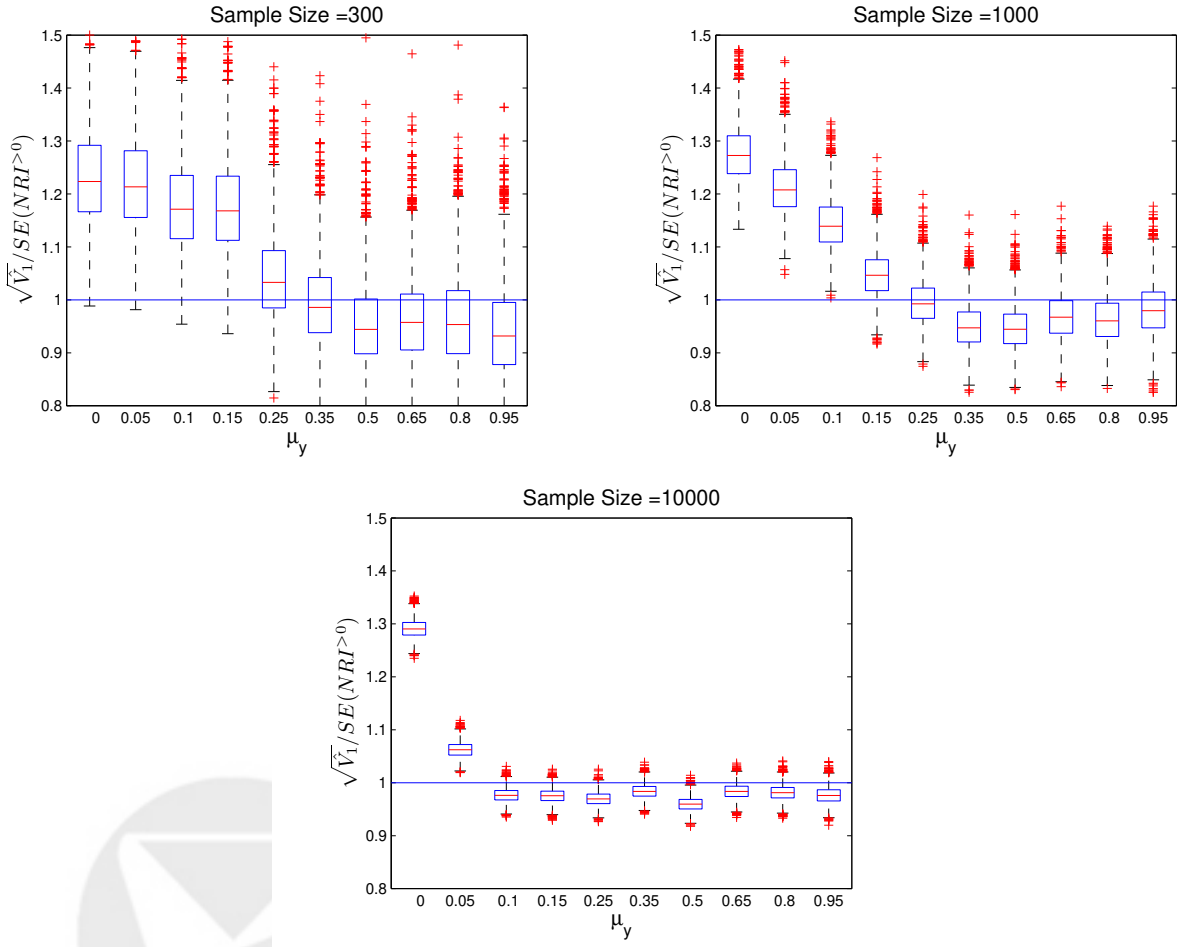


Figure 2: $\hat{V}_1$ as an estimate of the variance of $\widehat{NRI}^{>0}$. Results here are based on 4000 simulations for each μ_Y with $\rho = 0.1$ and $r = 0$. The sample size of the simulated datasets is given over each set of boxplots. The boxplots show the ratio of $\sqrt{\hat{V}_1}$ divided 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.