Statistics
in Medicine

Research Article

Received: 5 January 2017, Accepted: 20 April 2017 Published online 19 June 2017 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7333

Asymptotic distribution of ΔAUC, NRIs, and IDI based on theory of U-statistics

Olga V. Demler,[a*†] Michael J. Pencina,[b] Nancy R. Cook[a] and Ralph B. D'Agostino Sr[c]

The change in area under the curve (ΔAUC), the integrated discrimination improvement (IDI), and net reclassification index (NRI) are commonly used measures of risk prediction model performance. Some authors have reported good validity of associated methods of estimating their standard errors (SE) and construction of confidence intervals, whereas others have questioned their performance. To address these issues, we unite the ΔAUC, IDI, and three versions of the NRI under the umbrella of the U-statistics family. We rigorously show that the asymptotic behavior of ΔAUC, NRIs, and IDI fits the asymptotic distribution theory developed for U-statistics. We prove that the ΔAUC, NRIs, and IDI are asymptotically normal, unless they compare nested models under the null hypothesis. In the latter case, asymptotic normality and existing SE estimates cannot be applied to ΔAUC, NRIs, or IDI. In the former case, SE formulas proposed in the literature are equivalent to SE formulas obtained from U-statistics theory if we ignore adjustment for estimated parameters. We use Sukhatme–Randles–deWet condition to determine when adjustment for estimated parameters is necessary. We show that adjustment is not necessary for SEs of the ΔAUC and two versions of the NRI when added predictor variables are significant and normally distributed. The SEs of the IDI and three-category NRI should always be adjusted for estimated parameters. These results allow us to define when existing formulas for SE estimates can be used and when resampling methods such as the bootstrap should be used instead when comparing nested models. We also use the U-statistic theory to develop a new SE estimate of ΔAUC.

Keywords: AUC; NRI; IDI; risk prediction; U-statistics

1. An introduction and a motivating example

In current medical research, risk prediction is viewed as an objective way to assess the risk of a patient to develop a disease and is often used by clinicians in making treatment decisions. The Framingham [1] and ATP III models for 10-year risk of cardiovascular outcomes [2] and the Gail model for 5-year risk of breast cancer [3] are among the first widely used risk prediction models. Moreover, in recent years, risk-prediction models have played an increasingly important role in medical decision making and have been directly incorporated into updates of existing treatment guidelines. For instance, the U.S. Preventive Services Task Force recently issued updated guidelines on aspirin use in prevention of cardiovascular events [4]. Based on the results of a microsimulation model, that used the American College of Cardiology/American Heart Association (ACC/AHA) risk equations for 10-year cardiovascular disease risk [5]. Therefore, the quality of the performance of a risk prediction model is often crucial for assigning the most beneficial treatment and making correct policy decisions.

Risk prediction models are often evaluated in terms of calibration and discrimination. Discrimination measures how well a given model separates events from non-events; calibration measures the closeness of the model-based and observed risks of the outcome. The area under the receiver operating characteristics curve (AUC of ROC) [6,7] is a widely used measure of discrimination. In 2008, several new intuitively appealing measures of discrimination were introduced such as the net reclassification index (NRI) and integrated discrimination improvement (IDI) [8,9].

[a] *Division of Preventive Medicine, Brigham and Women's Hospital, 900 Commonwealth Avenue, Boston, MA, 02115, U.S.A.*
[b] *Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, 27708, U.S.A.*
[c] *Department of Mathematics and Statistics, Boston University, 111 Cummington Mall, Boston, MA, 02215, U.S.A.*
**Correspondence to: Olga V. Demler, Division of Preventive Medicine, Brigham and Women's Hospital, 900 Commonwealth Avenue, Boston, MA 02115, U.S.A.*
† *E-mail: olgademler@gmail.com*

They rapidly gained popularity and at the time of writing this paper had been referenced more than 2800 times. Simple estimators for variance and asymptotic distributional behavior were proposed to allow construction of confidence intervals.

While some papers reported good validity of the methods for confidence intervals and variance estimators of ΔAUC, NRIs, and IDI [8,10,11], others questioned their performance [10,12–14]. To illustrate these conflicting views, we ran some simulations and summarize the results in Table I. For two nested models with binary outcome and multivariate normal predictor variables, we compare observed and theoretical standard errors of ΔAUC, three types of NRIs (continuous ($NRI_{>0}$), 2-category NRI at event rate threshold (NRI(r)) and 3-category NRI (3cNRI)), and IDI. AUC is a measure of discrimination. It is equal to the probability that the risk of a randomly picked event is greater than for randomly picked non-event [6,7]. ΔAUC measures improvement in quality of discrimination between events and non-event by the new model relative to the old one [11]. $NRI_{>0}$, another measure of discrimination, calculates the difference between fractions of correct and incorrect movements of predicted probabilities among events and adds to it a similar quantity calculated for non-events [9]. Categorical NRIs are similar to $NRI_{>0}$ but consider only movements across categories. NRI(r) uses two categories defined by event rate threshold [15]. 3cNRI uses three categories defined by any thresholds [16]. IDI combines average change in probabilities among events and among non-events [8]. For comparison, we included in Table I the regression coefficient (β) for the new predictor variable $\boldsymbol{x}_2$. The relative bias of standard error estimate is calculated as $\frac{(\text{theoretical se–observed se})}{\text{observed se}}100\%$. Shaded areas in Table I indicate scenarios in which the relative bias is 5% or more in our simulations, while white areas indicate when standard errors have very low bias (<5%). Asymptotic theory developed for three of the five statistics performed very well in most situations, while the bias of the 3cNRI is comparable with that of the standard error estimator of the Kaplan–Meier survival probability (when sample size is small) [17], and the standard error estimator of the IDI has the strongest bias of the five statistics.

Confidence intervals for ΔAUC, $NRI_{>0}$, categorical NRIs, and IDI proposed to date rely on asymptotic normality [8,9,11,18,19]. In Figure 1, we show an example in which the IDI can be asymptotically normally distributed under the alternative hypothesis of meaningful effect (left panel) and right-skewed under the null hypothesis of no meaningful effect (right panel) [20].

This paper is a validity study of previously proposed asymptotic distribution results of ΔAUC, IDI, and three types of NRIs (continuous ($NRI_{>0}$), 2-category NRI at event rate threshold (NRI(r)), and 3cNRI) [8,9,11,18,19] when comparing two nested models. Using U-statistics theory, we explicitly specify conditions when asymptotic results are valid and when resampling methods such as the bootstrap should be used instead. These results help us disentangle several reports of the asymptotic distribution and performance of variance estimators of ΔAUC, IDI, and three types of the NRI. The paper is structured as follows: Notation is introduced in Section 2; the main result is stated and proved in Section 3; in Sections 4 and 5, we apply theoretical findings to the Framingham Heart Study (FHS) data; and the implications of these findings are discussed in Section 6.

Table I. Relative bias (%) of standard errors of ΔAUC, $NRI_{>0}$, 3-category NRI, and IDI.

			Relative bias (%) of the standard error estimate of				
Effect size	Size	β_{x_2}	ΔAUC	$NRI_{>0}$	NRI(r)	3cNRI	IDI
0	30,000	3	7	29	−7	−7	−31
0	2,000	0	8	26	−12	−13	−37
0	500	0	8	26	−12	−13	−37
0.2	30,000	3	1	−3	1	−5	−42
0.2	2,000	1	−1	−3	−1	−10	−45
0.2	500	1	−1	−3	−1	−10	−45
0.7	30,000	2	0	−1	0	−26	−38
0.7	2,000	1	3	−2	1	−26	−39
0.7	500	1	3	−2	1	−26	−39

($rel.bias = \frac{se_{formula.based} - se_{bootstrap}}{se_{bootstrap}} \times 100\%$). We evaluated the performance of two nested risk prediction models: a logistic regression model with two multivariate normally distributed predictor variables (x_1 and x_2) and a baseline logistic regression model with only one of the predictors (x_1). We considered several simulations scenarios: effect size by the new predictor (x_2) of 0, 0.2, and 0.7; effect size by commonly used predictor variable (x_1) is 0.7; sample sizes of 30,000, 2,000, and 500 observations; 0.1 event rate; and B = 1,000 simulated datasets. 2% and 10% cutoffs were used for 3cNRI calculation.

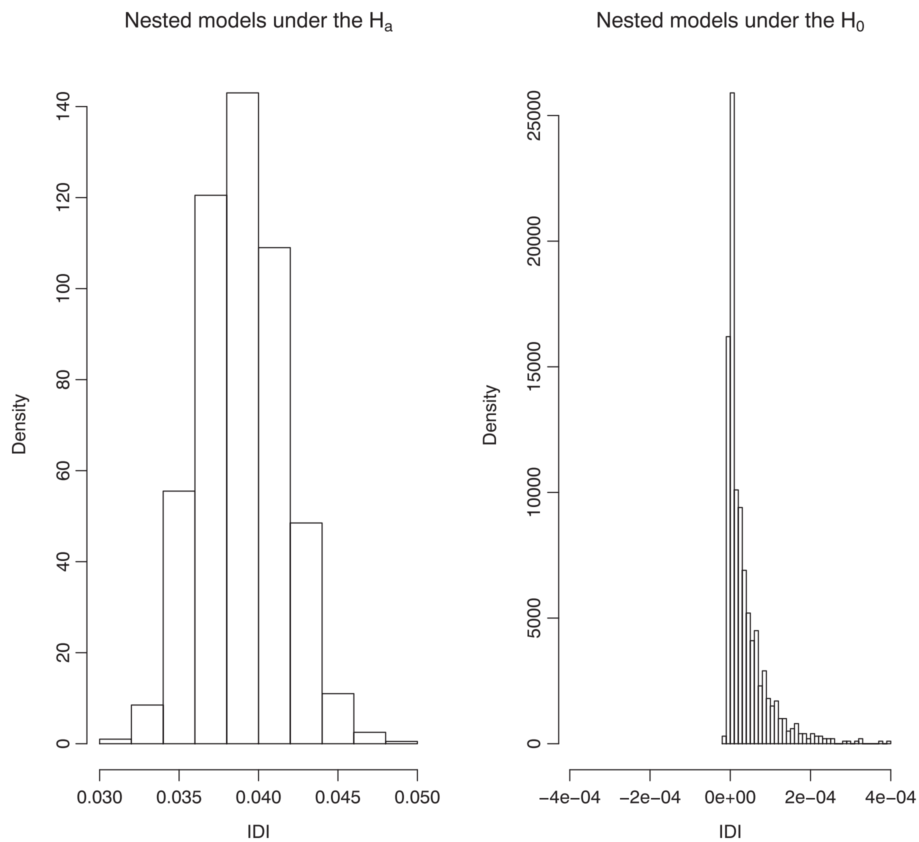

Figure 1. Histograms of IDI when comparing nested models under the alternative (left panel) and under the null (right panel). x_1,x_2 are predictors from the full model; $x_1$is the predictor from the reduced model. Left panel: simulated nested models under the alternative $x_1,x_2 \mid D = 1 \sim N(\mu,\Sigma)$ and $x_1,x_2 \mid D = 0 \sim N(0,\Sigma)$. Right panel: simulated nested models under the null $x_1 \mid D = 1 \sim N(\mu,\sigma 2)$, $x_1 \mid D = 0 \sim N(0,\sigma^2)$, and $x_2 \sim N(0,\sigma^2)$. x_2 is an uninformative predictor.

2. Notation

Let D be an outcome of interest, with $D = 1$ for events and $D = 0$ for non-events. Our goal is to predict the event status using p predictor variables. Conditioning on the event status, predictor variables follow two (potentially different) distribution functions: $\boldsymbol{x}|D = 0 \sim \mathbf{F}(.)$, $\boldsymbol{y}|D = 1 \sim \mathbf{G}(.)$. Assume that for each of N patients, their disease status D and vector of predictor variables are available. There are n_0 non-events and n_1 events. The prediction based on the full set of p predictor variables is to be compared with that based on a reduced number of predictor variables, $p - 1$. We assume that the linear model is true and that one of the linear models for binary outcome is employed (logistic regression, linear discriminant analysis (LDA), etc). We use this model to estimate linear coefficients in order to combine multiple predictor variables into one metric, the risk score. Unless otherwise specified, we assume that the models are nested, so the new model adds k new predictors to the old model. The regression technique of choice produces coefficients estimates $\boldsymbol{a}^{*\prime} = \left(a_1^*, \ldots, a_{p-k}^*, 0, \ldots, 0\right)$ (reduced model) and $\boldsymbol{a}' = (a_1 \ldots, a_p)$ (full model). Corresponding risk scores are calculated as $\boldsymbol{a}'\boldsymbol{x}$ and $\boldsymbol{a}^{*\prime}\boldsymbol{x}$ for non-events and $\boldsymbol{a}'\boldsymbol{y}$ and $\boldsymbol{a}^{*\prime}\boldsymbol{y}$ for events, with the symbol * always denoting the reduced model. We sought to test whether the risk prediction model with p predictors performs better than the model with only the first p-k predictors. We will consider ΔAUC, three varieties of the NRI, and the IDI as measures of model performance. They are often used in current medical research on risk prediction. Analysis of their performance, advantages, and disadvantages is an active area of methodological research on risk prediction. In the following, we review standard formulas [8,9,11] for ΔAUC, continuous NRI ($NRI_{>0}$), 3cNRI, NRI(r), and IDI.

2.1. ΔAUC

The area under ROC curve (AUC) can be interpreted as the probability that the risk score of a randomly picked event is higher than a randomly picked non-event. The AUC is estimated by the Mann–Whitney statistic [6,7] – a non-parametric unbiased estimator, often referred to as the *c*-statistic [21,22] and can be written as follows:

$AUC = \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} I\left[\boldsymbol{a}'\boldsymbol{x_i} < \boldsymbol{a}'\boldsymbol{y_j}\right]$, where $I[.]$ is the indicator function.

The AUC for the reduced model is as follows: $AUC^* = \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} I\left[\boldsymbol{a}^{*\prime}\boldsymbol{x_i} < \boldsymbol{a}^{*\prime}\boldsymbol{y_j}\right]$.

Then ΔAUC is as follows:

$$\Delta AUC = \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} I\left[\boldsymbol{a}'\boldsymbol{x_i} < \boldsymbol{a}'\boldsymbol{y_j}\right] - \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} I\left[\boldsymbol{a}^{*'}\boldsymbol{x_i} < \boldsymbol{a}^{*'}\boldsymbol{y_j}\right]$$

ΔAUC is one of the most widely used measures of discrimination.

2.2. Continuous NRI ($NRI_{>0}$)

$\text{NRI}_{>0}$ [9] is the difference of proportions of individuals with events and non-events whose predicted probabilities moved up:

$$\begin{aligned} NRI_{>0} &= \frac{\sum_{i=1}^{n_1} Sign[p_{new\ ev} - p_{old\ ev}]}{n_1} - \frac{\sum_{i=1}^{n_0} Sign[p_{new\ nonev} - p_{old\ nonev}]}{n_0} \\ &= \frac{\#events\ up}{n_1} - \frac{\#nonevents\ up}{n_0} \end{aligned}$$

2.3. Three-category NRI

Three-category NRI [16] is very close to the original definition of categorical NRI [9] but takes into account the size of the jump from category to category (number of categories moved). It is defined as follows:

$$3cNRI = \frac{1}{n_1}\sum_{i=1}^{n_1} \#categories\ up_i - \ \#categories\ down_i - \frac{1}{n_0}\sum_{j=1}^{n_0} \#categories\ up_j - \ \#categories\ down_j$$

This definition of categorical NRI is preferable over its original 2008 version [8], because of several attractive properties [16], including the fact that 3cNRI = 0 if marginal cells of the reclassification table stay the same for the two models. By using weights, it treats jumps across one versus two categories differently, and the event rate has a limited impact on the magnitude of the 3cNRI. Therefore, it successfully resolves several criticisms of the original definition of categorical NRI [23,24].

2.4. NRI at the event rate (NRI(r))

In their 2016 paper, Pencina, Steyerberg, and D'Agostino [15] investigate the properties of a 2-category NRI with categories defined by the proportion of cases in the sample (r) and show that it has several advantages: like ΔAUC, it is invariant to the event rate and has intuitive interpretation as the proportion of correct reclassifications.

2.5. IDI

IDI [8] is defined as

$$IDI = \frac{\sum_{i=1}^{n_1} p_{new\ ev\ i} - p_{old\ ev\ i}}{n_1} - \frac{\sum_{i=1}^{n_0} p_{new\ nonev\ i} - p_{old\ nonev\ i}}{n_0}$$

Discrimination improvement is related asymptotically to the rescaled Brier score and to the difference in discrimination slope [25]. We mentioned some criticisms of IDI earlier, and in the following, we address some of them.

Now we can formulate the following null hypotheses for the six statistics defined earlier:

$$\begin{array}{ll} H_0^{AUC}: \quad \Delta\text{AUC} = 0 \quad \text{vs} \quad H_a^{AUC}: \quad \Delta\text{AUC} \neq 0 \\ H_0^{NRI}: \quad \text{NRI}_{>0} = 0 \quad \text{vs} \quad H_a^{NRI}: \quad \text{NRI}_{>0} \neq 0 \\ H_0^{NRI}: \quad \text{NRI(r)} = 0 \quad \text{vs} \quad H_a^{NRI}: \quad \text{NRI(r)} \neq 0 \\ H_0^{3cNRI}: 3c\text{NRI} = 0 \quad \text{vs} \quad H_a^{3cNRI}: 3c\text{NRI} \neq 0 \\ H_0^{IDI}: \quad \text{IDI} = 0 \quad \text{vs} \quad H_a^{IDI}: \quad \text{IDI} \neq 0 \end{array} \tag{1}$$

Pepe *et al*. [26,27] showed that each of the five hypotheses in (1) is equivalent to testing the significance of the set of the new predictors in the new regression model (2).

$$H_0 : \text{a}_{p-k+1}, \ldots, \text{a}_p = 0 \text{ vs } H_a : \text{a}_{p-k+1}, \ldots, \text{a}_p \neq 0 \tag{2}$$

Therefore, when we consider data under the null, we can without loss of generality assume that the null is formulated in terms of non-significance of the linear coefficient by the new predictor variable, that is, the hypothesis in (2).

3. Main result

We formulate our main results as follows.
ΔAUC, $\text{NRI}_{>0}$*, NRI(r), 3cNRI, and IDI:*

STATEMENT 1. *are generalized U-statistics with estimated parameters.*
STATEMENT 2. *belong to non-degenerate subclass if and only if they compare any non-nested models or nested models under the alternative hypothesis in (2). As non-degenerate U-statistics,*

a. *they follow normal distribution asymptotically.*
b. *Available variance formulas are algebraically equal to the variance estimators provided by U-statistics theory if we ignore adjustment for estimated parameters.*
c. *Variance of ΔAUC, $NRI_{>0}$, and NRI(r) does not need to be adjusted for estimated parameters if predictor variables are normally distributed.*
d. *Variance of IDI and 3-category NRI should always be adjusted for estimated parameters.*

STATEMENT 3. *belong to the* ***degenerate subclass*** *if and only if they compare nested models under the null hypothesis in (2). As degenerate U-statistics, they do not follow normal distribution and available variance estimators do not apply for them.*

3.1. *ΔAUC, $NRI_{>0}$, NRI(r), 3CNRI, and IDI belong to the U-statistics family*

In Appendix B, we prove Statement 1 showing that statistics considered in this paper belong to a U-statistics family [28]. Rigorous asymptotic distribution theory of U-statistics has been developed by Hoeffding [29], Lehman [30], Sukhatme [31], and others. The form of the U-statistics' distribution depends on whether the U-statistics are degenerate. Non-degenerate U-statistics are normally distributed, and formulas for their standard errors are available. Degenerate U-statistics are distributed as an infinite sum of weighted Chi-square random variables, and derivation of their standard error is challenging.

In Appendix B, we show that ΔAUC, $\text{NRI}_{>0}$, NRI(r), 3cNRI, and IDI are degenerate if and only if they compare nested models under the null. In all other situations, they belong to the non-degenerate class of U-statistics. Degeneracy and non-degeneracy conditions are listed in Table II.

Degenerate and non-degenerate U-statistics form very different classes in terms of their asymptotic behavior. In the following sections, we will consider these two situations separately.

3.2. *Non-degenerate case*

ΔAUC, NRIs, and IDI are non-degenerate if they evaluate the performance of two non-nested models or of nested models under the alternative. This is the most practically interesting case because only in this

Table II. Non-degeneracy conditions of ΔAUC, $\text{NRI}_{>0}$, NRI(r), 3cNRI, and IDI.

	Models are under the null	Models are under the alternative
Nested models	Degenerate*	Non-degenerate
Non-nested models	Always non-degenerate	

*Null is defined as in (2) in the previous section. $H_0: a_{p-k+1}, \ldots, a_p = 0$.

situation we need to construct confidence intervals for ΔAUC, NRIs, and IDI. Hoeffding [29] and Lehman [30] showed that non-degenerate U-statistics are asymptotically normally distributed. U-statistics theory also provides their variance formulas [28] but notes that variances should be adjusted for estimated parameters. Adjustment has been studied by Sukhatme [32], Randles [33], and de Wet [34] and is summarized in [28].

3.3. Available variance estimators are identical to U-statistics theory-based variance estimators if we ignore an adjustment for estimated parameters

In the Appendix, we derived variances of ΔAUC, NRIs, and IDI based on U-statistics theory, ignoring adjustment for estimated parameters and presented them in Table III. The standard errors of $\text{NRI}_{>0}$ and NRI(r) based on the U-statistics theory are exactly the same as the ones derived by Pencina *et al*. in [10,11]. The standard error formula for ΔAUC is new. It is equal to the variance of the change in ranks. This representation is more intuitive, but it assumes no tied ranks.

U-statistics theory adds one more layer to variance calculations, namely that when U-statistic relies on estimated parameters, its variance in general should be adjusted for estimated parameters. In many cases ΔAUC, NRIs, and IDI rely on estimated parameters (linear coefficients of regression models), their variances may need to be adjusted for estimated parameters, or we need to show that such adjustment is not necessary. In the following section, we prove that for some of the statistics under certain assumptions, adjustment for estimated parameters is unnecessary.

3.4. Variances of ΔAUC, $\text{NRI}_{>0}$, and NRI(r) do not need to be adjusted for estimated parameters if predictor variables are normally distributed

Sometimes, adjustment for estimated parameters can be avoided. Sukhatme [32], Randles [33], and de Wet [34] showed that adjustment for estimated parameters is unnecessary if and only if a certain condition is met [28]. In the following, we check this condition and show that under normality of predictor variables, standard error estimates of ΔAUC, continuous NRI, and NRI(r) do not need to be adjusted for estimated parameters.

Table III. Variance formulas in non-degenerate case, unadjusted for estimated parameters.

	$\widehat{\sigma^2}$, ignoring the adjustment for estimated parameters	Requires adjustment?
$\hat{\sigma}^2_{\Delta AUC}$ no tied ranks	$\frac{Var\left(\text{rank}_e^*(\boldsymbol{a}^*\boldsymbol{T}\boldsymbol{x_i})-\text{rank}_e(\boldsymbol{a}^T\boldsymbol{x_i})\right)}{n_0}+\frac{Var\left(\text{rank}_{ne}^*(\boldsymbol{a}^*\boldsymbol{T}\boldsymbol{y_j})-\text{rank}_{ne}(\boldsymbol{a}^T\boldsymbol{y_j})\right)}{n_1}$	No
$\hat{\sigma}^2_{\Delta AUC}$ tied ranks	Use DeLong formula [11]	No
$\hat{\sigma}^2_{NRI_{>0}}$	$\frac{\hat{p}_{ne}^{up}\left(1-\hat{p}_{ne}^{up}\right)}{n_0}+\frac{\hat{p}_{e}^{up}\left(1-\hat{p}_{e}^{up}\right)}{n_1}$	No
$\hat{\sigma}^2_{NRI(r)}$	$\frac{\hat{p}_{ne}^{up}+\hat{p}_{ne}^{down}-\left(\hat{p}_{ne}^{up}-\hat{p}_{ne}^{down}\right)^2}{n_0}+\frac{\hat{p}_{ev}^{up}+\hat{p}_{ev}^{down}-\left(\hat{p}_{ev}^{up}-\hat{p}_{ev}^{down}\right)^2}{n_1}$	No
$\hat{\sigma}^2_{3cNRI}$	$\frac{4\left(\hat{p}_{ne}^{2up}+\hat{p}_{ne}^{2down}\right)+\hat{p}_{ne}^{1up}+\hat{p}_{ne}^{1down}-\left(2\left(\hat{p}_{ne}^{2up}-\hat{p}_{ne}^{2down}\right)+\hat{p}_{ne}^{1up}-\hat{p}_{ne}^{1down}\right)^2}{n_0}+\frac{4\left(\hat{p}_{ev}^{2up}+\hat{p}_{ev}^{2down}\right)+\hat{p}_{ev}^{1up}+\hat{p}_{ev}^{1down}-\left(2\left(\hat{p}_{ev}^{2up}-\hat{p}_{ev}^{2down}\right)+\hat{p}_{ev}^{1up}-\hat{p}_{ev}^{1down}\right)^2}{n_1}$	Yes
$\hat{\sigma}^2_{IDI}$	$\frac{Var(\Delta predp(\boldsymbol{x_i}))}{n_0}+\frac{Var\left(\Delta predp\left(\boldsymbol{y_j}\right)\right)}{n_1}$	Yes

STATEMENT 2.C *If ΔAUC, $NRI_{>0}$, and NRI(r) when comparing nested models are non-degenerate (Table* II*) and if predictor variables are normally distributed, then standard errors of ΔAUC, continuous NRI, and NRI(r) do not need to be adjusted for estimated parameters.*

Proof
We restate here the condition for adjustment for estimated parameters:

> *Suhkatme–Randles–de Wet condition:*
> *Standard errors for a U-statistic with estimated parameters do not need to be adjusted for estimated parameters if and only if the derivative of the expected value of the U-statistic with respect to parameters is zero.*

For example, for ΔAUC, this condition is written as $\frac{\partial}{\partial a}E[\Delta AUC] = 0$.

In our assumptions, predictors are normally distributed; therefore, LDA is the most efficient way to estimate regression coefficients [35]. Su and Liu [36] also showed that under these assumptions, LDA coefficients maximize Mahalanobis distance [37] (M^2) between risk scores of events and non-events. Therefore, the gradient of Mahalanobis distance with respect to parameters is zero. When comparing nested models, ΔAUC is a function of the Mahalanobis distances ($\Delta AUC = \Phi\left(\sqrt{\frac{M_p^2}{2}}\right) - \Phi\left(\sqrt{\frac{M_{p-k}^2}{2}}\right)$) [36], where p is the number of predictor variables in a model. Hence, the gradient of ΔAUC with respect to parameters is zero as well. Therefore, the standard error of ΔAUC under the assumption of normality of predictor variables does not need to be adjusted for estimated parameters.

Similarly, we can use a closed-form formula for $NRI_{>0}$ [38] for nested models: ($NRI_{>0} = 4\Phi\left(\frac{\sqrt{M_p^2 - M_{p-k}^2}}{2}\right) - 2$ to show that gradient of $NRI_{>0}$ is also zero at the LDA coefficients. Therefore, the standard error of $NRI_{>0}$ also does not need to be adjusted for estimated parameters.

Pencina, Steyerberg, and D'Agostino [15] showed that NRI(r) under normality assumptions when comparing nested models can be written as follows:

$NRI(r) = 2\cdot\left(\Phi\left(\frac{\sqrt{M_p^2}}{2}\right) - \Phi\left(\frac{\sqrt{M_{p-k}^2}}{2}\right)\right)$. The same reasoning can be applied to $NRI(r)$ to show that $\frac{\partial}{\partial a}NRI(r) = 0$. Therefore, NRI(r) does not need to be adjusted for estimated parameters under the assumptions of this statement.

Q.E.D.

STATEMENT 2.D *Variances of IDI and 3cNRI should always be adjusted for estimated parameters.*

Note that the IDI and 3cNRI also can be expressed in closed form under normality of predictor variables [16,38] (please see the Appendix), but their closed-form expression does not rely exclusively on the Mahalanobis distance. It also depends on the estimated rate of events, which becomes one of the parameters. Under normality of predictor variables, LDA solution maximizes Mahalanobis distance, and therefore, the derivative of M_p^2 with respect to regression parameters is zero. However, there is no such result for partial derivative of the closed-form formulas of IDI and 3cNRI with respect to event rate. Derivatives of closed-form formulas of 3cNRI and IDI with respect to event rate were calculated in Appendix D. Both derivatives are nonlinear in r, and both are in general non-zero. For example, derivatives of 3cNRI and IDI are 2.02 and 1.04 correspondingly for event rate observed in FHS of 7.65%, when comparing models with Mahalanobis distances of 0.7 and 0.8 and using 5% and 7.5% cutoffs to calculate 3cNRI. Therefore, the Sukhatme–Randles–deWet condition is not satisfied for IDI and 3cNRI, and standard errors of IDI and 3cNRI should be adjusted for estimated parameters.

Our empirical results in Table I support the main theoretical results proven in this paper. Variances of ΔAUC, $NRI_{>0}$, and NRI(r) calculated from unadjusted formulas have on average very small relative bias compared with those of the IDI and 3cNRI whose variances must be adjusted for estimated parameters.

Also, the top three rows of Table I are calculated for the degenerate case (when comparing two nested models under the null). All five statistics are degenerate, and theoretical formulas for their variance estimator are not applicable: Existing variance formulas have strong bias for all five statistics when comparing nested models under the null.

To illustrate further the main theoretical findings of this paper, we simulated a binary model with normally distributed predictor variables. In Figure 2, we plot histograms of ΔAUC, NRIs, and IDI calculated for nested models under the alternative and overlay two normal distribution curves with empirical (dotted line) and theoretical (solid) variances. ΔAUC, $NRI_{>0}$, and NRI(r) are in the top row. They do not need to be adjusted for estimated parameters, and the dotted and solid curves almost completely overlap. IDI and 3cNRI are in the bottom row. They require adjustment for estimated parameters, and the two curves do not overlap because the theoretical variance is an incorrect estimate of the actual variance of 3cNRI and IDI.

3.5. Statement 2.C and 2.D for logistic regression and non-normal data

We showed in the proof of Statement 2.C that by estimating parameters with LDA, we ensured that Sukhatme–Randles–deWet condition holds true. What would happen if we had used logistic regression to estimate parameters instead of the LDA? To use theoretical variance formulas, we need to show that adjustment for parameters estimated by logistic regression is not required. Therefore, we need to satisfy the Sukhatme–Randles–deWet condition. Parameter estimates produced by logistic regression and the LDA are both consistent under assumption of normality [35]; therefore, when sample size is sufficiently large, the two estimates are very close. In Table I, we used logistic regression to estimate parameters for

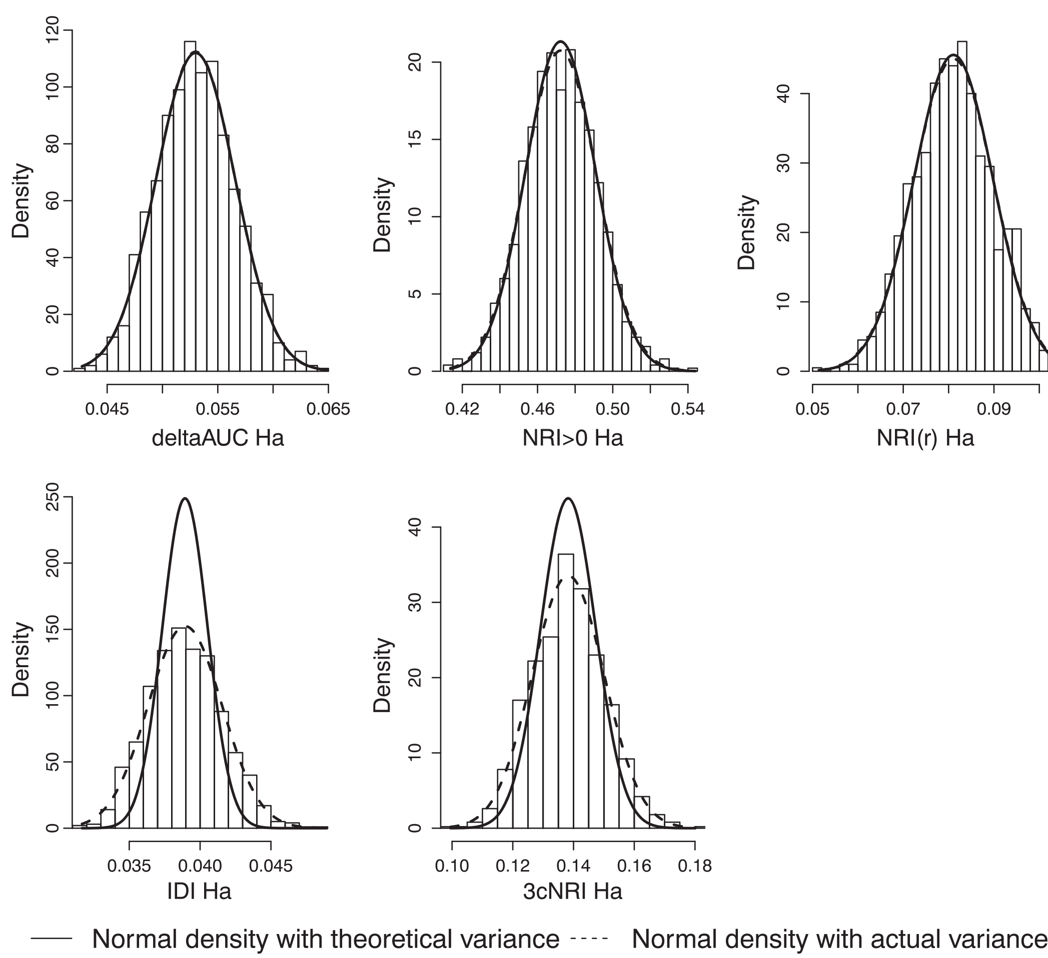

Figure 2. Two normal density curves with empirical (dotted line) and theoretical from Table II (solid line) variances overlaid on the histograms of the five statistics calculated for nested models under the alternative (non-degenerate case). Simulated two predictor variables and binary outcomes: x_1,x_2 |D = 1 ~ N(μ,Σ) and x_1, x_2 |D = 0 ~ N(0,Σ). x_1,x_2 are predictors from the full model; $x_1$is the predictor from the reduced model.

simulated normal data. Table I supports the theoretical findings of Statement 2.C and 2.D despite the use of logistic regression.

The proof of Statement 2.C and the discussion earlier rely on normality of predictor variables. An important question is how sensitive these results are to the normality assumption. In Section 4, we apply the results of this section to real-life non-normal data using logistic regression and discuss the implications.

3.6. Degenerate case

In the Appendix, we show that when comparing nested models under the null, ΔAUC, NRIs, and IDI belong to a degenerate class of U-statistics. They are distributed as an infinite sum of weighted Chi-square distributions. Histograms in Figure 3 demonstrate why any test that assumes normality is invalid for ΔAUC and IDI.

3.7. Injecting random noise to remedy degeneracy

In previous sections, we discussed problems induced by the degenerate state of ΔAUC, NRIs, and IDI when they compare nested models under the null. Their asymptotic distribution and variance estimators become practically intractable. In their non-degenerate state, ΔAUC, NRIs, and IDI follow a normal distribution asymptotically, and variance formulas are available. In this section, we show how degeneracy is at the root of the problem. We will artificially move ΔAUC, NRIs, and IDI away from degeneracy and show that their distribution functions shift to normal distribution. This will shed some

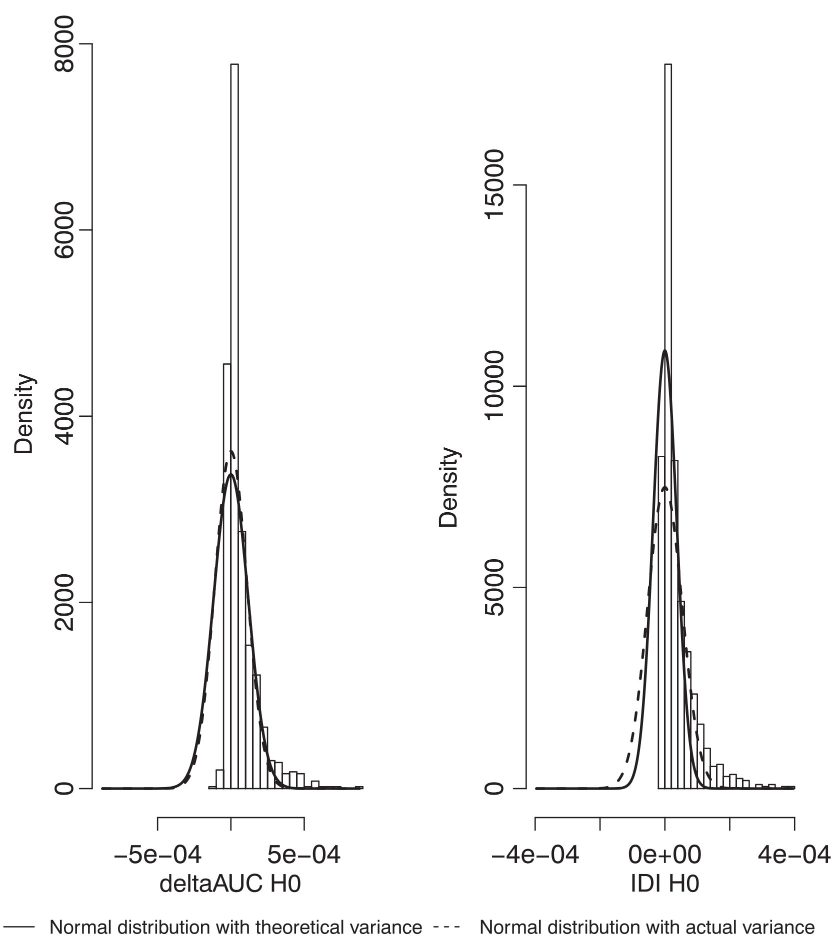

Figure 3. Histograms of ΔAUC and IDI when comparing nested models under the null. Two normal density curves with empirical (dotted line) and theoretical (solid line) variances overlaid on the histograms of ΔAUC and IDI calculated for nested models under the null (degenerate case). Simulated two predictor variables and binary outcome: x_1 |D = 1 ~ N(μ, σ^2), x_1 |D = 0 ~ N(0, σ^2), and x_2 ~ N(0,σ^2). x_2 is an uninformative predictor. x_1,x_2 are predictors from the full model; $x_1$is the predictor from the reduced model.

light on other aspects of NRI behavior that we will discuss later in the section. In the Appendix, the degeneracy condition is formulated in mathematical terms, and it follows that the nested models under the null are the fundamental reason for the degeneracy of ΔAUC, $NRI_{>0}$, IDI, and all categorical versions of the NRI. So let us consider two nested models under the null. ΔAUC, NRIs, and IDI calculated for these two models will be in a degenerate state. To force them to move away from the degeneracy, we need to violate the degeneracy condition: one way is to force the models away from the null, and an alternative way is to un-nest them. In practical situations, we have no control over a model being under the null or under the alternative. However, we can try to un-nest the two models by injecting random noise, that is, add a weak predictor to the smaller model and another independent weak predictor of the same strength to the other model. Histograms of these statistics for the same models but with injected noise are in the right column of Figure 4. Their distributions shift to asymptotic normality. Results for variance estimators hold in this example too: Variance estimate of ΔAUC is still satisfactory, and the variance of the IDI is underestimated by existing formulas. Our simulations indicate that de-degenrating these two U-statistics comes at the price of a substantial increase of variance and leads to a loss of power. However, this exercise helps to explain why the distribution of $NRI_{>0}$, NRI(r), and 3cNRI appears more Gaussian for the degenerate state in our simulations (Figure 5). The IDI can be written as follows: $IDI = \frac{\sum_{i=1}^{n_1} p_{new\ ev} - p_{old\ ev}}{n_1} - \frac{\sum_{i=1}^{n_0} p_{new\ nonev} - p_{old\ nonev}}{n_0}$. The $NRI_{>0}$ uses the same definition as IDI but dichotomizes the change in predictive probability:

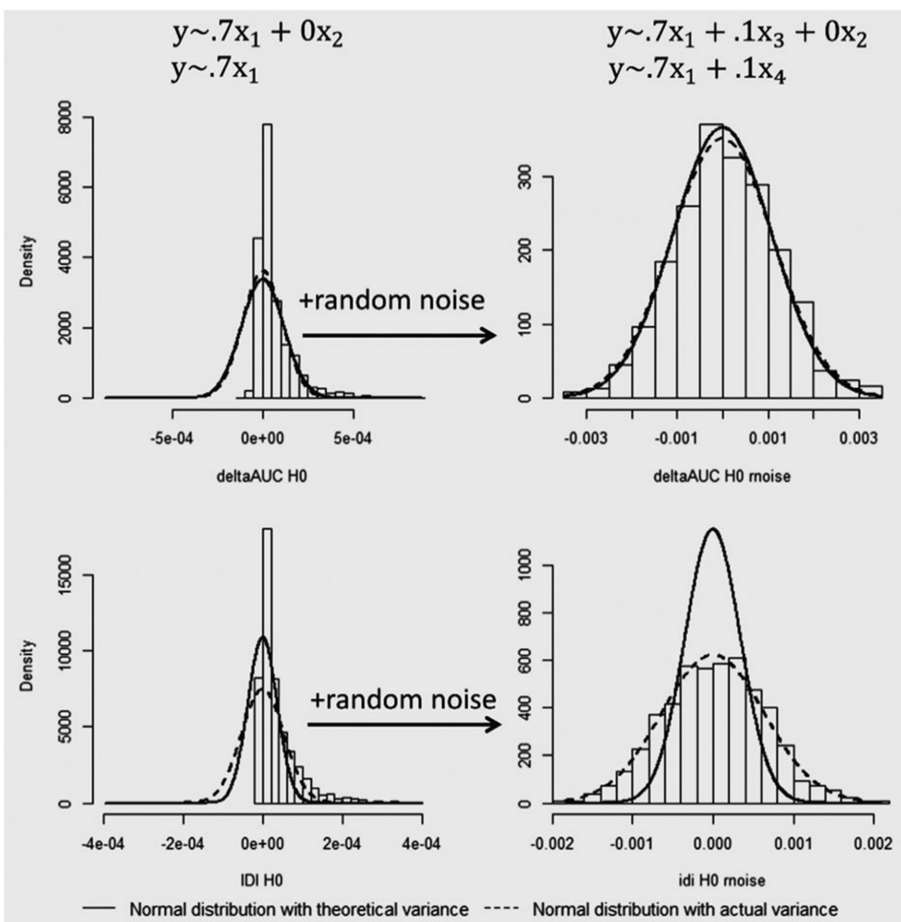

Figure 4. Left column: two nested models under the H0; right column: the two models after un-nesting, preserving the H_0. Left panel models: x_1, x_2 are predictors from the full model; $x_1$is predictor from the reduced model. $x_1 \mid D = 1 \sim N(\mu, \sigma^2)$, $x_1 \mid D = 0 \sim N(0, \sigma^2)$, and $x_2 \sim N(0, \sigma^2)$. x_2 is an uninformative predictor. Right panel models: x_1,x_2, x_3 are predictors from the full model; x_1,x_4 are predictors from the reduced model. $x_1 \mid D = 1 \sim N(\mu,\sigma\text{^}2)$, $x_1 \mid D = 0 \sim N(0, \sigma^2)$, and $x_2 \sim N(0, \sigma^2)$. $x_{3,4} \mid D = 1 \sim N(\varepsilon, I\sigma^2)$ and $x_{3,4} \mid D = 0 \sim N(0,I\ \sigma^2)$. x_2 is an uninformative predictor, and x_3, x_4 are added 'noise' – independent simulated weak predictors.

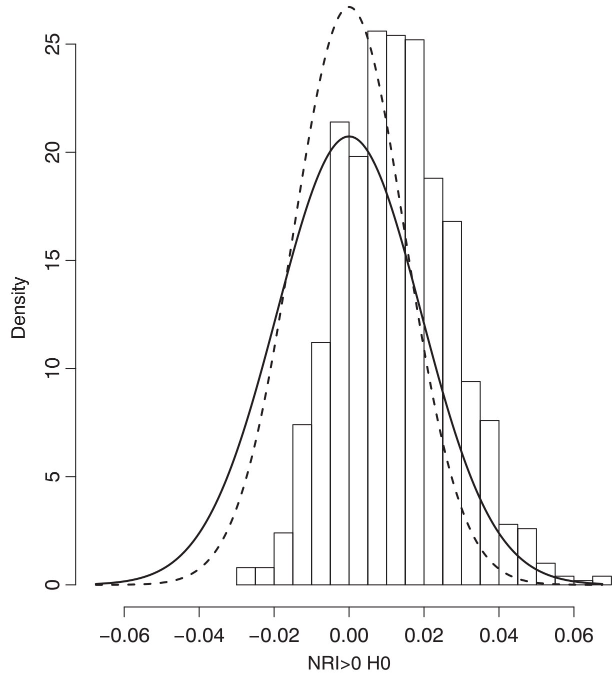

Figure 5. Histogram of NRI > 0 under H0. Simulated two predictor variables and binary outcomes: x_1 | D = 1 ~ N(μ,σ^2), x_1 |D = 0 ~ N(0,σ^2), and x_2 ~ N(0,σ^2). x_2 is an uninformative predictor. x_1,x_2 are predictors from the full model; x_1 is the predictor from the reduced model.

$$NRI_{>0} = \frac{\sum_{i=1}^{n_1} Sign[p_{new\ ev} - p_{old\ ev}]}{n_1} - \frac{\sum_{i=1}^{n_0} Sign[p_{new\ nonev} - p_{old\ nonev}]}{n_0}$$

Therefore, we can view the $\mathrm{NRI}_{>0}$ as an IDI that adds to each summand a random component that complements it to the nearest of the values of 1 or −1. This random component operates as injected noise in Figure 4. It adds enough noise so that $\mathrm{NRI}_{>0}$ transitions to non-degeneracy and its histogram looks Gaussian, even when adding a predictor variable of interest (x_2) does not improve the performance of the model (Figure 5). Note that $\mathrm{NRI}_{>0}$ remains biased. Its bias is studied in [39].

4. Practical example

We apply our results to FHS [1,40] data. Full information about this data set and the study including the enrollment criteria is reported in [40]. Briefly, 8365 people free of cardiovascular disease at baseline examination were followed for 12 years. The outcome of interest was coronary heart disease (CHD), and 640 people developed CHD during follow-up (7.7%). Predictor variables in this example include age, total and high-density lipoprotein cholesterol, systolic and diastolic blood pressure, baseline diabetes status, and current smoking. All continuous variables are log-transformed. We use logistic regression to run the full model with all the predictors. We also ran a series of smaller nested models, which we obtained by omitting from the full model one of the predictor variables.

The bootstrap estimator of the standard error is consistent for a wide range of statistics under mild regularity conditions [41–43]. Therefore, we can use the bootstrap estimate of the standard error of ΔAUC, $\mathrm{NRI}_{>0}$, and IDI as a proxy for the gold standard, that is, as an estimator with established consistency. For this reason, we define the relative bias of the formula-based standard error as the difference between the average of a formula-based and bootstrap-based variance estimates divided by the bootstrap-based variance estimate.

In this practical example, all predictors are statistically significant; therefore according to results of this paper, ΔAUC, $\mathrm{NRI}_{>0}$, and IDI are non-degenerate U-statistics, and according to Statement 2.C, we would expect low bias of the theoretical standard error formulas for ΔAUC, $NRI_{>0}$, and NRI(r) and high bias for those that require adjustment for estimated parameters: 3cNRI and IDI.

5. Results

Relative bias of the standard error was calculated for FHS data using bootstrap as described in the previous section. Results are presented in Table IV.

As we anticipate, the two statistics that require adjustment for estimated parameters (IDI and 3cNRI) have a stable strong bias in Table III. However, contrary to our expectations, bias of the theoretical standard error estimates of the three statistics that should not require adjustment for estimated parameters (ΔAUC, $NRI_{>0}$, and NRI(r)) varies greatly. For example, the DeLong formula for standard error of ΔAUC often underestimates it by as much as 23% and the formula for $NRI_{>0}$ by as much as 56%. Statement 2 is proved under assumption of normally distributed predictors, and this result is consistent with empirical simulations in Table I. But some of the simulations with real-life data in Table IV still show substantial bias. To further explore this phenomenon, we first check the stability of our results in Table III. We replicate bootstrap analysis several times with the FHS data set but with different random seed. Relative bias remains present across replications. Second, we use the result obtained by Harrell *et al*. [21], that is, that tests of c-index (a survival analysis version of AUC) have very low power. We hypothesize that ΔAUC, $NRI_{>0}$, and NRI(r) experience similar loss of power. We observed in our simulations that transition to non-degeneracy is gradual (Appendix Figure A1), so lack of power may be explained by degenerate behavior of the ΔAUC, NRIs, and IDI even for moderately strong predictor variables; therefore, we cannot use standard error formulas developed under the assumption of non-degeneracy. This reasoning implies that if we artificially inflate the strength of the added predictor variable, ΔAUC, $NRI_{>0}$, and NRI(r) should move further away from the null and the relative bias of standard error estimates of ΔAUC, $NRI_{>0}$, and NRI(r) will go down. Standard error estimates of 3cNRI and IDI have another problem: they require adjustment for estimated parameters. This problem cannot be solved by artificial inflation of effect size, so we expect bias of their standard error estimates

Table IV. Relative bias (bottom table) (rel.bias $= \frac{se_{formula.based} - se_{bootstrap}}{se_{bootstrap}} \times 100\%$) and averaged bootstrap estimates were calculated by bootstrapping FHS dataset by simple random sampling without replacement. The full model included baseline values of age, TCL, HDL, SBP,DBP, diabetes status, and current smoking. The first row compares the full model to the same model without SBP as a predictor. The row 'AGE' compares the full model to the same model but with age omitted from the list of predictors.

	Parameter estimates					
	β[a]	ΔAUC	$NRI_{>0}$	NRI(r)	3cNRI[c]	IDI
SBP	1.39	0.00	0.03	0.01	0.00	0.00
HDL	−1.65	0.03	0.44	0.06	0.12	0.02
TCL	1.57	0.01	0.25	0.02	0.06	0.01
AGE	2.73	0.02	0.46	0.06	0.16	0.01
DBP	1.05	0.00	0.20	0.01	0.01	0.00
SMOKING	0.50	0.01	0.16	0.02	0.05	0.00
DIABETES	0.55	0.00	−0.10	0.00	0.00	0.00
	Relative bias (%) of standard error estimate of					
	z-score(β)[b]	ΔAUC	$NRI_{>0}$	NRI(r)	3cNRI[c]	IDI
SBP	3.05	−23	−3	−24	−24	−45
HDL	−10.93	−2	−6	−8	−28	−38
TOT	6.70	−2	−4	−13	−26	−29
AGE	10.41	9	−5	−6	−34	−36
DBP	2.16	20	−36	−19	−36	−16
SMOKING	5.62	−13	−1	−16	−29	−27
DIABETES	3.61	−3	−56	−16	−19	−43

[a]β is the linear coefficient by added predictor variable (larger model). Continuous predictors were log-transformed but not standardized.
[b]z–scores of β coefficients (β/se(β)).
[c]About 2% and 10% cutoffs were used to calculate 3cNRI.

to stay strong. Table V shows the results of the bootstrap for the same data as in Table IV, but with artificially inflated effect sizes of added predictor variables.

Because we have artificially forced predictor variables away from the null, results presented in Table V now support Statement 2. As expected, formula-based standard error estimates of ΔAUC, $NRI_{>0}$, and NRI(r) have low bias, and 3cNRI and IDI have high bias because the latter group requires adjustment for estimated parameters.

In Figure 6, we illustrate the relationship between relative bias of formula-based standard error and effect size of the added predictor.

Results of this bootstrap analysis using real-life data suggest that Statement 2 is sensitive to the assumption of non-degeneracy. Statistical significance at the 0.05 level of added predictor variable is not sufficient to guarantee non-degeneracy, and associations with stronger effect sizes are required for asymptotic formulas to become consistent. In our example, when *p*-values of added predictor variables are weaker than 10^{-5}, ΔAUC, $NRI_{>0}$, and NRI(r) are too close to degeneracy. Figure A1 in the Appendix illustrates very slow gradual transition away from degeneracy of ΔAUC as the added predictor variable gets stronger. That is, the distribution of ΔAUC is still non-normal when the z-score of the added predictor variable is less than 4.0 (p-value $\leq 6{\cdot}10^{-5}$). Much stronger effect sizes are needed to achieve non-degeneracy. This observation explains why formula-based standard error estimators of $NRI_{>0}$, NRI(r), and ΔAUC are biased in Table IV when *p*-values of the added predictor variable are less than 0.05 but greater than 10^{-5}.

Figure 6 illustrates that in FHS data, as the added predictor variable gets stronger, bias of standard error estimates of ΔAUC, $NRI_{>0}$, and NRI(r) decreases. With z-scores of beta coefficient ≥4.0 relative bias of formula-based standard error estimates of $NRI_{>0}$, NRI(r) falls below 15% while standard error of ΔAUC is still overestimated by the asymptotic formula and requires an even stronger predictor to lower its relative bias below 15%. When the z-score of the added predictor in nested models framework is less than 4.0 (p-value $>10^{-5}$), standard errors of ΔAUC, $NRI_{>0}$, and NRI(r) should be estimated using resampling methods. Electronic Health Records, pooled genetic cohorts, social networks data, etc. can result in very large sample sizes and potentially very low *p*-values. For such large sample sizes, traditional resampling technique can become time consuming. Our results show that in this situation, formula-based standard error estimates of ΔAUC, $NRI_{>0}$, and NRI(r) may have low bias and may be estimated by using the formulas presented in Table II. Table V implies that bias of added dichotomous predictors may remain strong in all scenarios. Standard errors of 3cNRI and IDI always require adjustment for estimated

Table V. Analysis presented in Table III was repeated, but effect size of each added predictor variable was artificially inflated. rel.bias $= \frac{se_{formula.based} - se_{bootstrap}}{se_{bootstrap}} \times 100\%$. Effect size of dichotomous variables was inflated by artificially increasing their prevalence among events.

	Parameters Estimates					
	β	ΔAUC	$NRI_{>0}$	NRI(r)	3cNRI	IDI
SBP	21.70	0.20	1.50	0.44	0.95	0.43
HDL	5.63	0.17	1.16	0.34	0.70	0.26
TCL	14.35	0.21	1.63	0.48	1.08	0.55
AGE	2.72	0.02	0.46	0.06	0.16	0.01
DBP	36.90	0.22	1.83	0.55	1.23	0.77
SMOKING	1.34	0.04	0.56	0.09	0.18	0.04
DIABETES	2.32	0.06	0.62	0.06	0.12	0.11
	Relative Bias (%) of Standard Error Estimate of					
	zscore(β)	ΔAUC	$NRI_{>0}$	NRI(r)	3cNRI	IDI
SBP	29.46	−2	−12	0	−36	−26
HDL	27.46	2	−7	−2	−37	−25
TCL	27.90	0	−7	−7	−38	−22
AGE	10.41	7	−4	−8	−35	−35
DBP	22.58	1	−9	−2	−40	−27
SMOKING	14.63	−2	−2	−3	−24	−39
DIABETES	21.34	0.02	−0.23	0.08	−0.11	−0.38

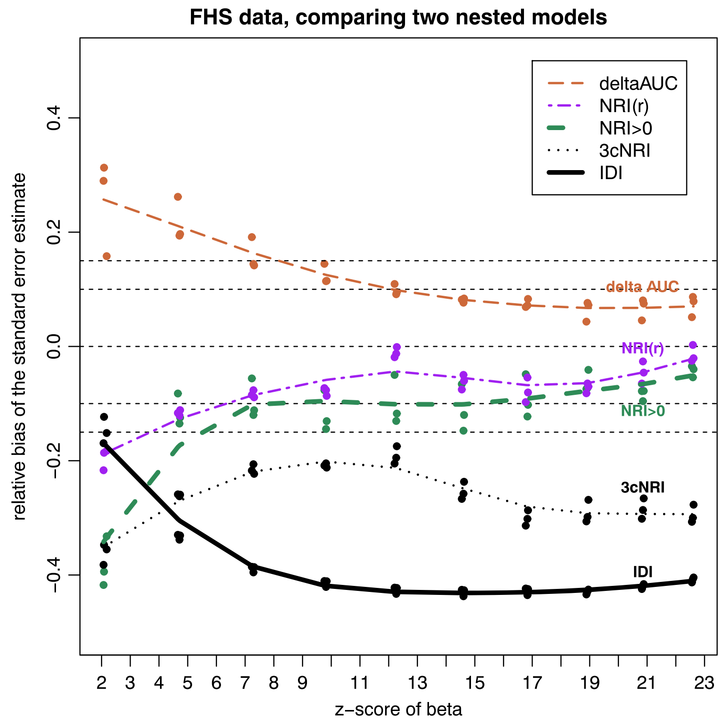

Figure 6. Relative bias of standard error estimate as a function of strength of the added predictor variable (z-score of βDBP) using Framingham Heart Study data. Reduced model included predictor variables: age, HDL and total cholesterol, systolic blood pressure, smoking, and diabetes status. Full model = reduced model + diastolic blood pressure (DBP). We artificially varied the strength of added DBP variable and calculated relative bias of variance estimate using theoretical formula relative to its bootstrapped value. z-score(β_{DBP}) = β_{DBP}/se(β_{DBP}). rel.bias = ($se_{formula.based}$-$se_{bootstrap}$)/$se_{bootstrap}$. [Colour figure can be viewed at wileyonlinelibrary.com]

parameters. As illustrated in Figure 6, their bias stays strong. For these reasons, resampling methods should be preferred in all situations to estimate standard errors of 3cNRI and IDI. We recommend similar strategies in estimating confidence intervals for ΔAUC, $NRI_{>0}$, NRI(r), 3cNRI, and IDI.

6. Discussion

This validation study shows that the behavior of ΔAUC, $NRI_{>0}$, NRI(r), 3cNRI, and IDI is affected by the interplay of several factors including the shift to degeneracy (non-normality) when comparing two nested models under the null and the lack of adjustment for estimated parameters for 3cNRI and IDI.

Our results explicitly specify conditions under which normal distribution theory can and cannot be applied to ΔAUC, $NRI_{>0}$, IDI, and categorical versions of the NRI when comparing two nested models. A few tests of these statistics have been proposed, and all with the exception of [20] rely on asymptotic normality. Our results imply that tests that rely on asymptotic normality are invalid for nested models and should not be used. Fortunately, testing is unnecessary: Pepe *et al.* [26] proved that testing of several measures of model performance is redundant because improvement in most of these statistics is equivalent to the significance of the new predictor variable. Therefore, the recommended strategy is to establish the significance of the regression coefficient first and then evaluate improvement in model performance by producing confidence intervals for measures of performance such as ΔAUC, NRIs, and IDI.

Using U-statistics theory, we proved that when the added predictor variable is significant, the distribution of ΔAUC, $NRI_{>0}$, NRI(r), 3cNRI, and IDI is normal; therefore, asymptotic confidence intervals can rely on the normal distribution. We considered their variance estimators and showed in Statement 2 that theoretical standard error estimates of ΔAUC, $NRI_{>0}$, and NRI(r) are valid when predictor variables are normally distributed. Our practical example using FHS data demonstrated that when the added predictor is significant but the *p*-value is not particularly low, the variance of $NRI_{>0}$

and NRI(r) is still underestimated by the formula and the variance of ΔAUC is overestimated. Our simulations demonstrated that a stronger added predictor variable is required to reach non-degeneracy, a necessary condition for validity of the formulas. We offer an example in which the *p*-value of added predictor variable $<10^{-5}$ is needed for ΔAUC, $NRI_{>0}$, and NRI(r) to fully transition to non-degeneracy and for the relative bias of the standard error of $NRI_{>0}$ and NRI(r) to drop to below 15% (Figure 6). Such high effect sizes and significance levels might be common in Big Data studies.

While formula-based standard errors of ΔAUC, $NRI_{>0}$, and NRI(r) are valid in the situations described previously, formula-based standard error estimators of 3cNRI and IDI are not. Unless they are adjusted for estimated parameters, they underestimate actual variance. Therefore, existing standard errors formulas for 3cNRI and IDI should not be used, and bootstrap or other resampling technique should be employed instead.

Additionally, using U-statistics theory, we showed that the standard error estimator of ΔAUC can be calculated as the variance of the change in ranks of predicted probabilities (Table III). In our numerical simulations, the new variance estimator was identical to the one produced by DeLong *et al*. [11] and the two are likely algebraically equivalent when there are no ties in predicted probabilities. However, rigorous proof of this result is beyond the scope of this paper.

In summary, when comparing two nested models after establishing the significance of the regression coefficient of an added predictor variable, we recommend estimating formula-based standard errors and confidence intervals of ΔAUC, $NRI_{>0}$, and NRI(r) when the significance of predictor variables is strong enough (*p*-value $<10^{-5}$, z-score >4.0 in our FHS data example). In other situations, the CIs of ΔAUC are too conservative; while CIs for all other statistics are too narrow, therefore, resampling techniques (such as bootstrap) should be used to estimate these. Standard errors of IDI and 3-calegory NRI should always be estimated by the bootstrap or other resampling technique.

Appendix A

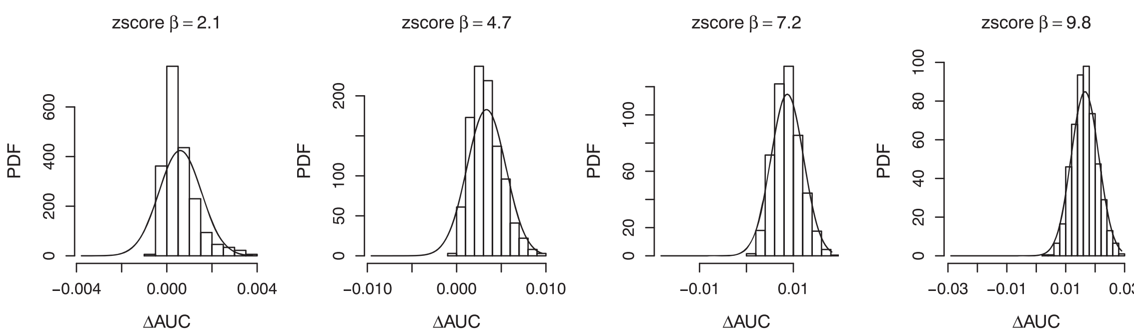

Figure A1. Transition of ΔAUC from degeneracy to non-degeneracy. Framingham Heart Study data. Reduced model included predictor variables: age, high-density lipoprotein and total cholesterol, systolic blood pressure, smoking, and diabetes status. Full model = reduced model + diastolic blood pressure (DPF). We artificially varied the strength of added DPF variable. Strength is measured by z-score(β_{dpf}) = β_{dpf}/se(β_{dpf}).

A.1. Brief review of U-statistics theory definitions and results

U-statistics theory can be viewed as an extension of Central Limit Theorem for sums of correlated variables [1]. Generalized U-statistics can be defined for a sample generated from *k* different distributions with no restriction on *k*, but for our purposes, we focus on the two-sample generalized U-statistics.

Definition 1

Generalized two-sample U-statistics [1]. Define $x_1, \ldots, x_{n_0}$ and $y_1, \ldots, y_{n_1}$ as samples of size n_0 and n_1, respectively, from two different (possibly multivariate) distribution functions $\mathbf{x} \sim \mathbf{F}(.)$ and $\mathbf{y} \sim \mathbf{G}(.)$.

Assume that $x_1, \ldots, x_{n_0}$ and $y_1, \ldots, y_{n_1}$ are independent within their samples and also across the two samples. Then *generalized U-statistic* is defined as follows:

$$U_{n_0,n_1} = \binom{n_0}{c}^{-1}\binom{n_1}{d}^{-1}\sum_{(n_0,c)}\sum_{(n_1,d)}\psi\left(x_{i_1},\ldots,x_{i_c};y_{j_1},\ldots,y_{j_d}\right), \tag{a1}$$

where *ψ(.;.) is a real-valued kernel* symmetric in each set of arguments and c and d are dimension constants satisfying $1 \le c \le n_0, 1 \le d \le n_1$. Note that $c = d = 1$ for ΔAUC, NRI > 0, and IDI (will be demonstrated in the following). Therefore, it is enough to consider $c = d = 1$. In this case, the symmetry condition is automatically satisfied and formula (a1) simplifies to

$$U_{n_0,n_1} = \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\psi\left(\boldsymbol{x_i};\boldsymbol{y_j}\right) \tag{a2}$$

U-statistics can be viewed as a sum of non-i.i.d. (where i.i.d. is independent and identically distributed) random variables for which Central Limit Theorem can be extended

Theorem [1]

$\sqrt{N}\left(U_{n_0,n_1} - \Theta\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0,\delta^2\right), N = n_0 + n_1 \to \infty$, unless U_{n_0,n_1} is a degenerate U-statistic.

See Lee [1] for a definition of δ^2.

Therefore, this Theorem fully defines asymptotic distribution of non-degenerate U-statistics. If we show that ΔAUC, NRI > 0, and IDI are non-degenerate U-statistics, we can use this Theorem in order to create their tests and confidence intervals. Degeneracy condition is defined in the following.

A.2. Brief review of degeneracy definitions for U-statistics

To utilize the Theorem, we need first to check degeneracy condition and then calculate variance; both of them are defined in terms of H-decomposition of U-statistics. Therefore, we first introduce H-decomposition and formulate degeneracy condition for any U-statistics and then calculate H-decomposition and define degeneracy condition of ΔAUC, $\text{NRI}_{>0}$, and IDI in the following.

Definition 2

First recall that $x_1, \ldots, x_{n_0}$ and $y_1, \ldots, y_{n_1}$ are samples of size n_0 and n_1, respectively, from two different (possibly multivariate) distribution functions $\mathbf{x} \sim \mathbf{F}(\cdot)$ and $\mathbf{y} \sim \mathbf{G}(\cdot)$.

H-decomposition of a generalized two-sample U-statistics [1–2]. Any generalized two-sample U-statistic (defined in (a2)) where $\mathbf{x} \sim F(.)$ and $\mathbf{y} \sim G(.)$ can be decomposed into a summation of several sums of i.i.d. random variables as follows:

$$\begin{aligned} U_{n_0,n_1} &\overset{def}{=} \frac{1}{n_0}\frac{1}{n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\psi\left(x_i;y_j\right) \\ &= h^{(0,\ 0)} + \frac{1}{n_1}\sum_{i=1}^{n_0}h^{(1,\ 0)}(x_i) + \frac{1}{n_1}\sum_{j=1}^{n_1}h^{(0,\ 1)}\left(y_j\right) + \frac{1}{n_0}\frac{1}{n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}h^{(1,\ 1)}\left(x_i;y_j\right), \end{aligned} \tag{a3}$$

where $h^{(.,.)}(.)$ are real-valued kernels. Formulas to calculate the kernels $h^{(.,.)}(.)$ are given by Lee [1]:

$$\begin{aligned} &h^{(0,0)} = \iint\psi(x;y)dF(x)dG(y) \\ &h^{(1,0)}(x_i) = \int\psi(x_i;y)dG(y) - \iint\psi(x;y)dF(x)dG(y) \\ &h^{(0,1)}\left(y_j\right) = \int\psi\left(x;y_j\right)dF(x) - \iint\psi(x;y)dF(x)dG(y) \\ &h^{(1,1)}\left(x_i;y_j\right) = \psi\left(x_i;y_j\right) - \int\psi(x_i;y)dG(y) - \int\psi\left(x;y_j\right)dF(x) + \iint\psi(x;y)dF(x)dG(y) \end{aligned} \tag{a4}$$

where F(.) and G(.) are cumulative distribution functions of random vectors x and y, $\mathbf{x} \sim F(.)$ and $\mathbf{y} \sim G(.)$.

Definition 3

Degeneracy of generalized U-statistics. Using the previous notation, if $\text{Var}(h^{(0,1)}(\mathbf{y})) = 0$ and $\text{Var}(h^{(1,0)}(\mathbf{x})) = 0$ for some family of distribution functions $F(.) \in \Phi(.)$ and $G(.) \in \Gamma(.)$, then generalized U-statistic H-decomposed as in Definition 2 is called degenerate.

It follows that in order to check degeneracy, it is enough to calculate the first two terms $h^{(1,0)}(x_i)$, $h^{(0,1)}(y_j)$ of H-decomposition. Let us calculate $h^{(1,0)}(x_i)$, $h^{(0,1)}(y_j)$ for ΔAUC, NRI > 0, IDI, and 3cNRI and check their degeneracy conditions.

For two nested models, ΔAUC, $\mathrm{NRI}_{>0}$, 3cNRI, and IDI

1. are generalized U-statistics.
2. are degenerate generalized U-statistics if models are under the null. If they are degenerate, they do not follow normal distribution.
3. are non-degenerate generalized U-statistics under the alternative and therefore are asymptotically normally distributed.
4. If we do not adjust for estimated parameters, then the variance formulas derived from U-statistics theory are same as the published formulas [3–4].

For two non-nested models, ΔAUC, NRI > 0, 3cNRI, and IDI

1. are generalized U-statistics.
2. are non-degenerate generalized U-statistics.
3. are asymptotically normally distributed with the variance formulas provided in the following.

Variance formulas:

$$\hat{\sigma}^2_{\Delta AUC} = \frac{Var\left(rank^*_e(\boldsymbol{a}^{*T}\boldsymbol{x_i}) - rank_e(\boldsymbol{a}^T\boldsymbol{x_i})\right)}{n_0} + \frac{Var\left(rank^*_{ne}\left(\boldsymbol{a}^{*T}\boldsymbol{y_j}\right) - rank_{ne}\left(\boldsymbol{a}^T\boldsymbol{y_j}\right)\right)}{n_1},$$

$$\hat{\sigma}^2_{NRI} = \frac{\hat{p}^{up}_{ne}\left(1-\hat{p}^{up}_{ne}\right)}{n_0} + \frac{\hat{p}^{up}_{e}\left(1-\hat{p}^{up}_{e}\right)}{n_1},$$

$$\hat{\sigma}^2_{3cNRI} = \frac{4\left(\hat{p}^{2up}_{ne}+\hat{p}^{2down}_{ne}\right)+\hat{p}^{1up}_{ne}+\hat{p}^{1down}_{ne}-\left(2\left(\hat{p}^{2up}_{ne}-\hat{p}^{2down}_{ne}\right)+\hat{p}^{1up}_{ne}-\hat{p}^{1down}_{ne}\right)^2}{n_0}$$
$$+\frac{4\left(\hat{p}^{2up}_{ev}+\hat{p}^{2down}_{ev}\right)+\hat{p}^{1up}_{ev}+\hat{p}^{1down}_{ev}-\left(2\left(\hat{p}^{2up}_{ev}-\hat{p}^{2down}_{ev}\right)+\hat{p}^{1up}_{ev}-\hat{p}^{1down}_{ev}\right)^2}{n_1},$$

$$\hat{\sigma}^2_{IDI} = \frac{Var(\Delta predp(\boldsymbol{x_i}))}{n_0} + \frac{Var\left(\Delta predp\left(\boldsymbol{y_j}\right)\right)}{n_1},$$

where

$\mathrm{rank}_e(\boldsymbol{a}^T\boldsymbol{x_i})$ is the rank of $\boldsymbol{a}^T\boldsymbol{x_i}$ among $\boldsymbol{a}^T\boldsymbol{y}$ (risk scores of events),
$\mathrm{rank}_{ne}(\boldsymbol{a}^T\boldsymbol{y_j})$ is the rank of $\boldsymbol{a}^T\boldsymbol{y_j}$ among $\boldsymbol{a}^T\boldsymbol{x_i}$(risk scores of events) and as before * means that similar functions and coefficients were calculated for the old model.

$$\hat{p}^{up}_{ne} = \frac{\#non-events\ moving\ up}{n_0},$$
$$\hat{p}^{up}_{e} = \frac{\#events\ moving\ up}{n_1},$$

$\Delta predp(\boldsymbol{x_i})$ is the difference in predicted probabilities from two models for non-event x_i
$\Delta predp(\boldsymbol{y_j})$ is the difference in predicted probabilities from two models for event $\boldsymbol{y_j}$.

Appendix B: Proof of the main result

B.1. Main result

ΔAUC, $\mathrm{NRI}_{>0}$, *NRI(r), 3cNRI and IDI*

STATEMENT 1 *are **generalized U-statistics with estimated parameters**.*
STATEMENT 2 *belong to **non-degenerate subclass** if they compare any non-nested models or nested models under the alternative. As non-degenerate U-statistics, they*

a *follow normal distribution asymptotically.*

b *Available variance formulas are algebraically equal to the variance estimates provided by U-statistics theory if we ignore adjustment for estimated parameters.*
c *Variances of ΔAUC, $NRI_{>0}$, and NRI(r) do not need to be adjusted for estimated parameters if predictor variables are normally distributed.*
d *Variance of IDI and 3-category NRI always should be adjusted for estimated parameters.*

STATEMENT 3 ΔAUC, $NRI_{>0}$, *NRI(r), 3cNRI, and IDI belong to* ***degenerate subclass*** *if they compare nested models under the null. As degenerate U-statistics, they do not follow normal distribution and available variance estimates do not apply for them.*

B.1.1. STATEMENT 1. ΔAUC, NRI > 0, and IDI are generalized two-sample U-statistics.

Lemma 1
ΔAUC, NRI $>$ 0, IDI, and categorical NRIs are generalized two-sample U-statistics.

Proof
We need to show that each of the three statistics can be written in the form (a2).

B.1.1.1. ΔAUC is a generalized two-sample U-statistics. We assume that AUC is estimated by the Mann–Whitney statistic [5] – a non-parametric unbiased estimator, often referred to as the *c*-statistic [6–7]. Mann–Whitney statistic for the full model is as follows:

$$AUC = \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} I\left[\boldsymbol{a}'\boldsymbol{x_i} < \boldsymbol{a}'\boldsymbol{y_j}\right],$$

where I[.] is the indicator function.

AUC for the reduced model is as follows:

$$AUC^* = \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} I\left[\boldsymbol{a}^{*'}\boldsymbol{x_i} < \boldsymbol{a}^{*'}\boldsymbol{y_j}\right]$$

If we write ΔAUC as (a2), then we will show that ΔAUC is a U-statistic.

$$\begin{aligned}\Delta AUC &= \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} I\left[\boldsymbol{a}'\boldsymbol{x_i} < \boldsymbol{a}'\boldsymbol{y_j}\right] - \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} I\left[\boldsymbol{a}^{*'}\boldsymbol{x_i} < \boldsymbol{a}^{*'}\boldsymbol{y_j}\right] \\ &= \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\left(I\left[\boldsymbol{a}'\boldsymbol{x_i} < \boldsymbol{a}'\boldsymbol{y_j}\right] - I\left[\boldsymbol{a}^{*'}\boldsymbol{x_i} < \boldsymbol{a}^{*'}\boldsymbol{y_j}\right]\right)\end{aligned} \tag{a5}$$

Denoting $\psi_{\Delta AUC}(x,y) = I\left[\boldsymbol{a}'\boldsymbol{x_i} < \boldsymbol{a}'\boldsymbol{y_j}\right] - I\left[\boldsymbol{a}^{*'}\boldsymbol{x_i} < \boldsymbol{a}^{*'}\boldsymbol{y_j}\right]$ as the kernel, we see that ΔAUC satisfies the properties required in Definition 1 for generalized two-sample U-statistics.

B.1.1.2. $NRI_{>0}$ is a generalized two-sample U-statistics. $\mathrm{NRI}_{>0}$ is defined as follows:

$$\begin{aligned}NRI_{>0} &= \frac{\#up, events}{n_1} - \frac{\#up, nonevents}{n_0} = \frac{1}{n_1 n_0}\left(n_0\cdot \#up, events - n_1\cdot \#up, nonevents\right) \\ &= \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\left(I\left[\left(\boldsymbol{a}' - \boldsymbol{a}^{*'}\right)\boldsymbol{y_j} > 0\right] - I\left[\left(\boldsymbol{a}' - \boldsymbol{a}^{*'}\right)\boldsymbol{x_i} > 0\right]\right)\end{aligned} \tag{a6}$$

Denoting kernel as $\psi_{\mathrm{NRI}_{>0}}(x,y) = I\left[\left(\boldsymbol{a}' - \boldsymbol{a}^{*'}\right)\boldsymbol{y_j} > 0\right] - I\left[\left(\boldsymbol{a}' - \boldsymbol{a}^{*'}\right)\boldsymbol{x_i} > 0\right]$, we see that continuous NRI (NRI $>$ 0) satisfies the properties required in Definition 1 for generalized two-sample U-statistics.

B.1.1.3. Three-category NRI is a generalized two-sample U-statistics. We can define categories as low (L), intermediate (I), and high (H), defined by cutoffs c1 and c2. Then jump size adjusted NRI is defined as follows:

$$3cNRI = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \left(I_{3c}\left[\boldsymbol{y_j}\right] - I_{3c}[\boldsymbol{x_i}] \right), \quad \text{(a9)}$$

where

$$I_{3c}[\boldsymbol{x_i}] = \begin{cases} 2, & \textit{if predicted prob of non} - \textit{event i moved two categoriesup} \\ 1, & \textit{if predicted prob of non} - \textit{event i moved one category up} \\ 0, & \textit{if predicted prob of non} - \textit{event i did not change} \\ -1, & \textit{if predicted prob of non} - \textit{event i moved one category down} \\ -2, & \textit{if predicted prob of non} - \textit{event i moved two categories down} \end{cases}$$

Similar definition for $I_{js}[\boldsymbol{y_j}]$.

Denoting kernel as $\psi_{3\mathrm{cNRI}}(x, y) = I_{jump}[\boldsymbol{y_j}] - I_{jump}[\boldsymbol{x_i}]$, we see that (a9) satisfies the properties required in Definition 1 for generalized two-sample U-statistics.

B.1.1.4. IDI is a generalized two-sample U-statistics. Denoting $f(\cdot) = inv.logit(\cdot)$, IDI is defined as follows:

$$IDI = \frac{\sum_1^{n_1} p_j - p_j^*}{n_1} - \frac{\sum_1^{n_0} p_i - p_i^*}{n_0} = \frac{\sum_1^{n_1} f\left(\boldsymbol{a}'\boldsymbol{y_j}\right) - f\left(\boldsymbol{a}^{*\prime}\boldsymbol{y_j}\right)}{n_1} - \frac{\sum_1^{n_0} f\left(\boldsymbol{a}'\boldsymbol{x_i}\right) - f\left(\boldsymbol{a}^{*\prime}\boldsymbol{x_i}\right)}{n_0}$$
$$= \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \left(f\left(\boldsymbol{a}'\boldsymbol{y_j}\right) - f\left(\boldsymbol{a}^{*\prime}\boldsymbol{y_j}\right) - \left(f\left(\boldsymbol{a}'\boldsymbol{x_i}\right) - f\left(\boldsymbol{a}^{*\prime}\boldsymbol{x_i}\right) \right) \right) \quad \text{(a7)}$$

Denoting kernel as $\psi_{\mathrm{IDI}}(x, y) = f\left(\boldsymbol{a}'\boldsymbol{y_j}\right) - f\left(\boldsymbol{a}^{*\prime}\boldsymbol{y_j}\right) - \left(f\left(\boldsymbol{a}'\boldsymbol{x_i}\right) - f\left(\boldsymbol{a}^{*\prime}\boldsymbol{x_i}\right)\right)$, we see that IDI satisfies the properties required in Definition 1 for generalized two-sample U-statistics.

Q.E.D.

B.2. Degeneracy condition of ΔAUC, NRIs, and IDI

B.2.1. STATEMENT 2. ΔAUC, NRIs, and IDI belong to non-degenerate subclass if they compare any non-nested models or nested models under the alternative.

Proof

First, let us write H-decomposition of ΔAUC, NRI > 0, IDI, and categorical NRIs.

B.2.1. H-decomposition and degeneracy condition of ΔAUC.

$$\begin{aligned} \mathrm{h}^{(1,0)}(x_i) &= \int \psi(x_i; y)\mathrm{dG}\,(y) - \iint \psi(x; y)\mathrm{dF}\,(x)\mathrm{dG}\,(y) \\ \mathrm{h}^{(0,1)}\left(y_j\right) &= \int \psi\left(x; y_j\right)\mathrm{dF}\,(x) - \iint \psi(x; y)\mathrm{dF}\,(x)\mathrm{dG}\,(y) \end{aligned} \quad \text{(a9)}$$

Recall that kernel of ΔAUC is defined as follows:

$$\psi_{\Delta AUC}(x,\ y) = I\left[\boldsymbol{a}'\boldsymbol{x_i} < \boldsymbol{a}'\boldsymbol{y_j}\right] - I\left[\boldsymbol{a}^{*\prime}\boldsymbol{x_i} < \boldsymbol{a}^{*\prime}\boldsymbol{y_j}\right]$$

Plugging it into (a5), we obtain the following:

$$\begin{aligned} h^{(1,\ 0)}(x_i) &= \int I\left[\boldsymbol{a}'\boldsymbol{x_i} < \boldsymbol{a}'\boldsymbol{y}\right] - I\left[\boldsymbol{a}^{*\prime}\boldsymbol{x_i} < \boldsymbol{a}^{*\prime}\boldsymbol{y}\right] d\boldsymbol{G}(\boldsymbol{y}) - h^{(0,\ 0)} \\ &= \Pr\left(\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{x_i} < \boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{y}\right) - \Pr\left(\boldsymbol{a}^*\boldsymbol{T}\boldsymbol{x_i} < \boldsymbol{a}^*\boldsymbol{T}\boldsymbol{y}\right) - h^{(0,\ 0)} \\ &= G^*\left(\boldsymbol{a}^*\boldsymbol{T}\boldsymbol{x_i}\right) - G\left(\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{x_i}\right) - h^{(0,\ 0)} \end{aligned} \quad \text{(a10)}$$

$$\begin{aligned} h^{(0,\ 1)}(y_i) &= \int I\left[\boldsymbol{a}'\boldsymbol{x} < \boldsymbol{a}'\boldsymbol{y_j}\right] - I\left[\boldsymbol{a}^{*'}\boldsymbol{x} < \boldsymbol{a}^{*'}\boldsymbol{y_j}\right] d\boldsymbol{F}(\boldsymbol{x}) - h^{(0,\ 0)} \\ &= \Pr\left(\boldsymbol{a^T x} < \boldsymbol{a^T y_j}\right) - \Pr\left(\boldsymbol{a}^*\boldsymbol{Tx} < \boldsymbol{a}^*\boldsymbol{Ty_j}\right) - h^{(0,\ 0)} \\ &= F\left(\boldsymbol{a^T y_j}\right) - F^*\left(\boldsymbol{a}^*\boldsymbol{Ty_j}\right) - h^{(0,\ 0)} \end{aligned} \tag{a11}$$

where $h^{(0,0)} = \iint \psi_{\Delta AUC}(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{F}(\boldsymbol{x})d\boldsymbol{G}(\boldsymbol{y})$ in (a5), and it is a constant; $\boldsymbol{a^T x_i} \sim F(\cdot)$, $\boldsymbol{a^{*T} x_i} \sim F^*(\cdot)$, $\boldsymbol{a^T y_i} \sim G(\cdot)$, $\boldsymbol{a^{*T} y_i} \sim G^*(\cdot)$.

B.2.1.1.1. Degeneracy condition of ΔAUC. ΔAUC is a degenerate U-statistics if $Var(h^{(1,0)}(x_i)) = 0$ and $Var(h^{(0,1)}(y_j)) = 0$. From (a6), it means that $G^*(\boldsymbol{a^{*T} x_i}) - G(\boldsymbol{a^T x_i}) \equiv constant\,\forall \boldsymbol{x_i}$ and $F^*(\boldsymbol{a^{*T} y_i}) - F(\boldsymbol{a^T y_i}) \equiv constant\,\forall \boldsymbol{y_j}$, which can happen only if $\boldsymbol{a}^* = \boldsymbol{a}$, which is true for nested model under the null.

B.2.1.2. H-decomposition and degeneracy condition of $NRI_{>0}$. Recall that kernel of $\mathrm{NRI}_{>0}$ is as follows:

$$\psi_{NRI_{>0}}(x,y) = I\left[\left(\boldsymbol{a}' - \boldsymbol{a}^{*'}\right)\boldsymbol{y_j} > 0\right] - I\left[\left(\boldsymbol{a}' - \boldsymbol{a}^{*'}\right)\boldsymbol{x_i} > 0\right]$$

Plugging it into (a10) we obtain the following:

$$\begin{aligned} h^{(1,\ 0)}(x_i) &= \int I\left[\left(\boldsymbol{a}' - \boldsymbol{a}^{*'}\right)\boldsymbol{y_j} > 0\right] - I\left[\left(\boldsymbol{a}' - \boldsymbol{a}^{*'}\right)\boldsymbol{x_i} > 0\right] d\boldsymbol{G}(\boldsymbol{y}) - h^{(0,\ 0)} \\ &= \Pr((\boldsymbol{a^T} - \boldsymbol{a}^*\boldsymbol{T})\boldsymbol{y} > 0) - \mathrm{I}[(\boldsymbol{a^T} - \boldsymbol{a}^*\boldsymbol{T})\boldsymbol{x_i} > 0] - h^{(0,\ 0)} \\ h^{(0,\ 1)}(y_i) &= \int I\left[\left(\boldsymbol{a}' - \boldsymbol{a}^{*'}\right)\boldsymbol{y_j} > 0\right] - I\left[\left(\boldsymbol{a}' - \boldsymbol{a}^{*'}\right)\boldsymbol{x_i} > 0\right] d\boldsymbol{F}(\boldsymbol{x}) - h^{(0,\ 0)} \\ &= \mathrm{I}\left[(\boldsymbol{a^T} - \boldsymbol{a}^*\boldsymbol{T})\boldsymbol{y_j} > 0\right] - \Pr((\boldsymbol{a^T} - \boldsymbol{a}^*\boldsymbol{T})\boldsymbol{x} > 0) - h^{(0,\ 0)} \end{aligned} \tag{a12}$$

Where $h^{(0,0)} = \iint \psi_{\mathrm{NRI}}(\boldsymbol{x},\boldsymbol{y})d\boldsymbol{F}(\boldsymbol{x})d\boldsymbol{G}(\boldsymbol{y})$ in (a5), and it is a constant.

B.2.1.2.1. Degeneracy condition of $NRI_{>0}$. $NRI > 0$ is a degenerate U-statistics if $Var(h^{(1,0)}(x_i)) = s0$ and $Var(h^{(0,1)}(y_j)) = 0$. From (a7), it is equivalent to $I[(\boldsymbol{a^T} - \boldsymbol{a^{*T}})\boldsymbol{x_i} > 0] \equiv constant\,\forall \boldsymbol{x_i}$ and $I[(\boldsymbol{a^T} - \boldsymbol{a^{*T}})\boldsymbol{y_j} > 0] \equiv constant\,\forall \boldsymbol{y_j}$, which can happen only if $\boldsymbol{a}^* = \boldsymbol{a}$, which is true for nested model under the null.

B.2.1.3. H-decomposition and degeneracy condition of 3-category NRI. From (a9),

$$\begin{aligned} \psi_{3\mathrm{cNRI}}(x,y) &= I_{3c}\left[\boldsymbol{y_j}\right] - I_{3c}[\boldsymbol{x_i}], \ , \\ h^{(1,\ 0)}(x_i) &= -I_{3c}[\boldsymbol{x_i}] + const \\ h^{(0,\ 1)}\left(y_j\right) &= I_{3c}\left[\boldsymbol{y_j}\right] + const \end{aligned} \tag{a14}$$

B.2.1.3.1. Degeneracy condition of three-category NRI. Three-category NRI is a degenerate U-statistics if $Var(h^{(1,0)}(x_i)) = 0$ and $Var(h^{(0,1)}(y_j)) = 0$. From (a9), it is equivalent to $I_{3c}[\boldsymbol{y_j}] \equiv constant\,\forall \boldsymbol{y_j}$ and $I_{3c}[\boldsymbol{x_i}] \equiv constant\,\forall \boldsymbol{x_i}$, which can happen only if $\boldsymbol{a}^* = \boldsymbol{a}$, which is true for any nested models under the null.

B.2.1.4. H-decomposition and degeneracy condition of IDI. Recall that kernel of IDI is as follows:

$\psi_{\mathrm{IDI}}(x,y) = f\left(\boldsymbol{a}'\boldsymbol{y_j}\right) - f\left(\boldsymbol{a}^{*'}\boldsymbol{y_j}\right) - \left(f\left(\boldsymbol{a}'\boldsymbol{x_i}\right) - f\left(\boldsymbol{a}^{*\prime}\boldsymbol{x_i}\right)\right)$, where *f(.)* is the predicted probability.

Plugging it into (a10), we obtain

$$\begin{aligned} h^{(1,\ 0)}(x_i) &= \int f\left(\boldsymbol{a}'\boldsymbol{y}\right) - f\left(\boldsymbol{a}^{*'}\boldsymbol{y}\right) - \left(f\left(\boldsymbol{a}'\boldsymbol{x_i}\right) - f(\boldsymbol{a}^{*}\boldsymbol{'x_i})\right) d\boldsymbol{G}(\boldsymbol{y}) - h^{(0,\ 0)} \\ &= f\left(\boldsymbol{a}'\boldsymbol{x_i}\right) - f\left(\boldsymbol{a}^{*'}\boldsymbol{x_i}\right) + const \\ h^{(1,\ 0)}(x_i) &= \int f\left(\boldsymbol{a}'\boldsymbol{y_j}\right) - f\left(\boldsymbol{a}^{*'}\boldsymbol{y_j}\right) - \left(f\left(\boldsymbol{a}'\boldsymbol{x}\right) - f(\boldsymbol{a}^{*}\boldsymbol{'x})\right) d\boldsymbol{F}(\boldsymbol{x}) - h^{(0,\ 0)} \\ &= f\left(\boldsymbol{a}'\boldsymbol{y_j}\right) - f\left(\boldsymbol{a}^{*'}\boldsymbol{y_j}\right) + const \end{aligned} \tag{a13}$$

B.2.1.4.1. Degeneracy condition of IDI. *IDI* is a degenerate U-statistics if $Var(h^{(1,0)}(x_i))=0$ and $Var(h^{(0,1)}(y_j))=0$. From (a8), it is equivalent to $I[(\boldsymbol{a^T}-\boldsymbol{a^{*T}})\boldsymbol{x_i}>0]\equiv constant\,\forall\boldsymbol{x_i}$ and $I[(\boldsymbol{a^T}-\boldsymbol{a^{*T}})\boldsymbol{y_j}>0]\equiv constant\,\forall\boldsymbol{y_j}$, which can happen only if $\boldsymbol{a}^*=\boldsymbol{a}$, which is true for any nested models under the null.

Q.E.D.

So all statistics considered in this paper are degenerate if and only if we compare nested models under the null. When the three statistics are non-degenerate, we can apply the Theorem and calculate their variance.

B.2.2. STATEMENT 2a. When ΔAUC, NRIs, and IDI belong to non-degenerate subclass, they follow normal distribution asymptotically..

$$\sqrt{\mathbf{N}}(\Delta\mathbf{AUC} - \mathbf{E}[\Delta\mathbf{AUC}]) \xrightarrow{D} N\left(0, \mathbf{N}\boldsymbol{\sigma}^2_{\Delta\mathbf{AUC}}\right),$$

$$\sqrt{\mathbf{N}}(\mathbf{NRI}_{>0} - \mathbf{E}[\mathbf{NNRI}_{>0}]) \xrightarrow{D} N\left(0, \mathbf{N}\boldsymbol{\sigma}^2_{\mathbf{NRI}_{>0}}\right),$$

$$\sqrt{\mathbf{N}}(\mathbf{IDI} - \mathbf{E}[\mathbf{IDI}]) \xrightarrow{D} N\left(0, \mathbf{N}\boldsymbol{\sigma}^2_{\mathbf{IDI}}\right)$$

$$\sqrt{\mathbf{N}}(3\mathbf{cNRI} - \mathbf{E}[3\mathbf{cNRI}]) \xrightarrow{D} N\left(0, \mathbf{N}\boldsymbol{\sigma}^2_{3\mathbf{cNRI}}\right),$$

$$\sqrt{\mathbf{N}}(\mathbf{NRI(p)} - \mathbf{E}[\mathbf{NRI(p)}]) \xrightarrow{D} N\left(0, \mathbf{N}\boldsymbol{\sigma}^2_{\mathbf{NRI(p)}}\right)$$

To derive $\sigma^2_{\Delta AUC}, \sigma^2_{\mathrm{NRI}_{>0}}, \sigma^2_{IDI}, \sigma^2_{3cNRI}$ and $\sigma^2_{NRI(p)}$, we will rewrite the Theorem in a more detailed way:

Theorem [Lee [1] pg 140].
$\sqrt{N}\left(U_{n_0,n_1} - \Theta\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \delta^2\right)$ *as* $N = n_0 + n_1 \to \infty$, unless U_{n_0,n_1} is a degenerate U-statistic. Above $\delta^2 = \frac{1}{1-p}\delta^2_{1,0} + \frac{1}{p}\delta^2_{0,1}, p = \lim\limits_{N\to\infty}\frac{n_1}{N}$

and

$$\delta^2_{1,0} = Var\left(h^{(1,\ 0)}(x_i)\right),$$

$$\delta^2_{0,1} = Var\left(h^{(0,\ 1)}\left(y_j\right)\right).$$

Note that $\delta^2 = \frac{1}{1-p}\delta^2_{1,0} + \frac{1}{p}\delta^2_{0,1} \approx \frac{N}{n_0}\delta^2_{1,0} + \frac{N}{n_1}\delta^2_{0,1}$

B.2.2.1. Derivation of $\sigma^2_{\Delta\mathrm{AUC}}$.

$$\hat{\sigma}^2_{\Delta AUC} = \frac{Var\left(rank^*_e(\boldsymbol{a}^*\boldsymbol{Tx_i}) - rank_e(\boldsymbol{a^T x_i})\right)}{n_0} + \frac{Var\left(rank^*_{ne}\left(\boldsymbol{a}^*\boldsymbol{Ty_j}\right) - rank_{ne}\left(\boldsymbol{a^T y_j}\right)\right)}{n_1}$$

$$\sigma^2_{\Delta AUC} = \frac{1}{N}\left(\frac{1}{1-p}\delta^2_{1,0} + \frac{1}{p}\delta^2_{0,1}\right) \approx \frac{1}{N}\left(\frac{N}{n_0}\delta^2_{1,0} + \frac{N}{n_1}\delta^2_{0,1}\right) = \frac{1}{n_0}\delta^2_{1,0} + \frac{1}{n_1}\delta^2_{0,1}$$

From (a12),

$$\delta^2_{1,0} = Var\left(h^{(1,\ 0)}(\boldsymbol{x_i})\right) = Var\left(G^*(\boldsymbol{a}^*\boldsymbol{T}\boldsymbol{x_i}) - G\left(\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{x_i}\right)\right)$$

$$\delta^2_{0,1} = Var\left(h^{(0,\ 0)}\left(\boldsymbol{y_j}\right)\right) = Var\left(G^*(\boldsymbol{a}^*\boldsymbol{T}\boldsymbol{x_i}) - G\left(\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{x_i}\right)\right)$$

$G(\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{x_i})$ can be estimated as the rank of $\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{x_i}$ among $\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{y}$ (risk scores of events).

$$G\left(\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{x_i}\right) \approx rank_e\left(\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{x_i}\right),\quad G^*(\boldsymbol{a}^*\boldsymbol{T}\boldsymbol{x_i}) \approx rank^*_e(\boldsymbol{a}^*\boldsymbol{T}\boldsymbol{x_i})$$
$$F\left(\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{y_j}\right) \approx rank_{ne}\left(\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{y_j}\right),\quad F^*\left(\boldsymbol{a}^*\boldsymbol{T}\boldsymbol{y_j}\right) \approx rank^*_{ne}\left(\boldsymbol{a}^*\boldsymbol{T}\boldsymbol{y_j}\right)$$

Therefore,

$$\delta^2_{1,0} = Var\left(h^{(1,\ 0)}(\boldsymbol{x_i})\right) \approx Var\left(rank^*_e(\boldsymbol{a}^*\boldsymbol{T}\boldsymbol{x_i}) - rank_e\left(\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{x_i}\right)\right)$$

$$\delta^2_{0,1} = Var\left(h^{(0,\ 1)}\left(\boldsymbol{y_j}\right)\right) \approx Var\left(rank^*_{ne}\left(\boldsymbol{a}^*\boldsymbol{T}\boldsymbol{y_j}\right) - rank_{ne}\left(\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{y_j}\right)\right)$$

$$\hat{\sigma}^2_{\Delta AUC} = \frac{Var\left(rank^*_e(\boldsymbol{a}^*\boldsymbol{T}\boldsymbol{x_i}) - rank_e(\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{x_i})\right)}{n_0} + \frac{Var\left(rank^*_{ne}\left(\boldsymbol{a}^*\boldsymbol{T}\boldsymbol{y_j}\right) - rank_{ne}\left(\boldsymbol{a}^{\boldsymbol{T}}\boldsymbol{y_j}\right)\right)}{n_1}$$

B.2.2.2. Derivation of $\boldsymbol{\sigma}^2_{\boldsymbol{NRI}_{>0}}$.

$$\hat{\sigma}^2_{NRI_{>0}} = \frac{\hat{p}^{up}_{ne}\left(1-\hat{p}^{up}_{ne}\right)}{n_0} + \frac{\hat{p}^{up}_{e}\left(1-\hat{p}^{up}_{e}\right),}{n_1}$$

$$\sigma^2_{NRI_{>0}} = \frac{1}{n_0}\delta^2_{1,0} + \frac{1}{n_1}\delta^2_{0,1}$$

From (a13),

$$\delta^2_{1,0} = Var\left(h^{(1,\ 0)}(\boldsymbol{x_i})\right) = Var\left(I\left[\left(\boldsymbol{a}^{\boldsymbol{T}} - \boldsymbol{a}^*\boldsymbol{T}\right)\boldsymbol{x_i} > 0\right]\right) \approx \hat{p}^{up}_{ne}\left(1-\hat{p}^{up}_{ne}\right)$$

$$\delta^2_{0,1} = Var\left(h^{(0,\ 1)}\left(\boldsymbol{y_j}\right)\right) = Var\left(I\left[\left(\boldsymbol{a}^{\boldsymbol{T}} - \boldsymbol{a}^*\boldsymbol{T}\right)\boldsymbol{y_j} > 0\right]\right) \approx \hat{p}^{up}_{e}\left(1-\hat{p}^{up}_{e}\right),$$

where $\hat{p}^{up}_{ne} = \dfrac{\#non-events\ moving\ up}{n_0}$ and $\hat{p}^{up}_{e} = \dfrac{\#events\ moving\ up}{n_1}$

$$\hat{\sigma}^2_{NRI_{>0}} = \frac{\hat{p}^{up}_{ne}\left(1-\hat{p}^{up}_{ne}\right)}{n_0} + \frac{\hat{p}^{up}_{e}\left(1-\hat{p}^{up}_{e}\right),}{n_1}$$

B.2.2.3. Derivation of $\sigma^2_{\mathrm{3cNRI}}$.

$$\hat{\sigma}^2_{3cNRI} = \frac{4\left(\hat{p}^{2up}_{ne} + \hat{p}^{2down}_{ne}\right) + \hat{p}^{1up}_{ne} + \hat{p}^{1down}_{ne} - \left(2\left(\hat{p}^{2up}_{ne} - \hat{p}^{2down}_{ne}\right) + \hat{p}^{1up}_{ne} - \hat{p}^{1down}_{ne}\right)^2}{n_0}$$
$$+ \frac{4\left(\hat{p}^{2up}_{ev} + \hat{p}^{2down}_{ev}\right) + \hat{p}^{1up}_{ev} + \hat{p}^{1down}_{ev} - \left(2\left(\hat{p}^{2up}_{ev} - \hat{p}^{2down}_{ev}\right) + \hat{p}^{1up}_{ev} - \hat{p}^{1down}_{ev}\right)^2}{n_1},$$

where $\hat{p}_{ev}^{k\ up}$ is the fraction of events that moved k categories up in the full model.

$$\sigma^2_{3cNRI} = \frac{1}{n_0}\delta^2_{1,0} + \frac{1}{n_1}\delta^2_{0,1}$$

From (a15),

$$\begin{aligned} h^{(1,\ 0)}(x_i) &= -I_{js}[\boldsymbol{x_i}] + const \\ h^{(0,\ 1)}\left(y_j\right) &= I_{js}\left[\boldsymbol{y_j}\right] + const \end{aligned} \tag{a14}$$

$\delta^2_{1,\ 0} = Var\left(h^{(1,\ 0)}(x_i)\right) = Var\left(I_{js}[x_i]\right)$ is the variance of a multinomial random variable that takes on values in (2, 1, 0, -1, -2).

$$\begin{aligned} \hat{\delta}^2_{1,0} &\approx 4\left(\hat{p}^{2up}_{ne} + \hat{p}^{2down}_{ne}\right) + \hat{p}^{1up}_{ne} + \hat{p}^{1down}_{ne} - \left(2\left(\hat{p}^{2up}_{ne} - \hat{p}^{2down}_{ne}\right) + \hat{p}^{1up}_{ne} - \hat{p}^{1down}_{ne}\right)^2 \\ \hat{\delta}^2_{0,1} &\approx 4\left(\hat{p}^{2up}_{ev} + \hat{p}^{2down}_{ev}\right) + \hat{p}^{1up}_{ev} + \hat{p}^{1down}_{ev} - \left(2\left(\hat{p}^{2up}_{ev} - \hat{p}^{2down}_{ev}\right) + \hat{p}^{1up}_{ev} - \hat{p}^{1down}_{ev}\right)^2 \end{aligned}$$

Then
$\hat{\sigma}^2_{3cNRI} = \frac{\hat{\delta}^2_{1,0}}{n_0} + \frac{\hat{\delta}^2_{0,1}}{n_1}$, where $\hat{\delta}^2_{1,0}$ and $\hat{\delta}^2_{0,1}$ are defined earlier.

$$\begin{aligned} \hat{\sigma}^2_{3cNRI} = {} & \frac{4\left(\hat{p}^{2up}_{ne} + \hat{p}^{2down}_{ne}\right) + \hat{p}^{1up}_{ne} + \hat{p}^{1down}_{ne} - \left(2\left(\hat{p}^{2up}_{ne} - \hat{p}^{2down}_{ne}\right) + \hat{p}^{1up}_{ne} - \hat{p}^{1down}_{ne}\right)^2}{n_0} \\ & + \frac{4\left(\hat{p}^{2up}_{ev} + \hat{p}^{2down}_{ev}\right) + \hat{p}^{1up}_{ev} + \hat{p}^{1down}_{ev} - \left(2\left(\hat{p}^{2up}_{ev} - \hat{p}^{2down}_{ev}\right) + \hat{p}^{1up}_{ev} - \hat{p}^{1down}_{ev}\right)^2}{n_1} \end{aligned}$$

B.2.2.4. Derivation of σ^2_{IDI}.

$$\begin{aligned} \hat{\sigma}^2_{IDI} &= \frac{Var(\Delta predp(\boldsymbol{x_i}))}{n_0} + \frac{Var\left(\Delta predp\left(\boldsymbol{y_j}\right)\right)}{n_1}, \\ \sigma^2_{IDI} &= \frac{1}{n_0}\delta^2_{1,0} + \frac{1}{n_1}\delta^2_{0,1} \end{aligned}$$

From (a14),

$$\begin{aligned} \delta^2_{1,0} &= Var\left(h^{(1,\ 0)}(\boldsymbol{x_i})\right) = Var\left(f\left(\boldsymbol{a}'\boldsymbol{x_i}\right) - f\left(\boldsymbol{a}^{*'}\boldsymbol{x_i}\right)\right) \approx Var(\Delta predp(\boldsymbol{x_i})) \\ \delta^2_{0,1} &= Var\left(h^{(0,\ 1)}\left(\boldsymbol{y_j}\right)\right) = Var\left(f\left(\boldsymbol{a}'\boldsymbol{y_j}\right) - f\left(\boldsymbol{a}^{*'}\boldsymbol{y_j}\right)\right) \approx Var\left(\Delta predp\left(\boldsymbol{y_j}\right)\right) \end{aligned}$$

where
$\Delta predp(\boldsymbol{x_i})$ is the difference in predicted probabilities from two models for non-event $x_i$and
$\Delta predp(\boldsymbol{y_j})$ is the difference in predicted probabilities from two models for event $\boldsymbol{y_j}$

$$\hat{\sigma}^2_{IDI} = \frac{Var(\Delta predp(\boldsymbol{x_i}))}{n_0} + \frac{Var\left(\Delta predp\left(\boldsymbol{y_j}\right)\right)}{n_1},$$

Variances derived here are summarized in Table III of the paper and are reproduced here (Table A1). q.e.d.

Table A1. Variance formulas in non-degenerate case, unadjusted for estimated parameters.

	$\widehat{\sigma^2}$, ignoring the adjustment for estimated parameters	Requires adjustment?
$\hat{\sigma}^2_{\Delta AUC}$ no tied ranks	$\frac{Var\left(rank_e^*(\boldsymbol{a}^*\boldsymbol{T}\boldsymbol{x}_i)-rank_e(\boldsymbol{a}^T\boldsymbol{x}_i)\right)}{n_0}+\frac{Var\left(rank_{ne}^*\left(\boldsymbol{a}^*\boldsymbol{T}\boldsymbol{y}_j\right)-rank_{ne}\left(\boldsymbol{a}^T\boldsymbol{y}_j\right)\right)}{n_1}$	No
$\hat{\sigma}^2_{\Delta AUC}$ tied ranks	Use DeLong formula [8]	No
$\hat{\sigma}^2_{NRI_{>0}}$	$\frac{\hat{p}_{ne}^{up}\left(1-\hat{p}_{ne}^{up}\right)}{n_0}+\frac{\hat{p}_{e}^{up}\left(1-\hat{p}_{e}^{up}\right)}{n_1}$	No
$\hat{\sigma}^2_{NRI(r)}$	$\frac{\hat{p}_{ne}^{up}+\hat{p}_{ne}^{down}-\left(\hat{p}_{ne}^{up}-\hat{p}_{ne}^{down}\right)^2}{n_0}+\frac{\hat{p}_{ev}^{up}+\hat{p}_{ev}^{down}-\left(\hat{p}_{ev}^{up}-\hat{p}_{ev}^{down}\right)^2}{n_1}$	No
$\hat{\sigma}^2_{3cNRI}$	$\frac{4\left(\hat{p}_{ne}^{2up}+\hat{p}_{ne}^{2down}\right)+\hat{p}_{ne}^{1up}+\hat{p}_{ne}^{1down}-\left(2\left(\hat{p}_{ne}^{2up}-\hat{p}_{ne}^{2down}\right)+\hat{p}_{ne}^{1up}-\hat{p}_{ne}^{1down}\right)^2}{n_0}+\frac{4\left(\hat{p}_{ev}^{2up}+\hat{p}_{ev}^{2down}\right)+\hat{p}_{ev}^{1up}+\hat{p}_{ev}^{1down}-\left(2\left(\hat{p}_{ev}^{2up}-\hat{p}_{ev}^{2down}\right)+\hat{p}_{ev}^{1up}-\hat{p}_{ev}^{1down}\right)^2}{n_1}$	Yes
$\hat{\sigma}^2_{IDI}$	$\frac{Var(\Delta predp(\boldsymbol{x}_i))}{n_0}+\frac{Var\left(\Delta predp\left(\boldsymbol{y}_j\right)\right)}{n_1}$	Yes

Appendix C

C.1. Closed-form expression for 3cNRI under the assumption of normality of predictor variables

Closed-form expression of 3cNRI under the assumption of normality of predictor variables was derived in [9].

$$
\begin{aligned}
3cNRI = {} & \Phi\left(\frac{\frac{M_{new}^2}{2}-\ln\left(\frac{t_H(1-r)}{(1-t_H)r}\right)}{\sqrt{M_{new}^2}}\right)-\Phi\left(\frac{\frac{M_{old}^2}{2}-\ln\left(\frac{t_H(1-r)}{(1-t_H)r}\right)}{\sqrt{M_{old}^2}}\right)+\Phi\left(\frac{\frac{M_{new}^2}{2}-\ln\left(\frac{t_L(1-r)}{(1-t_L)r}\right)}{\sqrt{M_{new}^2}}\right) \\
& -\Phi\left(\frac{\frac{M_{old}^2}{2}-\ln\left(\frac{t_L(1-r)}{(1-t_L)r}\right)}{\sqrt{M_{old}^2}}\right)+\Phi\left(\frac{\frac{M_{new}^2}{2}+\ln\left(\frac{t_H(1-r)}{(1-t_H)r}\right)}{\sqrt{M_{new}^2}}\right)-\Phi\left(\frac{\frac{M_{old}^2}{2}+\ln\left(\frac{t_H(1-r)}{(1-t_H)r}\right)}{\sqrt{M_{old}^2}}\right) \\
& +\Phi\left(\frac{\frac{M_{new}^2}{2}+\ln\left(\frac{t_L(1-r)}{(1-t_L)r}\right)}{\sqrt{M_{new}^2}}\right)-\Phi\left(\frac{\frac{M_{old}^2}{2}+\ln\left(\frac{t_L(1-r)}{(1-t_L)r}\right)}{\sqrt{M_{old}^2}}\right)
\end{aligned}
\tag{A3.1}
$$

where M_{new}^2, M_{old}^2 are squared Mahalanobis distance for new and old models, t_H, t_L, r are high and low thresholds and an event rate.

C.2. Closed-form expression for IDI under the assumption of normality of predictor variables

Closed-form expression of IDI under the assumption of normality of predictor variables was derived in [39].

$$
\begin{aligned}
IDI = {} & \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi M_{new}^2}}\exp\left(\frac{-\left(x-0.5M_{new}^2\right)^2}{2M_{new}^2}\right)\left(\frac{1}{1+\frac{r}{1-r}\cdot\exp(-x)}-\frac{1}{1+\frac{r}{1-r}\cdot\exp(x)}\right)dx \\
& -\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi M_{old}^2}}\exp\left(\frac{-\left(x-0.5M_{old}^2\right)^2}{2M_{old}^2}\right)\left(\frac{1}{1+\frac{r}{1-r}\cdot\exp(-x)}-\frac{1}{1+\frac{r}{1-r}\cdot\exp(x)}\right)dx,
\end{aligned}
\tag{A3.2}
$$

where r is the event rate.

Appendix D

Derivative of closed-form formula of 3cNRI [formula (A3.1)] with respect to event rate r is as follows:

$$(3cNRI)'_r = \frac{1}{\sqrt{2\pi}(r-1)r} \times$$

$$\frac{1}{M_{new}} \left(\begin{array}{l} \exp\left(-\frac{\left(M_{new}^2 - 2\ln\left[\frac{(1-r)t_H}{r(1-t_H)}\right]\right)^2}{8M_{new}^2}\right) - \exp\left(-\frac{\left(M_{new}^2 + 2\ln\left[\frac{(1-r)t_H}{r(1-t_H)}\right]\right)^2}{8M_{new}^2}\right) + \\ \exp\left(-\frac{\left(M_{new}^2 - 2\ln\left[\frac{(1-r)t_L}{r(1-t_l)}\right]\right)^2}{8M_{new}^2}\right) - \exp\left(-\frac{\left(M_{new}^2 + 2\ln\left[\frac{(1-r)t_L}{r(1-t_l)}\right]\right)^2}{8M_{new}^2}\right) \end{array} \right)$$

$$\frac{1}{M_{old}} \left(\begin{array}{l} -\exp\left(-\frac{\left(M_{old}^2 - 2\ln\left[\frac{(1-r)t_H}{r(1-t_H)}\right]\right)^2}{8M_{old}^2}\right) + \exp\left(-\frac{\left(M_{old}^2 - 2\ln\left[\frac{(1-r)t_H}{r(1-t_H)}\right]\right)^2}{8M_{old}^2}\right) - \\ \exp\left(-\frac{\left(M_{old}^2 - 2\ln\left[\frac{(1-r)t_L}{r(1-t_l)}\right]\right)^2}{8M_{old}^2}\right) + \exp\left(-\frac{\left(M_{old}^2 + 2\ln\left[\frac{(1-r)t_L}{r(1-t_l)}\right]\right)^2}{8M_{old}^2}\right) \end{array} \right)$$

Derivative of closed-form formula of IDI [formula (A3.2)] with respect to event rate r is as follows:

$$(IDI)'_r = \frac{1}{\sqrt{2\pi}M_{new}M_{old}} \times \int_{-\infty}^{+\infty} \frac{(-1+e^{2x})(2r-1)}{(r+e^x(1-r))^2(1+(e^x-1)r)^2} \exp\left(x - \frac{(x-0.5M_{old}^2)^2}{2M_{old}^2} - \frac{(x-0.5M_{new}^2)^2}{2M_{new}^2}\right)$$
$$\left(M_{new}\exp\left(\frac{(x-0.5M_{new}^2)^2}{2M_{new}^2} - \frac{(x-0.5M_{new}^2)^2}{2M_{new}^2}\right) - M_{old}\exp\left(\frac{(x-0.5M_{old}^2)^2}{2M_{old}^2} - \frac{(x-0.5M_{new}^2)^2}{2M_{new}^2}\right)\right)dx$$

1. Lee J. U-Statistics: Theory and Practice. 1990.
2. Hoeffding W. A class of statistics with asymptotically normal distribution. *The annals of mathematical statistics* 1948:293–325.
3. Pencina MJ, D'Agostino RB, and Vasan RS. Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond. *Stat Med* 2008; **2**:157–172.
4. Pencina MJ, D'Agostino RB, and Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med* 2011; **1**:11–21.
5. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, USA, 2003.
6. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology* 1975; **4**:387–415.
7. Hanley JA, and McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **1**:29–36.
8. DeLong ER, DeLong DM, and Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988:837–845.
9. Pencina KM, Pencina MJ, and D'Agostino RB. What to expect from net reclassification improvement with three categories. *Stat Med* 2014; **28**:4975–4987.

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; **18**:1837–1847.
2. Expert Panel on Detection E. Executive summary of the third report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel III). *JAMA* 2001; **19**:2486.
3. 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; **24**:1879–1886.
4. Dehmer SP, Maciosek MV, Flottemesch TJ. Aspirin use to prevent cardiovascular disease and colorectal cancer. *Evidence Syntheses* 2015; **2015**:131s.
5. Stone NJ, Robinson J, Lichtenstein AH, Bairey Merz CN, Blum CB, Eckel RH, Goldberg AC, Gordon D, Levy D, Lloyd-Jones DM, McBride P, Schwartz JS, Shero ST, Smith SC Jr, Watson K, Wilson PWF. 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the american college of cardiology/american heart association task force on practice guidelines. *Journal of the American College of Cardiology* 2014; 25_PA:2889–2934.
6. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* 1975; **4**:387–415.
7. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **1**:29–36.
8. Pencina MJ, D'Agostino 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; **2**:157–172.
9. Pencina MJ, D'Agostino RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine* 2011; **1**:11–21.
10. Pencina MJ, Neely B, Steyerberg EW. Re: net risk reclassification P values: valid or misleading? *Journal of the National Cancer Institute* 2015; **1** dju355.
11. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44**:837–845.
12. Demler OV, Pencina MJ, D'Agostino RB Sr. Misuse of DeLong test to compare AUCs for nested models. *Statistics in Medicine* 2012; **23**:2577–2587.
13. Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. Net reclassification indices for evaluating risk-prediction instruments: a critical review. *Epidemiology (Cambridge, Mass.)* 2014; **1**:114.
14. Seshan VE, Gönen M, Begg CB. Comparing ROC curves derived from regression models. *Statistics in Medicine* 2013; **9**:1483–1493.
15. Pencina MJ, Steyerberg EW, D'Agostino RB. Net reclassification index at event rate: properties and relationships. *Statistics in Medicine* 2017. DOI: 10.1002/sim.7041
16. Pencina KM, Pencina MJ, D'Agostino RB. What to expect from net reclassification improvement with three categories. *Statistics in Medicine* 2014; **28**:4975–4987.
17. Klein JP. Small sample moments of some estimators of the variance of the Kaplan–Meier and Nelson–Aalen estimators. *Scandinavian Journal of Statistics* 1991; **18**:333–340.
18. Antolini L, Nam B-H, D'Agostino RB. Inference on correlated discrimination measures in survival analysis: a nonparametric approach. *Communications in statistics-Theory and Methods* 2004; **9**:2117–2135.
19. Lee MLT, Rosner BA. The average area under correlated receiver operating characteristic curves: a nonparametric approach based on generalized two-sample Wilcoxon statistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2001; **3**:337–344.
20. Kerr KF, McClelland RL, Brown ER, Lumley T. Evaluating the incremental value of new biomarkers with integrated discrimination improvement. *American Journal of Epidemiology* 2011; **3**:364–374.
21. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine* 1984; **2**:143–152.
22. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press: USA, 2003.
23. Pepe M, Janes H. Methods for evaluating prediction performance of biomarkers and tests. In Risk Assessment and Evaluation of Predictions. Springer Science+Business Media: New York, 2013; 107–142.
24. So H-C, Sham PC. A unifying framework for evaluating the predictive power of genetic variants based on the level of heritability explained. *PLoS Genetics* 2010; **12** e1001230.
25. Pencina MJ, Fine JP, D'Agostino RB. Discrimination slope and integrated discrimination improvement–properties, relationships and impact of calibration. *Statistics in Medicine* 2016 in press.
26. Pepe MS, Kerr KF, Longton G, Wang Z. Testing for improvement in prediction model performance. *Statistics in Medicine* 2013; **9**:1467–1482.
27. Demler OV, Pencina MJ, D'Agostino RB Sr. Equivalence of improvement in area under ROC curve and linear discriminant analysis coefficient under assumption of normality. *Statistics in Medicine* 2011; **12**:1410–1418.
28. Lee J. U-statistics: theory and practice. Marcel Dekker: New York, 1990.
29. Hoeffding W. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics* 1948; **19**:293–325.
30. Lehmann EL. Consistency and unbiasedness of certain nonparametric tests. *The Annals of Mathematical Statistics* 1951; **22**:165–179.
31. Serfling RJ. Approximation Theorems of Mathematical Statistics, Vol. **162**. John Wiley & Sons: New York, 2009.

32. Sukhatme BV. Testing the hypothesis that two populations differ only in location. *The Annals of Mathematical Statistics* 1958; **29**:60–78.
33. Randles RH. On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics* 1982; **10**:462–474.
34. de Wet T, Randles RH. On the effect of substituting parameter estimators in limiting X2 U and V statistics. *The Annals of Statistics* 1987; **15**:398–412.
35. Efron B. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association* 1975; **352**:892–898.
36. Su JQ, Liu JS. Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association* 1993; **424**:1350–1355.
37. Mardia KV, Kent JT, Bibby JM. Multivariate Analysis (Probability and Mathematical Statistics). Academic Press: London, 1980.
38. Pencina MJ, D'Agostino RB Sr, Demler OV. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. *Statistics in Medicine* 2012; **2**:101–113.
39. Paynter NP, Cook NR. A bias-corrected net reclassification improvement for clinical subgroups. *Medical Decision Making* 2013; **2**:154–162.
40. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care the framingham heart study. *Circulation* 2008; **6**:743–753.
41. Van Der Vaart AW, Wellner JA. Weak Convergence. In Weak Convergence and Empirical Processes. Springer: New York, 1996; 16–28.
42. Babu GJ, Singh K. On one term Edgeworth correction by Efron's bootstrap. *Sankhyā: The Indian Journal of Statistics, Series A* 1984; **46**:219–232.
43. Singh K. On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics* 1981; **9**:1187–1195.

