New Metrics for Assessing Diagnostic Potential of Candidate Biomarkers

John W. Pickering* and Zoltan H. Endre†

Summary
New tests should improve the diagnostic performance of available tests. The area under the receiver operator characteristic curve has been the “metric of choice” to quantify new biomarker performance. Two new metrics, the integrated discrimination improvement (IDI) and net reclassification improvement (NRI), have been rapidly adopted to quantify the added value of a biomarker to an existing test. These metrics require the development of risk prediction models that calculate the probability of an event for each individual. This study demonstrates the application of these metrics in 528 critically ill patients with risk models of AKI, sepsis, and 30-day mortality to which the biomarker urinary cystatin C was added. Analogous to the receiver operator characteristic curve, we present a new risk assessment plot for visualizing these metrics. The results showed that the NRI was sensitive to the choice of risk threshold. The risk assessment plot identified that the addition of urinary cystatin C to the model decreased the calculated risk for some who did not have sepsis but increased it for others. The category-free NRI for each outcome indicated that most of those without the event had reduced calculated risk. This was driven by very small changes in calculated risk in the AKI and death models. The IDI reflected those small changes. Of the new metrics, the IDI, reported separately for those with and without the events, best represents the value of a new test. The risk assessment plot identified differences in the models not apparent in any of the metrics.


Introduction
Risk prediction models are important clinical and research tools in many medical fields. A new candidate biomarker must improve the model to be of additional benefit. New statistical metrics have been developed to assess the incremental diagnostic value of a new biomarker. Clinicians reading research studies over the past decade have become familiar with the area under the receiver operator characteristic curve (AUC) as the primary tool to report diagnostic potential. The AUC is easy to interpret. An AUC of 1 represents perfect discrimination between the diseased and non-diseased patients—all patients are correctly classified by the test. An AUC of 0.5 represents no discrimination at all—patients are correctly classified no more frequently than can be attributed to chance. A recent AKI biomarker review raised concerns around the inadequacy of the AUC (1). Perhaps the most serious issue with the AUC is that it is an insensitive measure of the ability of a new marker to add value to a pre-existing risk prediction model (2). From a clinical perspective, it does not provide good information on whether adding this biomarker to the other relevant diagnostic information will more accurately identify individual risk.

Two new metrics, namely the integrated discrimination improvement (IDI) and net reclassification improvement (NRI), have been recently introduced to assess the added value of a candidate biomarker to pre-existing risk prediction models (3). A third metric, the category-free NRI (cfNRI) has recently been touted to overcome some of the shortcomings of the NRI (4). A systematic literature review of PubMed for reclassification studies showed that of 48 studies published between the publication of the NRI in 2008 and January 2010, 38 used the NRI and 19 the IDI (5). Nearly all studies assessed cardiovascular outcomes or mortality. The first use of the NRI in the nephrology literature seems to be a study of the added value of urinary neutrophil gelatinase-associated lipocalin to a clinical model for diagnosis of AKI (6). Since then, the NRI has been used in assessing the ability of biomarkers to predict dialysis after transplantation (7), development of CKD and microalbuminuria (8), recovery from dialysis-dependent AKI (9), the absence of AKI (10), worsening of AKI (11), and the presence of AKI (12–15). More recently, it has been used in its category-free (or continuous) form (16,17). Several studies have also reported the IDI (8,11,13–17). Although the primary responsibility for appropriate use of these new metrics lies with study authors, researchers and clinicians need to be able to interpret these metrics and understand their strengths and weaknesses. The interpretation and application of these metrics are likely to affect the choice of future biomarkers for further study, the eventual clinical use for diagnostic or predictive purposes, and the choice of AKI biomarkers for early intervention trials.

We have identified two related issues: (1) which statistics we should use to assess the added value of a biomarker to a risk model and (2) which risk categories we should adopt for clinical use. This article principally...
addresses the first issue. To aid interpretation of these metrics, we describe how they are defined and then apply them to a clinical example. To aid future application, we discuss some of their limitations, introduce a graphical technique to visualize the data, and make a series of recommendations.

Defining the New Metrics for Biomarker Assessment

The metrics require a reference risk prediction model that calculates the probability (calculated risk) of a patient having the event of interest (e.g., developing CKD, having AKI, etc.) and then a recalculated probability based on a new model compromising the reference model plus a new biomarker. In the nephrology literature to date all risk prediction models are determined from a statistical analysis of risk factors in the studied cohort. Normally, variables with a predetermined low P value under univariate analysis are included in a logistic regression model. An alternative approach is to use a model with prespecified variables such as the Framingham risk model for coronary heart disease, as discussed by Kivimäki et al. (18). Unfortunately, apart from the Thakar model for prediction of AKI after cardiac surgery (19), the field of AKI lacks a widely adopted model.

The NRI, cfNRI, and IDI each consider separately individuals who develop and who do not develop events. Therefore, they provide additional information not available from the AUC. For the NRI, each individual is assigned to a risk category—e.g., low (<5%), medium (5% to <20%), or high (≥20%)—based on the event probability calculated by the reference risk prediction model. A second model is constructed by adding the biomarker of interest to the reference model and each individual is reassigned to a new risk category. The net proportion of patients with events reassigned to a higher risk category (NRIf events) and of patients without events reassigned to a lower risk category (NRIf nonevents) is calculated (Appendix). The NRI is the sum of NRIf events and NRIf nonevents. It is interpreted as the proportion of patients reclassified to a more appropriate risk category. Among those with the event, if the addition of the biomarker of interest to the model results in an assignment to a higher risk category than to lower ones, then the NRI is positive. Conversely, among those without events, if more are assigned to lower than higher risk categories, then the NRI is positive. To illustrate how the metrics compare, we have constructed a table that illustrates the contribution to each metric for individuals with the event depending on their reference and new model probability of the event (calculated risk) (Table 1). Only those individuals for whom the addition of the new biomarker decreases or increases their calculated risk to the extent that they cross a category threshold contribute to the NRI.

The cfNRI, also called the continuous NRI, counts the direction of change for every individual rather than the crossing of a threshold. Each patient is counted as either +1 or −1 depending on whether the change in calculated risk was in the correct direction (higher for those with events, lower for those without events). The cfNRI is the sum of the cfNRIf events and cfNRIf nonevents, where the cfNRIf events is the proportion of patients with events who have an increase in calculated risk minus the proportion with a decrease and the cfNRIf nonevents is the proportion of patients without events who have a decrease in calculated minus the proportion with an increase (Table 1).

The IDI is independent of category and considers separately the actual change in calculated risk for each individual for those with and those without events (i.e., not merely the direction of change as with the cfNRI). In Table 1, patients 1 and 3 both show identical increases in calculated risk and both contribute equally to the IDI for those with events (IDI events), whereas only patient 1 contributes to the NRI because only patient 1 crosses the threshold between categories. Patients 2 and 4 contribute negatively to the IDI events because there is a lower calculated risk in the new model. Patient 4 contributes more than patient 2 because the change in calculated risk is greater. However, both equally contribute to the cfNRI because the cfNRI does not consider the magnitude of the change, only the direction. The IDI events is the difference between the mean of the new model risk probability for those with the event and the mean of the reference model probability for those with the event. Similarly, the IDI for those without events (IDI nonevents) is the difference in mean probability for those who do not have the event between the reference and new models (see the equation in the Appendix) (3). The IDI is the sum of IDI events and IDI nonevents. The IDI is also equal to the difference in the average sensitivity (normally termed the integrated sensitivity [IS]) of the two models across all risk thresholds and the IDI for those without events (IDI nonevents) is equal to the difference in the average 1-specificity (integrated 1-specificity [IP]). See the Appendix for details.

Interpretation of the New Metrics: A Worked Example

To illustrate the new metrics and how they are interpreted, we use data from the EARLYARF (Early intervention in Acute Renal Failure) trial, which has been described in detail elsewhere (20,21). Briefly, this study comprised 528 patients on entry to one of two general intensive care units (ICUs) who had multiple potential biomarkers of AKI measured including urinary cystatin C. We found cystatin C to be independently diagnostic of AKI and of sepsis and predictive of death using risk prediction models and the AUC statistic (22). AKI on entry was defined as a >0.3 mg/dl or 50% increase in plasma creatinine above baseline. Sepsis was determined independently by the attending ICU physicians referenced on the presence of ≥2 systemic inflammatory response syndrome criteria and confirmed or suspected viral or bacterial infection (22). Reference multivariate logistic regression models for each event (AKI, sepsis, 30-day mortality) were constructed with variables associated with the event at P<0.20 (Table 2) (22). Base 10-log(urinary cystatin C) was added to each model to find the new model. The integrals of sensitivity (IS) and 1-specificity (IP) over all possible threshold values for both those with and without events were also calculated. All analyses were performed using MATLAB 2011a software (MathWorks, Natick, MA).

Table 3 shows the event and nonevent classification tables for AKI, sepsis, and 30-day mortality for two-category
and they summarize each curve in a single metric similarly to the AUC. Ideally, IS equals 1 (all those with event have 100% risk of event) and IP equals 0 (all those without the event have 0% risk of event). The greater the separation between the reference and new model (dashed lines and solid lines, respectively, in Figure 1A) for those with events and those without events considered separately, the more the biomarker improves the diagnostic or predictive ability of the reference model.

**AKI**

The separation of the reference model event and non-event curves (dashed lines) in Figure 1A illustrates that the reference model reasonably discriminates between AKI and non-AKI. This is reflected in an AUC of 0.83. The non-event curve (IPref=0.19) more closely approaches the point (0, 0) than the events curve (ISref=0.50) approaches the point (1, 1). That is, the model overall better identifies those who do not have events than those who do. The addition of urinary cystatin C to the model makes very little difference. There was no significant change in the AUC, and we observed little difference between the reference and new model curves in Figure 1A. This is reflected by the confidence interval (CI) of the NRI at prevalence for both events and nonevents straddling 0 (Table 4). Interestingly, whereas the cnNRIevents behaved similarly, the cnNRInonevents showed substantial change (44%) resulting in a positive overall cnNRI (48%; CI, 24%–70%). This reflects changes in the relative order of calculated risk of those without events after addition of urinary cystatin C. The value of the IDI, which takes into account the size of those changes, is negligible with a 95% CI that straddles zero (IDInonevents=0.00052 to 0.00634). This means that the addition of cystatin C to the reference model reduced the calculated risk in 44% of those without events but not significantly.

**Sepsis**

Figure 1B illustrates that the sepsis model was less discriminatory than the AKI model (event and nonevent dashed curves are closer together), and that the addition of urinary cystatin C (new model) improved the ability of the model to quantify the risk of both those with and without the event (seen as the separation between the dashed and solid lines). The statistical metrics support this observation with all of the metrics other than NRIevents at prevalence demonstrating significant improvement. The
negative. Only the IDI events indicated any statistical difference between the models for those with events. Neither model calculated high risk for those with events.

Deaths leading in cases in which there is a cross-over in risk provide a good example of this. As shown in Figure 1B by the cross-over of the lines for the new model with the reference (dashed) model at a calculated risk of 0.3. For example, among those without events, the new model reduced the calculated risk <0.3 but increased the calculated risk >0.3. If the threshold for a two-category NRI was taken at, say, 0.5 rather than prevalence, then the NRI events was negative (−3.8%), whereas the NRI events was very large and positive (29%) because the reference model for the events was poor and total NRI positive (25%). Appendix Figures A1–A3 illustrate how both the NRI and IDI may be misleading in cases in which there is a cross-over in risk profiles between reference and new models.

**Death**

The reference and new models for events (solid lines) were poor with low IS (0.17 and 0.18). This is because neither model calculated high risk for those with events. Only the IDI events indicated any statistical difference between the models for those with events, although this was small (0.015; 95% CI, 0.00063–0.042). Although the

<table>
<thead>
<tr>
<th>Table 3. Event and nonevent tables for the addition of urinary cystatin C to reference models using event-specific prevalence as the threshold between two categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference + Urinary Cystatin C (New Model)</td>
</tr>
<tr>
<td>AKI reference events</td>
</tr>
<tr>
<td>&lt;27.8%</td>
</tr>
<tr>
<td>≥27.8%</td>
</tr>
<tr>
<td>new total</td>
</tr>
<tr>
<td>AKI nonevents</td>
</tr>
<tr>
<td>&lt;27.8%</td>
</tr>
<tr>
<td>≥27.8%</td>
</tr>
<tr>
<td>new total</td>
</tr>
<tr>
<td>Sepsis reference events</td>
</tr>
<tr>
<td>&lt;19.1%</td>
</tr>
<tr>
<td>≥19.1%</td>
</tr>
<tr>
<td>new total</td>
</tr>
<tr>
<td>Sepsis nonevents</td>
</tr>
<tr>
<td>&lt;19.1%</td>
</tr>
<tr>
<td>≥19.1%</td>
</tr>
<tr>
<td>new total</td>
</tr>
<tr>
<td>Death reference events</td>
</tr>
<tr>
<td>&lt;15.1%</td>
</tr>
<tr>
<td>≥15.1%</td>
</tr>
<tr>
<td>new total</td>
</tr>
<tr>
<td>Death nonevents</td>
</tr>
<tr>
<td>&lt;15.1%</td>
</tr>
<tr>
<td>≥15.1%</td>
</tr>
<tr>
<td>new total</td>
</tr>
</tbody>
</table>

*Prevalence was 27.8% for AKI, 19.1% for sepsis, and 15.1% for death (see text).”

The importance of visualizing the data is highlighted in Figure 1B by the cross-over of the lines for the new model with the reference (dashed) model at a calculated risk of 0.3. For example, among those without events, the new model reduced the calculated risk <0.3 but increased the calculated risk >0.3. If the threshold for a two-category NRI was taken at, say, 0.5 rather than prevalence, then the NRI events was negative (−3.8%), whereas the NRI events was very large and positive (29%) because the reference model for the events was poor and total NRI positive (25%). Appendix Figures A1–A3 illustrate how both the NRI and IDI may be misleading in cases in which there is a cross-over in risk profiles between reference and new models.

The use of the NRI in the general medical literature has recently been reviewed (5). The authors concluded that claims for improved reclassification are often spurious because of deficiencies in application. Pepe et al. described some of the problems with reclassification and recommended that the NRI for both those with and without events should be presented as well as the overall NRI (23,24). The NRI is sensitive to both the number of risk categories and the thresholds between categories. Using the simulated biomarker dataset of Pepe et al., we have illustrated this in the Appendix (23). The NRI tends to increase with increasing number of risk categories (25), whereas variation in NRI according to threshold increases with increasing difference in AUC between the models (26).

**How Best to Apply the New Metrics**

The NRI applied with multiple risk categories determined by individual research groups does not allow meaningful comparison between studies of the added value of AKI biomarkers to improve risk assessment because of the lack of broadly accepted and clinically meaningful risk categories. The five studies of AKI biomarkers that used the NRI each used a three-category model with different thresholds (6,12–15). The use of prevalence, if that prevalence fairly represents population prevalence, is one possible approach to a common threshold in which there are no agreed categories, although only for a two-category NRI. This is equivalent to the difference in the Youden index at the prevalence threshold (4). This threshold may or may not be of clinical relevance, but it does provide a way of comparing studies at similar prevalence. It is also analogous to using an objectively determined cut-point biomarker concentration derived from an receiver operating characteristic (ROC) curve (e.g., the

NRI nonevents at prevalence indicated a net 13.2% (6.3%–42%) more patients were assigned to the lower risk category and cNRI nonevents was 37%, the IDI nonevents was small with a 95% CI that straddled zero.

A two-category NRI with threshold chosen at prevalence did not reflect accurately the addition of cystatin C to risk models for AKI, sepsis, and death. In particular the NRI events and NRI nonevents for sepsis were only positive because of the choice of prevalence as the threshold, and the NRI nonevents for death was positive but did not reflect the very small difference in models. In contrast, the category-free NRI was positive in all cases. Nevertheless, this only indicated a reordering of cases and not the extent of the change in calculated risk. In all cases, the IDI statistics better represented the difference in models and the integrated sensitivities and specificities better represented the overall performance of the models. Graphically representing the data enhanced data interpretation because it revealed that in some cases the addition of cystatin C to the risk model improved the assessment of calculated risk above or below specific thresholds. This is analogous to a cross-over of receiver operator curves at different specificity.
The additional information provided by the cfNRI events will depend on the clinical importance of situations in which no established risk categories exist (4). AUCs between ROC curves. It has been recommended for absolute clinical utility and is similar to the difference in addition of a risk factor to an existing model rather than its required to determine these thresholds. Large epidemiologic studies are beyond the scope of this article. An alternative and preferable approach, in our opinion, is to use a consensus definition of risk threshold(s) for AKI. This would allow the choice of clinically meaningful thresholds. These may differ according to etiology and demographics (e.g., a cardiac surgery risk threshold may be different than a general ICU risk threshold). Large epidemiologic studies are required to determine these thresholds.

The category-free NRI is a measure of the validity of the addition of a risk factor to an existing model rather than its absolute clinical utility and is similar to the difference in AUCs between ROC curves. It has been recommended for situations in which no established risk categories exist (4). The additional information provided by the cfNRIevents and cfNRInonevents is more revealing than the total cfNRI because it allows for the assessment of the performance of the new model for those with and without events separately. The relative importance of the cfNRIevents and cfNRInonevents will depend on the clinical importance of detecting or excluding an event. The IDI is more meaningful than the cfNRI. Only the IDI incorporates both the direction of change in calculated risk and the extent of change. Reporting of both IDIevents and IDInonevents identifies whether the biomarker improves calculated risk more for those with or without events. In contrast, it is theoretically possible for the cfNRI to be 200%, even though the increase of risk for each patient with the event and decrease for each subject without the event is minimal.

A risk assessment plot is more informative than an ROC plot because it illustrates separately how good each model is for both those with and without events. The IS and IP summarize these plots analogous to the way AUC summarizes a receiver operator characteristic curve. As can be seen for patients with sepsis (Figure 1B), there is additional value in visualizing the ranges of risk over which a new biomarker may improve or diminish risk prediction.

Finally, the NRI and IDI for a sample are estimates of the population NRI and IDI values and thus are meaningless unless presented with a CI. We recommend bootstrapping methods, particularly when the number of events is small.

Some caveats remain. The addition of a new biomarker to a poor reference model is likely to result in greater IDI.

<table>
<thead>
<tr>
<th>Event</th>
<th>AKI</th>
<th>Sepsis</th>
<th>Death</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDIevents</td>
<td>0.01 (0.0002–0.03)</td>
<td>0.13 (0.08–0.20)</td>
<td>0.02 (0.0006–0.04)</td>
</tr>
<tr>
<td>IDInonevents</td>
<td>0.003 (0.0005 to 0.01)</td>
<td>0.03 (0.02–0.05)</td>
<td>0.004 (0.0004 to 0.01)</td>
</tr>
<tr>
<td>IDI</td>
<td>0.02 (0.002–0.05)</td>
<td>0.16 (0.10–0.25)</td>
<td>0.02 (0.003–0.05)</td>
</tr>
<tr>
<td>ISref</td>
<td>0.49 (0.40–0.55)</td>
<td>0.25 (0.19–0.29)</td>
<td>0.17 (0.12–0.20)</td>
</tr>
<tr>
<td>ISnew (reference + urinary cystatin C)</td>
<td>0.51 (0.43–0.57)</td>
<td>0.38 (0.30–0.45)</td>
<td>0.18 (0.13–0.22)</td>
</tr>
<tr>
<td>IPref (reference)</td>
<td>0.19 (0.16–0.23)</td>
<td>0.18 (0.15–0.21)</td>
<td>0.15 (0.12–0.18)</td>
</tr>
<tr>
<td>IPnew (reference + urinary cystatin C)</td>
<td>0.19 (0.16–0.22)</td>
<td>0.14 (0.12–0.18)</td>
<td>0.15 (0.12–0.18)</td>
</tr>
<tr>
<td>AUC</td>
<td>reference</td>
<td>0.83 (0.78–0.87)</td>
<td>0.70 (0.64–0.76)</td>
</tr>
<tr>
<td>reference + urinary cystatin C</td>
<td>0.84 (0.79–0.88)</td>
<td>0.82 (0.77–0.88)</td>
<td>0.65 (0.58–0.72)</td>
</tr>
<tr>
<td>difference (P value)</td>
<td>0.097</td>
<td>&lt;0.0001</td>
<td>0.066</td>
</tr>
</tbody>
</table>

The 95% confidence intervals are shown in parentheses. NRI, net reclassification improvement (NRIevents + NRInonevents); cfNRI, category-free NRI (cfNRIevents + cfNRInonevents); IS, integrated sensitivity (ideally IS=1); IP, integrated 1-specificity (ideally IP=0); IDI, integrated discrimination improvement ([ISnew – ISref] + [IPref – IPnew]); AUC, area under the curve of the receiver operator characteristic curve.

*A Hosmer–Lemeshow goodness of fit was used to test calibration of the model (5).
and NRI indices than when it is added to a good model. This highlights the importance of reporting AUC, IS, and IP for the reference model. The new metrics are applicable only when a model (often a logistic regression model) can be reasonably constructed from the dataset. When the dataset is small, fewer variables can legitimately be included in any model. As a rule of thumb, we use a maximum of one variable for every 10 participants in the cohort with the least number of participants (event or nonevent). These new metrics are not helpful for small cohorts in which risk prediction models cannot be constructed. Such small cohorts (phase 1 and 2) may be important in studies of a new biomarker or a biomarker in a new setting. The AUC and perhaps presentation of sensitivity and specificity at clinically relevant and/or optimal biomarker cut-points are still appropriate metrics in such studies.

Larger studies (phase 3 or 4) should evaluate the various metrics. Only after multiple studies in multiple settings are conducted will we truly understand what an NRI of 20% or IDI of 0.3 actually translate to in terms of clinical utility. The question that needs addressed in such studies is just how large a change in each metric needs to be for clinical relevance? For example, if the IDI_{events} increases by 0.1, is this a sufficient change to warrant the inclusion of the biomarker under investigation in a screening program? The answer will, of course, depend on the purpose of screening. Factors such as prevalence, and interventions with significant potential harm or cost will affect the answer. For example, if the screening is to include people in an intervention trial for AKI after cardiac surgery, where incidence is low, then it is likely to be necessary to also observe a reduction in calculated risk in those without AKI (large IDI_{nonevents}) so as to minimize unnecessarily treating individuals without the disease. If, on the other hand, the goal is screening patients on entry to the ICU so as to avoid the use of nephrotoxins in those with AKI, then a large IDI_{nonevents} score will not be as important as a large IDI_{events} score.

**Recommendations for Application and Reporting**

We recommend the following: that the IDI be reported, along with IDIs for events and nonevents; that cfNRI and cfNRIs for events and nonevents be reported if they provide additional information to the IDI; that NRI be used only for events in which there are clinically meaningful risk categories with broad acceptance; that determination of NRI risk categories be based on large-multicenter epidemiologic studies, and that the minimum number of categories be used that allow clinically meaningful separation between low and high risk categories. In addition, we recommend that all metrics be reported with 95% CIs; that NRI and cfNRI be reported as a proportion or percentage, but IDI as a raw number (i.e., not as a percentage); that risk assessment plots be used to represent the data graphically; that the IS and IP be reported as an summary performance metrics along with the AUC for each model; and that these metrics be evaluated and compared in large clinical data sets.

![Figure 1. Clinical model enhancement by adding urinary cystatin C in ICU patients.](image-url)
Appendix

Equations for Calculating NRI

For those with events:

\[
NRI_{\text{events}} = \frac{\# \text{ events moving up}}{\# \text{ events}} - \frac{\# \text{ events moving down}}{\# \text{ events}}
\]

For those without events:

\[
NRI_{\text{nonevents}} = \frac{\# \text{ nonevents moving down}}{\# \text{ nonevents}} - \frac{\# \text{ nonevents moving up}}{\# \text{ nonevents}}
\]

where moving up means a calculated risk increase for the individual moves them to a higher risk category on addition of the biomarker to the model. Similarly, moving down means that a calculated risk decrease moves them to a lower risk category.

The NRI is as follows:

![NRI graph](image_url)

Figure A1. Effect of varying the risk threshold between low and high risk categories on the NRI using a simulated data set (n=10,000). (A) Two-category (low-high risk) model. The NRI for a threshold of the prevalence of the events in the data set (10.17%) was 7.3%. (B) Three-category model showing variability in the threshold between low and medium, or medium and high risk categories changes the NRI (solid line). Data are the simulated data with an NRI of 17.4% (thin dashed line) for thresholds of 5% and 20% (risk categories <5%, 5%–20%, and >20%). NRI, net reclassification improvement.
RI = \text{NRI}_{\text{events}} + \text{NRI}_{\text{nonevents}}

The NRI requires only a definition of what constitutes a higher risk or lower risk reclassification. Rather than define this as crossing a threshold between categories, the cfNRI considers whether each individual moves up (to higher) or down in individual calculated risk. The cfNRI may be thought of as the NRI with a moving threshold of risk that is set to the event risk for each subject in the reference model. Each subject then either moves up in calculated risk (adding one to either the \#events moving up or \#nonevents moving down), stays at the same calculated risk, or moves down in calculated risk (adding one to either the \#events moving down or \#nonevents moving down). The maximum cfNRI is 200% (calculated risks for all subjects with events are increased, and all subjects without events are decreased).

IDI

The IDI is defined similarly to the cfNRI except that instead of adding 1, the actual difference in calculated risk between the two models for each individual is added. For example, the cfNRI treats a change of calculated risk of 0.005 and 0.5 identically, whereas the IDI gives more weight to a greater change in calculated risk.

For those with events:

\[
\text{IDI}_{\text{events}} = \sum \frac{\text{probability of event (new)} - \text{probability of event (reference)}}{\# \text{events}}
\]

For those without events:

\[
\text{IDI}_{\text{nonevents}} = \sum \frac{\text{probability of event (reference)} - \text{probability of event (new)}}{\# \text{nonevents}}
\]

The IDI is also the integral of the two-category NRI over all possible thresholds (see Supplemental Material for proof).

Category Thresholds and the NRI

To illustrate how the choice of thresholds influences the NRI, a larger data set was needed than clinically available. We used the simulated data of Pepe (23), which was specifically designed to model the effect of the addition of a positive biomarker on performance metrics. Data generation has been described in detail elsewhere (23). There were 10,000 subjects, of whom 1017 had events (23,24). Logistic regression was used to calculate the risk probabilities for a reference model and a new model with one more predictor variable. The reference model had an AUC of 0.88 (95% CI, 0.87–0.90), which increased to 0.92 (95% CI, 0.91–0.93) with the addition of the new predictor \( P < 0.001 \) [DeLong method of comparing AUCs (27)].

The NRI was calculated for a two-category (low and high risk) model with varying threshold from 0.5% to 50% and for a three-category (low, medium, and high risk) model where the lower and the upper thresholds were varied independently. For the clinical
data, a two-category model was used where the threshold was set at the prevalence for each outcome.

For a two-category analysis with varying threshold, the NRI varied between 6.3% and 13.2% (Figure A1A). If we vary one threshold (either 5% or 20%) and leave the other static for a three-category model (low, 5%–20% [medium], and high; see reference 24), the NRI varies from 12.9% to 22.0% (Figure A1B).

The cfNRIevents was 38.8% (95% CI: 34.6%–43.4%; i.e., 39% more of those with the event had an increase than had a decrease in risk), and cfNRInonevents was 41.1% (37.1%–44.7%; i.e., 41% more of those without the event had a decrease than had an increase in risk). The total cfNRI was 79.9% (95% CI, 73.1%–86.8%).

Risk Assessment Plots and Summary Statistics

The risk assessment plot (Figure A2) illustrates the NRI, IDI, IS, and IP (Figure A2). Each point on the plot represents the proportion of patients in the sample with risk above that of the calculated risk (sometimes called 1 – empirical cumulative distribution). For example, at a risk of 0.1, there are approximately 20% of patients without the event who have a risk of 0.1 determined by the reference model. Because the integral across all risk thresholds for a two-category NRI is equal to the IDI (Supplemental Material), the event curves in Figure A2 are the equivalent of plotting sensitivity against calculated risk (black curves) and the nonevent curves of plotting 1 – specificity against calculated risk (red curves). Both the difference between the reference and the new model and the performance of the models to stratify risk in those with and without events are illustrated. The NRI_{events} and NRI_{nonevents} at any level of risk may be read from the plot as the difference between the reference and new model event curves (dashed and solid black lines) and between the reference and new model nonevent curves (dashed and solid red lines). For example, at a risk of 0.1017 (the prevalence), NRI_{events} is 100 × (0.833 – 0.797) = 3.6% and NRI_{nonevents} is 100 × (0.198 – 0.161) = 3.7%, the sum of which is 7.3% (compare with Figure A1). The integrated sensitivity (IS) and integrated 1 – specificity (IP) are the areas under these curves. Ideally IS=1 (all those with the event have a risk of 1), and IP=0 (all those without the event have a risk of 0). IDI_{events} is, therefore, the area between the reference and new model event curves (IS_{new} – IS_{ref}) and IDI_{nonevents} is the area between the nonevent curves (IP_{ref} – IP_{new}), the sum of which is the IDI (see the appendix to Pencina et al. (3)).

The new metrics are potentially misleading when a reference and new model curve overlaps. Figure A3 is a risk assessment plot example for when the addition of a biomarker to a reference model results in, for those with the event, increased risk for subjects with a reference model risk below a certain risk (in this case, 0.7), but decreased risk for those with a reference model risk above that risk. For a two-category NRI when the threshold is low, there is a positive NRI_{events} (6% at threshold 0.35), and a negative NRI when threshold is high (16% at threshold 0.9). The IDI_{events} has a positive and negative component; positive below a risk of 0.7 and negative above 0.7. In this case, the resulting net is IDI_{events} of zero.

Acknowledgments

We thank Professor Margaret Pepe of the University of Washington for informative discussions and for providing access to the simulated dataset, and Dr Nick Cross of the Department of Nephrology at the Christchurch Hospital for useful suggestions regarding the presentation of these concepts.

J.W.P. was supported by an Australian and New Zealand Society of Nephrology infrastructure-enabling grant and by the Marsden Fund Council on government funding, administered by the Royal Society of New Zealand.

Disclosures

None.
References


Published online ahead of print. Publication date available at www.jasn.org.

This article contains supplemental material online at http://jasn.asnjournals.org/lookup/suppl/doi:10.2215/CJN.09909111/-/DCSupplemental.
Erratum

Correction

Pickering JW, Endre ZH: New Metrics for Assessing Diagnostic Potential of Candidate Biomarkers. *Clin J Am Soc Nephrol* 7: 1355–1364, 2012; published ahead of print June 7, 2012, doi:10.2215/CJN.09590911. In this paper, we wrote “The NRI is the sum of NRIevents and NRI_{nonevents}. It is interpreted as the proportion of patients reclassified to a more appropriate risk category.” Dr. Kathleen Kerr of the University of Washington and Dr. Margaret Pepe of the University of Washington and Fred Hutchison Cancer Research Center have kindly pointed out to us that the NRI is not a proportion; rather, it is a combination of four proportions. Therefore, we would like to clarify the interpretation of the metrics: NRI_{events} is the difference in two proportions, namely the proportion of patients with events moving to a higher-risk category minus the proportion with events moving to a lower-risk category. That is, the NRI_{events} is the net proportion, with events reclassified to a more appropriate risk category.

If NRI_{events} is negative, this indicates more patients with events have moved to a lower-risk category rather than a higher-risk category. The NRI_{nonevents} is also the difference in two proportions, namely the proportion of patients without events moving to a lower-risk category minus the proportion without events moving to a higher-risk category and may be interpreted as the net proportion, without events reclassified to a more appropriate risk category. If NRI_{nonevents} is negative, more patients without the event have moved to a higher-risk category than have moved to a lower-risk category [see the equations in the Appendix of our article and in the original article of Pencina et al. (1)].

Reference