Research Article

Statistics in Medicine

Received 9 December 2009, Accepted 23 August 2010 Published online 6 November 2010 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4085

Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers

Michael J. Pencina,[a,b*†] Ralph B. D'Agostino Sr[c] and Ewout W. Steyerberg[d]

Appropriate quantification of added usefulness offered by new markers included in risk prediction algorithms is a problem of active research and debate. Standard methods, including statistical significance and c statistic are useful but not sufficient. Net reclassification improvement (NRI) offers a simple intuitive way of quantifying improvement offered by new markers and has been gaining popularity among researchers. However, several aspects of the NRI have not been studied in sufficient detail.

In this paper we propose a prospective formulation for the NRI which offers immediate application to survival and competing risk data as well as allows for easy weighting with observed or perceived costs. We address the issue of the number and choice of categories and their impact on NRI. We contrast category-based NRI with one which is category-free and conclude that NRIs cannot be compared across studies unless they are defined in the same manner. We discuss the impact of differing event rates when models are applied to different samples or definitions of events and durations of follow-up vary between studies. We also show how NRI can be applied to case–control data. The concepts presented in the paper are illustrated in a Framingham Heart Study example.

In conclusion, NRI can be readily calculated for survival, competing risk, and case–control data, is more objective and comparable across studies using the category-free version, and can include relative costs for classifications. We recommend that researchers clearly define and justify the choices they make when choosing NRI for their application.

Keywords: discrimination; model performance; NRI; risk prediction; biomarker

1. Introduction

Risk prediction models are key components of prevention strategies adopted in a wide range of medical fields. Risk scores exist for cardiovascular disease (CVD) and its component diseases (coronary heart disease, stroke, atrial fibrillation), different forms of cancer, hypertension, diabetes and many others [1–7]. These algorithms have led to substantial advances in the reduction of disease burden by education, prevention and treatment. However, at the same time, there remains an ample opportunity for improvement, as in many instances intermediate and lower risk groups contribute the highest numbers of events. Hence, researchers have been looking for new biomarkers and genetic factors that could improve risk prediction. Multitudes of candidate genotypic and phenotypic markers had been postulated and it quickly became apparent that the existing statistical methods may not be sufficient to determine which of these new markers were actually useful. Standard methods proved either too liberal: significance of p-value is achieved easily in large sample or genetic studies, or too conservative: area under the curve (AUC) or C statistic hardly moves after a few good risk factors are already included in the model. Some researchers introduced a distinction between risk prediction and risk classification and suggested that measures going beyond statistical significance and C statistic are necessary [8].

To overcome these shortcomings Pencina and D'Agostino *et al.* [9] proposed two new measures to quantify the degree of correct reclassification: net reclassification improvement (NRI) and integrated discrimination improvement

[a]*Department of Biostatistics, Boston University, 801 Massachusetts Ave, Boston, MA 02118, U.S.A.*
[b]*Harvard Clinical Research Institute, 930 Commonwealth Ave, Boston, MA 02215, U.S.A.*
[c]*Department of Mathematics and Statistics, Boston University, 111 Cummington Street, Boston, MA 02215, U.S.A.*
[d]*Erasmus MC, Public Health, P.O. Box 1738, Rotterdam, 3000 DR, The Netherlands*
[*]*Correspondence to: Michael J. Pencina, Department of Biostatistics, Boston University, 801 Massachusetts Ave, Boston, MA 02118, U.S.A.*
[†]*E-mail: mpencina@bu.edu*

(IDI). They were developed for dichotomous outcomes and details can be found in [9]. Many interesting properties and extensions of the IDI have been presented in the statistical literature [10–12]. On the other hand, the NRI seems to have gained popularity in medical journals, most likely due to its simplicity [7, 13]. However, several issues have not been addressed with regard to NRI and its correct applications. These are the focus of the current paper.

Since many commonly used risk prediction models are based on time-to-event data, it is imperative to have an estimator of NRI applicable to survival data, which contains not only events and non-events, but also subjects who discontinue the study prematurely [14, 15]. Furthermore, extensions to competing risk models would be valuable given increasing interest in long-term risk prediction [16]. Similarly, extensions of NRI to case–control data are needed, as many biomarker studies are conducted using this design. The issue of weighting used in the calculation of NRI has been raised by several authors [12, 17, 18] and requires attention. The number and choice of categories and their impact on the magnitude and conclusions derived from the use of NRI merit special consideration. It is likely that these two may heavily influence the observed values of NRI. Moreover, it is possible to define NRI without the use of categories and properties of such metric should be explored. Finally, we need to address the effect of event rates on the magnitude of NRI, particularly in the context of calculations based on a validation sample.

In the following we propose a prospective formulation for the NRI which offers immediate application to survival and competing risk data as well as allows for easy weighting with observed or perceived costs. Furthermore, we address the issue of the number and choice of categories on NRI and contrast category-based NRI with one which is category-free. We discuss the issues of validation and event rate when models are applied to different samples or when incidence or prevalence and duration of follow-up vary between studies. We also show how NRI can be applied to case–control data. The concepts presented in the paper are illustrated in a Framingham Heart Study example with the same data and risk prediction model that was used in the original NRI paper [9] for easy comparisons.

2. Prospective form of NRI

Consider a situation in which predicted probabilities of a given event of interest come from two different risk prediction algorithms denoted here as 'new' and 'old'. Divide the predicted probabilities based on these two algorithms into a set of clinically meaningful ordinal categories of absolute risk and then cross-tabulate these two classifications. Define upward movement (*up*) as a change into higher category based on the new algorithm and downward movement (*down*) as a change in the opposite direction. The NRI is defined as:

$$\mathrm{NRI} = P(\mathrm{up}|\mathrm{event}) - P(\mathrm{down}|\mathrm{event}) + P(\mathrm{down}|\mathrm{non\text{-}event}) - P(\mathrm{up}|\mathrm{non\text{-}event}) \tag{1}$$

Using the Bayes rule we rewrite formula (1) in a different but equivalent form:

$$\begin{aligned}\mathrm{NRI} = &\frac{P(\mathrm{event}|\mathrm{up})\cdot P(\mathrm{up}) - P(\mathrm{event}|\mathrm{down})\cdot P(\mathrm{down})}{P(\mathrm{event})} \\ &+ \frac{P(\mathrm{non\text{-}event}|\mathrm{down})\cdot P(\mathrm{down}) - P(\mathrm{non\text{-}event}|\mathrm{up})\cdot P(\mathrm{up})}{P(\mathrm{non\text{-}event})}\end{aligned} \tag{2}$$

Substitute $P(\mathrm{non\text{-}event}) = 1 - P(\mathrm{event})$ and $P(\mathrm{non\text{-}event}|^*) = 1 - P(\mathrm{event}|^*)$, where * denotes 'up' or 'down', we obtain:

$$\begin{aligned}\mathrm{NRI} = &\frac{P(\mathrm{event}|\mathrm{up})\cdot P(\mathrm{up}) - P(\mathrm{event}|\mathrm{down})\cdot P(\mathrm{down})}{P(\mathrm{event})} \\ &+ \frac{(1 - P(\mathrm{event}|\mathrm{down}))\cdot P(\mathrm{down}) - (1 - P(\mathrm{event}|\mathrm{up}))\cdot P(\mathrm{up})}{1 - P(\mathrm{event})}\end{aligned} \tag{3}$$

After some simplification the above reduces to:

$$\mathrm{NRI} = \frac{(P(\mathrm{event}|\mathrm{up}) - P(\mathrm{event}))\cdot P(\mathrm{up}) + (P(\mathrm{event}) - P(\mathrm{event}|\mathrm{down}))\cdot P(\mathrm{down})}{P(\mathrm{event})\cdot(1 - P(\mathrm{event}))} \tag{4}$$

So far NRI (1) was presented as a sum of two components, one for events and one for non-events with weighting proportional to event incidence. This suggested a 'retrospective' interpretation of the index for incidence studies. The above expression changes to 'prospective' interpretation and allows for broad sets of definitions of reclassification that can be accommodated. This new formulation of the NRI requires a reclassification rule and then calculation of event rates among those who are reclassified upwards, downwards and the whole sample. It can be interpreted as a measure of event rate increase among those who are reclassified upwards and event rate decrease among those who are reclassified downwards.

The extension of formulas (3) or (4) to survival analysis is immediate, with $P(\text{event})$, $P(\text{event}|\text{up})$ and $P(\text{event}|\text{down})$ all estimated using the Kaplan–Meier approach. Proportions of people moving up and down (i.e. $P(\text{up})$ and $P(\text{down})$) are always available, so the computation of NRI according to formulas (3) or (4) is straightforward. Note that equations (3) and (4) also allow for application to competing risk models with Kaplan–Meier rates adjusted for the competing risk (cf. [19]).

Assuming that out of a total of n individuals, n_U are reclassified upwards and n_D downwards, (3) can be written as:

$$\text{NRI}=\frac{P(\text{event}|\text{up})\cdot n_U - P(\text{event}|\text{down})\cdot n_D}{n\cdot P(\text{event})}+\frac{(1-P(\text{event}|\text{down}))\cdot n_D-(1-P(\text{event}|\text{up}))\cdot n_U}{n\cdot(1-P(\text{event}))} \tag{5}$$

The quantities in the numerators represent expected numbers of events reclassified upwards and downwards (first numerator) and expected numbers of non-events reclassified downwards and upwards (second numerator). The denominators are the total expected cases of events and non-events, respectively. This is a representation analogous to the one proposed by Steyerberg and Pencina [15] for applications of NRI to survival models. Formula (5) does not depend on the number or even existence of risk categories as it assumes that probabilities of event among those reclassified upwards or downwards would be obtained pooling all individuals with the same reclassification. Read literally, the approach of Steyerberg and Pencina [15] assumed the existence of risk categories and proposed calculation of Kaplan–Meier estimators within each cell of the reclassification table. For large samples both approaches are equivalent; for smaller samples the one presented here might be preferable due to elimination of cells with very small numbers of events. As mentioned before, what we present here is more general, not requiring the existence of categories. We address this point further in Section 4.

3. Cost considerations and weighted NRI

In his commentary on the original NRI article, Greenland [17] suggested that the measure would be more meaningful if cost considerations were taken into account. Here, we propose a weighted form of the NRI (wNRI) which builds in cost considerations.

Assume that the savings associated with the upward reclassification of a person who eventually develops an event can be quantified by s_1, while the savings associated with downward reclassification of a person who does not develop an event can be quantified by s_2. In the simplest case of two risk categories s_1 can be viewed as the savings which result from starting treatment for a person bound to have an event and s_2 as savings from avoiding unnecessary treatment in non-events. When there are more than two categories, estimating s_1 and s_2 that are common for all upward and downward reclassification may not be possible and a finer partition of costs may be necessary leading to a definition that looks more like the one given in [15].

Here, however, the total savings expected from the use of the new rather than the old risk prediction algorithm can be calculated as:

$$s_1\cdot(P(\text{event}|\text{up})\cdot n_U - P(\text{event}|\text{down})\cdot n_D)+s_2\cdot(P(\text{non-event}|\text{down})\cdot n_D - P(\text{non-event}|\text{up})\cdot n_U)$$

Average savings per person are obtained by dividing the above by the total sample size, n, which produces a weighted form of the NRI, with weights related to cost savings:

$$\begin{aligned}\text{wNRI} = {} & s_1\cdot(P(\text{event}|\text{up})\cdot P(\text{up})-P(\text{event}|\text{down})\cdot P(\text{down}))\\ & +s_2\cdot(P(\text{non-event}|\text{down})\cdot P(\text{down})-P(\text{non-event}|\text{up})\cdot P(\text{up})),\end{aligned} \tag{6}$$

where we made use of the fact that $P(\text{up})=n_U/n$ and $P(\text{down})=n_D/n$.

The total savings in terms of those bound to experience events who now start treatment and those without events who now do not start treatment need to be compared to costs of measuring the marker. If costs are less than savings, then it is cost saving to measure the marker. If costs exceed savings, then further formal cost-effectiveness analysis is required.

We observe that taking $s_1=1/P(\text{event})$ and $s_2=1/P(\text{non-event})$, the above definition reduces to representation (2) of the previous section, a relationship already suggested in [12]. The expressions $1/P(\text{event})$ and $1/P(\text{non-event})$ may have little in common with true costs, especially if we think of cost as monetary, but it is of limited relevance, since what really matters is the ratio of s_1/s_2. We can always rescale wNRI or rescale the per-person cost of obtaining the new algorithm inputs over the cost of the old algorithm inputs.

In the case of two risk categories determined by one threshold, an alternative weighting might be proposed. If categories are sensibly defined, the decision theory suggests that the category threshold is the decision threshold [20, 21], which means weights are proportional to the sizes of the two intervals into which the (0, 1) interval is partitioned. For example, with the 0.20 threshold, 10-year risk $\geqslant$0.20 implies we take one decision (e.g. treatment with medication)

while <0.20 means only life-style intervention or no intervention at all. The relative weight of true positives to true negatives implied is then 0.80 to 0.20 or simply 4:1. Keeping with the convention of the original NRI, where the harmonic mean of weights s_1 and s_2 is equal to 1, we would take $s_1=5$ and $s_2=1.25$ in this example. Note that in the two-category case, if the threshold is equal to incidence (or prevalence), decision analytic and original weights are equal.

When there are more than two categories obtaining values of s_1 and s_2 that are not arbitrary and would be acceptable to a wide range of readers and reviewers constitutes a formidable task which usually will go beyond of what is expected from a medical research article. Thus the *ad hoc* 'statistical' costs suggested by $s_1=1/P(\text{event})$ and $s_2=1/P(\text{non-event})$, which imply that it is '$P(\text{non-event})/P(\text{event})$' times more important to reclassify upwards future event than it is to reclassify downwards future non-event, may not be unreasonable in many settings. Moreover, they afford a more objective measure that would not change from marker to marker, model to model or paper to paper, a property whose value should not be overlooked.

4. Problem of categories. Continuous NRI

The original article that introduced the NRI illustrated its application with a 3-category risk stratification [9]. However, others have applied it with 4 or no categories [13, 22, 23]. There is nothing implicit in the definition of the NRI which requires risk stratification into categories. The only requirement is that we define what upward and downward reclassification is.

While in some fields risk categories are firmly established and patient care depends on these categories (for example, primary prevention of CVD), other fields attempt to create meaningful risk categories but there is insufficient information to either justify or promote them (all cause mortality, diabetes, atrial fibrillation). Moreover, even when categories are firmly established (CVD prevention), their application is confused by different definitions of the endpoint of interest and thus different incidence rates for different models (hard CVD, full CVD, hard coronary heart disease (CHD), full CHD and so on). This can lead to different NRI values for the same marker added to different models.

What complicates matters further is the dependence of category-based NRI on the selection and number of categories. We illustrate this phenomenon with a very simple example. Assume 8 subjects, 4 events and 4 non-events with predicted probabilities of event based on a given old (and useless) model of 0.2, 0.4, 0.6 and 0.8 for events and 0.2, 0.4, 0.6 and 0.8 for non-events. Furthermore, assume that the addition of a new marker adds 0.16 to predicted probabilities for all event subjects and subtracts 0.16 for all non-event subjects. If we assume only two risk categories, below and above 0.5, the NRI equals $\frac{1}{4}+\frac{1}{4}=0.50$ (event subject with original probability 0.4 moves up and non-event subject with original probability 0.6 moves down). For NRI with three risk categories determined by cut-off points of 0.33 and 0.67 we get $\text{NRI}=\frac{2}{4}+\frac{2}{4}=1.00$ (event subjects with initial 0.2 and 0.6 move up and non-events with initial 0.4 and 0.8 move down). With four categories determined by cut-offs at 0.25, 0.50 and 0.75 we get $\text{NRI}=\frac{3}{4}+\frac{3}{4}=1.50$ and 'no category' NRI with upward and downward movement defined by any upward or downward change in predicted risks is equal to the maximum possible value of $1+1=2.00$. Similarly, it is not difficult to observe that NRI will also depend on the choice of categories. For this reason, it may not always be true in practice that more categories will mean higher NRI.

The above discussion suggests that the category-less or *continuous* NRI is the most objective and versatile measure of improvement in risk prediction. Its definition remains consistent with formulas (1)–(4) with the only difference in the meaning of upward and downward reclassification. In the following sections we show its alternative interpretations and invariance to changing event rates. We argue that in cases where no established categories exist, it is more prudent to use a version of NRI which does not require categories, rather than trying to create them for one particular application. Moreover, in cases where a priori categories do exist, it is still worth reporting the continuous NRI for comparison purposes with other applications.

In summary, two versions of NRI can be considered: one with categories which should be used if categories are already established in the field and influence care decisions and one without categories which can be used universally. We introduce the following notation:

(1) NRI(0.20) for two-category NRI with cut-off at 0.20;
(2) NRI(0.06,0.20) to denote NRI with three categories, established by cut-off points of 0.06 and 0.20;
(3) NRI(>0) or 'continuous NRI' for NRI with no categories;

Furthermore, 'event NRI' and 'non-event NRI' would indicate two useful subcomponents of the total NRI, with the former calculating the amount of correct reclassification among events and the latter among non-events. We recommend reporting these along with the single summary NRI for fuller interpretation:

$$\text{eventNRI}=P(\text{up}|\text{event})-P(\text{down}|\text{event})=\frac{P(\text{event}|\text{up})P(\text{up})-P(\text{event}|\text{down})P(\text{down})}{P(\text{event})}$$

$$\text{non-event NRI} = P(\text{down}|\text{non-event}) - P(\text{up}|\text{non-event}) = \frac{P(\text{non-event}|\text{down})P(\text{down}) - P(\text{non-event}|\text{up})P(\text{up})}{P(\text{non-event})}$$

Of course, NRI = event NRI + non-event NRI. We note that the original NRI was presented as a sum and not the average of the two subcomponents for 'historical reasons'—this way it matches the approach taken in the definition of IDI which in turn parallels the definition of Youden's index [24] and difference in logistic regression R-squares [10, 25]. However, an average ($\frac{1}{2}$NRI or $\frac{1}{2}$wNRI) could have an easier interpretation of average weighted improvement in classification (if categories are present) or in discrimination (if no categories are present).

In general we do not recommend using more than three categories unless they are already established and there is a justifiable need for that many. It seems to us that three categories offer sufficient categorization—high category for individuals with high risk (who should be treated), low category for those with low risk (who do not need treatment) and the middle category for everyone else. The use of categories can only be justified by explicit care recommendations for individuals in each category and it is often unlikely that these would materially differ between two middle categories (for example 0.05–0.10 and 0.10–0.20 in CVD prevention). If one feels a partition finer than three is needed, then the category-free NRI offers a better option.

We do realize that in some cases, category-based presentation may be more effective in terms of communication of results. Generally, we do not recommend an *ad hoc* creation of categories as they should be based on multiple factors, including cost considerations. However, if one feels they are absolutely necessary to convey the message and there are no categories established in the field, we recommend that categories are formed in a way which takes into account the event rate, severity of the disease under study and potential care recommendations based on the risk categories created.

We conclude this section with a comment about a modification of the 'original' NRI introduced by Cook *et al*. [26]. They define 'clinical NRI' as the amount of reclassification observed only in the 'middle risk' group. It is important to note that such 'clinical NRI' is meant to address a different question than the 'original NRI'. The 'original NRI' attempts to quantify the amount of improvement if the new marker was to be measured on everyone in the population of interest. 'Clinical NRI' quantifies the amount of improvement offered by a strategy in which only individuals in the middle risk group have the new marker obtained, have their risk recalculated based on a function which includes the new marker and are reclassified if the new probability leads to a different risk category. As these two NRIs are based on different groups of individuals they cannot be directly compared unless individuals in the high and low risk group who do not change categories in this strategy are included and 'clinical NRI' is translated into 'original NRI'. The latter approach might offer a more complete picture of the effect of the two-step strategy outlined above.

5. Relationships with other measures

In this section we show how in the binary case the two extreme versions of NRI, one without categories and one with only two categories, are related to the existing measures of diagnostic accuracy and model performance. First, we show how NRI(>0) is composed of the same building blocks as the difference in C statistics but arranged in a different way. Denote by Q predicted probabilities of event based on the 'new' risk prediction algorithm and by P based on the 'old' one. Then we have:

$$\begin{aligned} \tfrac{1}{2}\text{NRI}(>0) &= P(Q_i > P_i | i = \text{event}) - P(Q_j > P_j | j = \text{non-event}) \\ \Delta C &= P(Q_i > Q_j | i = \text{event}, j = \text{non-event}) - P(P_i > P_j | i = \text{event}, j = \text{non-event}) \end{aligned} \tag{7}$$

Equality (7) is proved in the Appendix. The comparisons used in the calculation of NRI(>0) are made between the two risk prediction rules but within event groups while for change in C statistics they are made within the two rules but between event groups. This illustrates the fact that NRI(>0) can be viewed as another measure of improvement in discrimination. Of note, similar to the C statistic, it is not affected by event incidence and thus can be compared across different studies.

The category-less NRI(>0) can also be viewed as the most appropriate summary measure for the 'reclassification plot' proposed by Steyerberg *et al*. [27]. It is constructed by plotting predicted probabilities based on the new risk prediction rule versus predicted probabilities based on the old rule, denoting events and non-events with different symbols. A 45° line of 'no change' is added to the graph for ease of visual inspection: for 'new' prediction rules which meaningfully improve reclassification, events are expected to lie above the 45° line (increase in predicted probability) and non-events below (decrease in predicted probability). An example of such plot for data presented in Section 8 is given in Figure 1.

Interestingly, the two-category (single cut-off) NRI is also related to the difference in C statistics, but this time these C statistics correspond to the areas under receiver operating characteristic (ROC) curves constructed in binary classification.

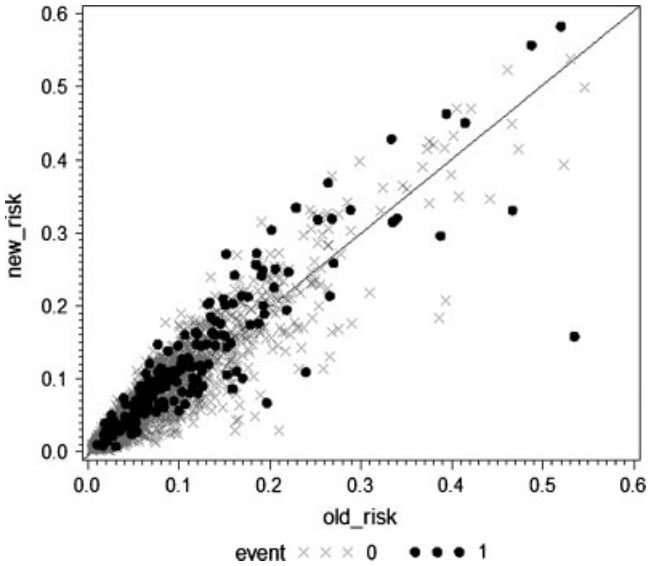

Figure 1. Reclassification plot.

For binary classification with threshold u, the area under the single-point ROC is equal to $\frac{1}{2}(\text{Sensitivity}(u)+\text{Specificity}(u))$ and hence, when comparing two models, we have:

$$\Delta C(u)=\tfrac{1}{2}(\Delta\text{Sensitivity}(u)+\Delta\text{Specificity}(u)).$$

At the same time:

$$\Delta\text{Sensitivity}(u)=P(\text{up}|\text{event})-P(\text{down}|\text{event}),$$

$$\Delta\text{Specificity}(u)=P(\text{down}|\text{non-event})-P(\text{up}|\text{non-event}),$$

and hence:

$$\Delta C(u)=\tfrac{1}{2}\Delta\text{Youden}(u)=\tfrac{1}{2}\text{NRI}(u),$$

where Youden(u) is used to denote the Youden [24] index for the cut-off u. Hence, the two-category NRI is equal to twice the difference in ROC areas for binary classification and to the increase in Youden index.

6. Differing event rates and validation

In Section 4 we alluded to the fact that category-based NRI is influenced by the relationship between category cut-offs and event rate. Hence, it may be misleading to apply the same fixed categories to events defined differently or time horizons of different durations which lead to varying incidence rates. This problem is absent when we use the category-free NRI which is not affected by event rates. When categories are present the situation is more complicated. First, one needs to ascertain that the endpoint of interest and duration of follow-up agree with those used to define the cut-off points. For example, in primary prevention of CVD the cut-off points of 6 and 20 per cent were established for hard coronary heart disease (defined as myocardial infarction or coronary death) with 10-year incidence rate of 0.09 for men and 0.03 for women (combined rate 0.06). When duration of follow-up is extended to 30 years or broader endpoint is considered, categories of 0.06 and 0.20 may no longer be meaningful. For example, 30-year incidence of hard CVD, which includes strokes, (Pencina *et al*. [16]) was 0.18 for men and 0.08 for women (combined rate 0.13). If category-based NRI was to be applied in this case, one might need to modify the categories. If proportionality of costs and event rates could be assumed, since the event rates are roughly 2:1 between 30-year HCVD and 10-year HCHD, categories of 0.12 and 0.40 could be an option. This is an *ad hoc* solution and further exploration is needed to assess its properties. On the other hand, a full cost effectiveness analysis would be required to formally justify the threshold selection. In summary, when using NRI with *a priori* established categories, one should not ignore the duration of follow-up and definition of the event of interest.

The above discussion has consequences for calculating NRI in a validation sample. It is not uncommon that the event rate of the validation sample will be a little different than in the development sample. This poses a potential

problem in distinguishing how much of the change in NRI is due to necessary correction for over-optimism and how much is contributed by slight misspecification of categories. The answer is straightforward for NRI(>0): all is due to the correction for over-optimism. The answer is not as simple for category-based NRIs—one might need to perform sensitivity analyses adjusting the categories. In general, the problem will be of smaller magnitude when NRI values are large (we have a useful new marker) or when the differences in incidence are small. At this point it might be helpful to observe that in some cases validated or cross-validated NRI may not be different or could even be higher than the one calculated on the development sample. This may be due to the issue with differing event rates raised above but also due to the fact that NRI is a measure of differences. It is conceivable that performance of models without and with the new marker goes down in the validation sample, but their difference is preserved or even goes up.

If an outside validation sample is not available, it is still important to correct for potential over-fitting. One way to accomplish this is through cross-validation of predicted probabilities based on the 'new' and 'old' risk prediction rules. A commonly used approach would randomly partition the sample into 5 or 10 equal subsamples (corresponding to 5 or 10-fold cross-validation), take out the first subsample and develop the prediction rules of interest on the remaining 4 or 9 subsamples combined. Then data from the subsample not used in development would be used to generate predicted probabilities according to the rule developed on all data except for this one subsample. The process is repeated for all subsamples. Thus we obtain a set of cross-validated predicted probabilities, two (based on the 'new' and 'old' rules) for each individual in the data set. These are to be used in the calculation of NRI. For smaller samples and to increase stability, one might repeat this process numerous times.

7. Case–control studies

So far our focus was primarily on NRI calculated using data coming from a study with prospective follow-up. However, many biomarker studies, especially at earlier stages of development, are conducted using the case–control design. It is natural to ask whether NRI can be calculated in these situations. Here, we present a method of extending NRI to case–control studies which use logistic regression as the analytic method. We adopt an approach analogous to the one used by Huang and Pepe for extending the predictiveness curve to case–control data [28].

NRI is based on predicted probabilities of event and event indicators. The latter are available 'by definition' in case–control studies but the former need some work. Predicted probabilities from a logistic regression conducted on case–control data are not meanigful as they do not represent the true risk for the population under study. However, the logistic regression coefficients are estimable in case–control data except for the intercept [29]. If we know the true disease incidence (or prevalence) we can adjust the case–control intercept to obtain a meaningful logistic regression model from which predicted probabilities of event can be easily derived.

Denote the logit of event probability in the case–control model with k predictors by

$$L(\beta_0, \beta_1, \ldots, \beta_k) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \tag{8}$$

where β_0 is the model intercept and $\beta_1, \ldots, \beta_k$ are regression coefficients for risk factors $X_1, \ldots, X_k$. Assume that in our sample there are n_E subjects who experienced the event and n_N who did not and the incidence or prevalence can be estimated based on a different sample and denoted by ρ. It can be shown [28] that adding $\log\left(\frac{\rho}{1-\rho}\cdot\frac{n_N}{n_E}\right)$ to the intercept of model (8) transforms it into a model that is scaled correctly for predicting risk with incidence or prevalenceρ. It is important to note that this method relies on the assumption that cases and controls are random (i.e. representative) samples of the underlying population. Otherwise, the intercept adjustment is invalid. This assumption might be violated in early case–control studies, where selection is often present, with extreme cases and very healthy controls constituting the analytic sample. When the above assumption is satisfied, we can suggest the following algorithm for estimating NRI from case–control data:

(1) Fit logistic two models to case–control data, one without and one with the new marker to obtain regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$ and $\beta'_0, \beta'_1, \ldots, \beta'_k, \beta'_{k+1}$.
(2) For every subject calculate adjusted predicted probabilities using the model without ('old') and with ('new') new marker as

$$p_{\text{old}} = \left(1+\exp\left(-L(\beta_0+\log\left(\frac{\rho}{1-\rho}\cdot\frac{n_N}{n_E}\right), \beta_1, \ldots, \beta_k)\right)\right)^{-1} \quad \text{and}$$

$$p_{\text{new}} = \left(1+\exp\left(-L(\beta'_0+\log\left(\frac{\rho}{1-\rho}\cdot\frac{n_N}{n_E}\right), \beta'_1, \ldots, \beta'_k, \beta'_{k+1})\right)\right)^{-1}$$

(3) Estimate NRI using p_{old} and p_{new} and event and non-event status in the case–control sample using sample estimates in formula (1).

Two observations need to be noted. First, the above procedure is not necessary for the category-less NRI(>0). Since the same constant is added to the logits of models without and with the new marker, and for each subject only the ordering of predicted probabilities matters, it is easy to show that this ordering remains unchanged after the adjustment. Thus NRI(>0) can be applied directly and remains meaningful for case–control data.

Second, the point estimator for NRI from adjusted case–control data should agree with the one that would have been obtained if prospective (or cross-sectional) data was available. However, the standard errors need not be the same due to different composition of the sample. Frequently, in case–control data the ratio of events to non-events is higher than in the true population. This means that the precision of estimation for events relative to non-events is higher in this type of case–control data. We note that in case–control as well as in any binary outcome data the asymptotic standard errors can be estimated as:

$$SE(\text{eventNRI}) = \sqrt{\frac{\hat{p}_{\text{up,events}} + \hat{p}_{\text{down,events}}}{n_E} - \frac{(\hat{p}_{\text{up,events}} - \hat{p}_{\text{down,events}})^2}{n_E}}$$

$$\text{SE}(\text{non-eventNRI}) = \sqrt{\frac{\hat{p}_{\text{down,non-events}} + \hat{p}_{\text{up,non-events}}}{n_N} - \frac{(\hat{p}_{\text{down,non-events}} - \hat{p}_{\text{up,non-events}})^2}{n_N}}$$

$$\text{SE}(\text{NRI}) = \sqrt{\text{SE}(\text{eventNRI})^2 + \text{SE}(\text{non-ventNRI})^2}$$

where the respective $\hat{p}$'s are estimated based on sample data as discussed by Pencina and D'Agostino *et al.* [9]:

$$\hat{p}_{\text{up,events}} = \frac{\#\text{events moving up}}{\#\text{events}}, \quad \hat{p}_{\text{down,events}} = \frac{\#\text{events moving down}}{\#\text{events}},$$

$$\hat{p}_{\text{up,non-events}} = \frac{\#\text{non-events moving up}}{\#\text{non-events}}, \quad \hat{p}_{\text{down,non-events}} = \frac{\#\text{non-events moving down}}{\#\text{non-events}}.$$

8. Practical applications

To illustrate the concepts described in the previous sections and to contrast the NRI from the original paper [9] with survival NRIs with and without categories, we will use the same data set. Some 3264 women and men aged 30–74 who attended the fourth Framingham Offspring cohort examination between 1987 and 1992 free of CVD were eligible for this analysis. All participants gave written informed consent and the study protocol was approved by the Institutional Review Board of the Boston Medical Center. The outcome of interest was 10-year incidence of coronary heart disease. Cox proportional hazards models were fit using sex, diabetes and smoking as dichotomous and age, systolic blood pressure (SBP), total (TCL) and HDL cholesterol (in one of the two models) as continuous predictors. C statistics for models without and with HDL as the new marker were calculated using the method described in [30]. The three NRIs of Section 4 were calculated on event probabilities obtained from the proportional hazards regression models. To correct for over-optimism we performed 10-fold cross-validation repeated 49 times and report median results. The 95% confidence intervals were estimated using 999 bootstrap replications following the approach outlined in [31]. NRI(0.20) was calculated with original and decision-analytic weights, as described in Section 3.

The cross-validated results are summarized in Tables I–III. We first observe that the 10-year incidence rate is 0.06, consistent with event rate used in the study that led to the derivation of 0.06 and 0.20 as cut-off points for risk categories. Hence these categories seem reasonable. We observe that in our example, NRI increases with increasing refinement of categories—it is the lowest when we look only at two categories: 4.5 per cent in Table I, and increases to 30.7 per cent for NRI without categories (Table III). This implies that one cannot compare NRIs with categories to NRIs without categories and should be clear which one is being reported. Of note, in our example, the category-based NRIs are attenuated more upon cross-validation than is the category-free NRI: without cross-validation the NRIs in Tables I–III would have been 6.1, 11.8 and 30.3 per cent, respectively. In fact, there is a marginal increase in cross-validated continuous NRI. We also note that the non-cross-validated survival-based NRI (0.06,0.20) of 11.8 per cent is very close to the 3-category NRI of 12.1% reported in the original paper [9] which ignored the survival nature of the data.

Tables I–III also give Kaplan–Meier (KM) rates for those reclassified up and down. In general, if the model is improved by the new marker, we would expect the KM rate for those reclassified upwards to exceed the overall KM rate

Table I. Reclassification table for two categories with cut-off point at 0.20.

	All	Reclassified upwards	Reclassified downwards	NRI	
Kaplan–Meier rate of CHD event	0.060	0.316	0.129		
Expected number of event subjects	196	12.9	4.1	Among event subjects	4.5 per cent
Expected number of non-event subjects	3068	28.1	27.9	Among non-event subjects	−0.0 per cent
Overall original (95 per cent bootstrap confidence interval)					4.5 per cent (0.2 per cent, 9.0 per cent)
Overall decision analytic (95 per cent bootstrap confidence interval)					1.2 per cent (−0.2 per cent, 2.7 per cent)

Table II. Reclassification table for two categories with cut-off points at 0.06 and 0.20.

	All	Reclassified upwards	Reclassified downwards	NRI	
KM rate of CHD event	0.060	0.141	0.057		
Expected number of event subjects	196	28.4	10.6	Among event subjects	9.1 per cent
Expected number of non-event subjects	3068	172.6	175.4	Among non-event subjects	0.1 per cent
Overall original (95 per cent bootstrap confidence interval)					9.2 per cent (2.8 per cent, 16.0 per cent)

Table III. Reclassification table with no categories.

	All	Reclassified upwards	Reclassified downwards	NRI	
KM rate of CHD event	0.060	0.078	0.043		
Expected number of event subjects	196	122.1	73.9	Among event subjects	24.6 per cent
Expected number of non-event subjects	3068	1440	1628	Among non-event subjects	6.1 per cent
Overall original (95 per cent bootstrap confidence interval)					30.7 per cent (15.6 per cent, 45.2 per cent)

and the rate for those who moved down to be below the overall KM rate. This is true in Tables II and III: for example, in Table II the KM rate for those moving up is $0.151 > 0.060$ while KM rate for those going down is $0.041 < 0.060$. When this relationship does not hold, NRI is likely to be smaller (cf. Table I). Even though the magnitudes of the 3 NRIs presented are different, when we compare event NRI and non-event NRI across all examples, we notice that in all of them the event NRI is larger than the non-event NRI, which might suggest that addition of HDL cholesterol helps increase predicted risk for those who experience events (in other words, 'catch events') to a larger degree than it does decrease predicted risk for those without events ('catch non-events').

In Table I we note that wNRI(0.20) with decision analytic weighting (weights 5.0 for events and 1.25 for non-events) is markedly lower than NRI(0.20) with original weights based on incidence (weights 16.7 for events, 1.1 for non-events). Our observation indicates how sensitive NRI is to the selection of weight. As mentioned in Section 3, in decision analysis, the choice of threshold reflects the weight: 0.20 is synonymous with the 4:1 weighting of importance of events versus non-events [20] and thus no other options would be considered (including weighting based on incidence). As observed above, in the HDL example presented here, the improvement in reclassification is more pronounced for events than for non-events. Moreover, we required the harmonic and not arithmetic mean of weights to be fixed, which resulted in event weights of 5.0 for decision analytic and 16.7 for original NRI(0.20) which explains the difference in magnitude. The important lesson from our example is that NRIs weighted differently are different quantities and need to be interpreted on their own scales and cannot be compared. It also gives further support to the recommendation to report event and non-event NRIs in addition to the combined measure.

Using the numbers from Table III, we can illustrate the relationship given in (7): on one hand $\frac{1}{2}\text{NRI}(>0)=\frac{1}{2}\cdot 0.307=0.153$ and on the other $P(Q_i>P_i|i=\text{event})-P(Q_j>P_j|j=\text{non-event})=\frac{122.1}{196}-\frac{1440}{3068}=0.622-0.469=0.153$. In contrast, the corresponding difference in C statistics is only $0.760-0.751=0.009$. This illustrates, albeit on different scales, how much more sensitive NRI(>0) is as a measure of improvement in discrimination.

9. Conclusions

In this paper we developed a general form for the net reclassification improvement which presents NRI as a prospective measure which quantifies the correctness of upward and downward reclassification or movement of predicted probabilities as a result of adding a new marker. The new form offers immediate extension to survival and competing risk data and allows for building in cost considerations. We have also contrasted NRIs which use categories with NRI that does not as well as NRIs which apply original versus decision analytic weights. We have shown that the category-less NRI(>0) which defines upward and downward movement as any change in the predicted probabilities is a measure of discrimination that is not influenced by correct scaling of the model and for binary data can be expressed in terms similar to the AUC. This makes NRI(>0) immediately applicable in validation sample and in case–control data. On the other hand category-based NRIs are measures of reclassification and are influenced by event rates and risk cut-offs. If established risk thresholds exist and treatment decisions are made based on risk categories, NRI which uses these categories can be useful. However, care needs to be taken to correctly define the event of interest and duration of follow-up. We have suggested a few possible ideas applicable to both observational and case–control data.

NRI is designed to quantify improvement in performance and hence its magnitude is more important than statistical significance. For this reason we recommend presenting NRI with its confidence interval rather than relying on p-values. Further research is needed to determine meaningful or sufficient degree of improvement as well as to suggest preferred methods to construct asymptotic or bootstrap confidence intervals.

Our theoretical considerations and practical examples have shown that NRI depends on the number and choice of categories (with no categories being one of the choices) and the weighting used. For this reason NRIs cannot be compared across studies unless they are defined in the same manner. In particular, the category-less NRI(>0) offers the widest and most standardized application. If presenting other versions, it is essential that researchers clearly define and justify their choices.

Appendix A

Here we prove equality (7): $\frac{1}{2}\text{NRI}(>0)=P(Q_i>P_i|i=\text{event})-P(Q_j>P_j|j=\text{non-event})$.

Assuming predicted probabilities follow a continuous distribution and any movement is considered meaningful (implying that every person has to move either up or down) we obtain

$$P(Q_i>P_i|i=\text{event})+P(Q_i<P_i|i=\text{event})=1 \text{ or equivalently:}$$

$$P(\text{up}|\text{event})+P(\text{down}|\text{event})=1 \text{ implying:}$$

$$P(\text{up}|\text{event})-P(\text{down}|\text{event})=2\cdot P(\text{up}|\text{event})-1.$$

Similarly: $P(\text{down}|\text{non-event})-P(\text{up}|\text{non-event})=1-2\cdot P(\text{up}|\text{non-event})$.

Thus:

$$\text{NRI}(>0)=P(\text{up}|\text{event})-P(\text{down}|\text{event})+P(\text{down}|\text{non-event})-P(\text{up}|\text{non-event})$$

$$=2\cdot P(\text{up}|\text{event})-1+1-2\cdot P(\text{up}|\text{non-event})=2\cdot(P(\text{up}|\text{event})-P(\text{up}|\text{non-event})).$$

Acknowledgements

This work was supported by the NIH/ARRA Risk Prediction of Atrial Fibrillation (1 RC1HL101056) and the National Heart, Lung, and Blood Institute's Framingham Heart Study (contract N01-HC-25195).

References

1. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care. *Circulation* 2008; **117**:743–753.
2. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation* 1998; **97**:1837–1847.
3. D'Agostino RB, Wolf PA, Belanger A, Kannel WB. Stroke risk profile: adjustment for antihypertensive medication. *Stroke* 1994; **25**:40–43.
4. Schnabel RB, Sullivan LM, Levy D, Pencina MJ, Massaro JM, D'Agostino RB Sr, Newton-Cheh C, Yamamoto JF, Magnani JW, Tadros TM, Kannel WB, Wang TJ, Ellinor PT, Wolf PA, Vasan RS, Benjamin EJ. Development of a risk score for atrial fibrillation (Framingham Heart Study): a community-based cohort study. *Lancet* 2009; **373**:739–745.
5. 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**:1879–1886.
6. Parikh NI, Pencina MJ, Wang TJ, Benjamin EJ, Lanier KJ, Levy D, D'Agostino Sr RB, Kannel WB, Vasan RS. A risk score for predicting near-term incidence of hypertension: the Framingham heart study. *Annals of Internal Medicine* 2008; **148**:102–110.
7. Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, Manning AK, Florez JC, Wilson PWF, D'Agostino RB, Cupples LA. Genotype score in addition to common risk factors for prediction of type 2 diabetes. *NEJM* 2008; **359**:2208–2219.
8. Cook NR. Use and misuse of the receiver operating characteristics curve in risk prediction. *Circulation* 2007; **115**:928–935.
9. 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.
10. Pepe MS, Feng Z, Gu JW. Commentary on 'Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond'. *Statistics in Medicine* 2008; **27**:173–181.
11. Uno H, Tian L, Cai T, Kohane IS, Wei LJ. Comparing risk scoring systems beyond the ROC paradigm in survival analysis. *Harvard University Biostatistics Working Paper Series* 2009; paper 107; accessed online on November 26, 2009.
12. Pencina MJ, D'Agostino Sr RB, D'Agostino Jr RB, Vasan RS. Comments on integrated discrimination and net reclassification improvements—practical advice. *Statistics in Medicine* 2008; **27**:207–212.
13. Ingelsson E, Schaefer EJ, Contois JH, McNamara JR, Sullivan L, Keyes MJ, Pencina MJ, Schoonmaker C, Wilson PW, D'Agostino RB, Vasan RS. Clinical utility of different lipid measures for prediction of coronary heart disease in men and women. *JAMA* 2007; **298**:776–785.
14. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. *Annals of Internal Medicine* 2009; **150**:795–802.
15. Steyerberg EW, Pencina MJ. Reclassification calculations with incomplete follow-up. *Annals of Internal Medicine* 2010; **152**:195–196.
16. Pencina MJ, D'Agostino Sr RB, Larson MG, Massaro JM, Vasan RS. Predicting the 30-year risk of cardiovascular disease. The Framingham heart study. *Circulation* 2009; **119**:3078–3084.
17. Greenland S. Evaluating the added predictive ability of a new marker: the need for reorientation toward cost-effective prediction. *Statistics in Medicine* 2008; **27**:199–206.
18. Vickers AJ, Elkin EB, Steyerberg E. Net reclassification improvement and decision theory. *Statistics in Medicine* 2009; **28**:525–526.
19. Gaynor JJ, Feuer EJ, Tan CC, Wu DH, Little CR, Straus DJ, Clarkson BD, Brennan MF. On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data. *Journal of the American Statistical Association* 1993; **88**:400–409.
20. Peirce CS. The numerical measure of the success of predictions. *Science* 1884; **4**:453–454.
21. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* 2006; **26**:565–574.
22. Ridker PM, Paynter NP, Rifai N. C-reactive protein and parental history improve global cardiovascular risk prediction: the Reynolds risk score for men. *Circulation* 2008; **118**:2243–2251.
23. Harrell FE. ImproveProb() routine in R statistical software. Accessed November 26, 2009.
24. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; **3**:32–35.
25. Hu B, Palta M, Shao S. Properties of *R*-square statistics for logistic regression. *Statistics in Medicine* 2006; **25**:1383–1395.
26. Cook NR. Comments on '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.
27. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models. A framework for traditional and novel measures. *Epidemiology* 2010; **21**:128–138.
28. Huang Y, Pepe M. Semiparametric methods for evaluating risk prediction markers in case–control studies. *Biometrika* 2009; **96**:991–997.
29. Breslow NE. Statistics in epidemiology: the case–control study. *JASA* 1996; **91**:14–28.
30. Pencina MJ, D'Agostino RB. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicne* 2004; **23**:2109–2123.
31. Hosmer DW, Lemeshow S. Confidence interval estimates of an index of quality performance based on logistic regression models. *Statistics in Medicine* 1995; **14**:2161–2172.