Eur J Epidemiol (2015) 30:299–304
DOI 10.1007/s10654-015-0001-1

METHODS

A confidence ellipse for the Net Reclassification Improvement

Kristin Mühlenbruch · Olga Kuxhaus · Michael J. Pencina · Heiner Boeing · Hannelore Liero · Matthias B. Schulze

Received: 3 November 2014 / Accepted: 12 February 2015 / Published online: 28 February 2015

Abstract The Net Reclassification Improvement (NRI) has become a popular metric for evaluating improvement in disease prediction models through the past years. The concept is relatively straightforward but usage and interpretation has been different across studies. While no thresholds exist for evaluating the degree of improvement, many studies have relied solely on the significance of the NRI estimate. However, recent studies recommend that statistical testing with the NRI should be avoided. We propose using confidence ellipses around the estimated values of event and non-event NRIs which might provide the best measure of variability around the point estimates. Our developments are illustrated using practical examples from EPIC-Potsdam study.

K. Mühlenbruch (✉) · O. Kuxhaus · M. B. Schulze
Department of Molecular Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Arthur-Scheunert-Allee 114-116, 14558 Nuthetal, Germany
e-mail: Kristin.Muehlenbruch@dife.de

K. Mühlenbruch · O. Kuxhaus · M. B. Schulze
German Center for Diabetes Research (DZD), Nuthetal, Germany

M. J. Pencina
Department of Biostatistics and Bioinformatics, Duke Clinical Research Institute, Duke University, Durham, NC, USA

H. Boeing
Department of Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany

H. Liero
Institute of Mathematics, University of Potsdam, Potsdam, Germany

Keywords Risk assessment · Risk model · Model comparison · Reclassification · Confidence intervals

Background

Risk prediction models have become a main focus in epidemiological research in the past years. Although a large number of prediction models exists, of which some have already been integrated in treatment strategies or health promotion programs, there is an ongoing effort to improve prediction models by the use of new risk markers. For the evaluation of such model extensions, the Net Reclassification Improvement (NRI) was proposed by Pencina et al. in 2008 as an addition to the evaluation of discrimination, e.g. by comparing receiver operating characteristic curves [1]. The NRI is based on the calculation of the amount of correctly and incorrectly reclassified cases and non-cases comparing classification of individuals into a priori defined risk categories in terms of their predicted risk between two nested models. Since its publication it has been used in a growing number of studies, however, there is a large heterogeneity in its use, presentation, and interpretation [2, 3]. Especially with regard to testing statistical significance of NRI estimates, there remains uncertainty. Pencina [4] discussed that even small NRI values (<0.01) might produce statistically significant *p* values and Pepe et al. [5] points out that valid methods for inference for the NRI do not exist. In a recent review of NRI measures, Kerr et al. [2] raise concerns about the proposed test statistic and variance formula. This suggests that statistical testing should be avoided for the NRI measure. However, confidence intervals provide precision estimates and are preferable, not only for the overall NRI, but also for its components. The NRI components do not reflect an overall improvement but rather improvement among cases and non-cases separately. Therefore, our aim was to introduce a method to calculate a confidence ellipse around the two components of the NRI which reflects the precision of the estimates and can help interpret the magnitude and variability of the observed effects.

Definition of NRI

Extension of prediction models with additional risk factors usually leads to changes in predicted risk for individual study participants. When predefined risk categories are used, this is reflected by upward and downward movements across these risk categories from the reference to the extended model. This reclassification is used for the calculation of the *NRI* which considers proportions of upward and downward movements separately for cases and non-cases (1) [1].

$$\begin{aligned} NRI_{cases} &= P(up|case) - P(down|case), \\ NRI_{non-cases} &= P(down|non{-}case) - P(up|non{-}case), \\ NRI &= NRI_{cases} + NRI_{non-cases} \\ &= P(up|case) - P(down|case) \\ &\quad + P(down|non{-}case) - P(up|non{-}case) \\ &= \left[p_{up,cases} - p_{down,cases}\right] \\ &\quad + \left[p_{down,non-cases} - p_{up,non-cases}\right] \end{aligned} \tag{1}$$

The corresponding standard error for the *NRI* and its components was defined by Pencina et al. [4] and depends on the standard error of cases, which often is a much smaller group:

$$SE\left(\widehat{NRI}\right) = \sqrt{SE\left(\widehat{NRI}_{cases}\right)^2 + SE\left(\widehat{NRI}_{non-cases}\right)^2},$$

$$SE\left(\widehat{NRI}_{cases}\right) = \sqrt{\frac{\hat{p}_{up,cases} + \hat{p}_{down,cases} - \left(\hat{p}_{up,cases} - \hat{p}_{down,cases}\right)^2}{N_{cases}}},$$

$$SE\left(\widehat{NRI}_{non-cases}\right) = \sqrt{\frac{\hat{p}_{up,non-cases} + \hat{p}_{down,non-cases} - \left(\hat{p}_{up,non-cases} - \hat{p}_{down,non-cases}\right)^2}{N_{non-cases}}}.$$

As such, the *NRI* is the sum of the single components (NRI_{cases}, $NRI_{non\text{-}cases}$) reflecting improvement among cases or improvement among non-cases or both. Thereby, the overall measure does not include evaluation of improvement among cases or non-cases separately. Absolute risks are derived from regression models; either logistic regression or Cox-regression with the disease as the outcome variable.

Confidence ellipse for two components of NRI

Pencina already suggested to report CIs for the *NRI* and used the bootstrap method for their construction [4]. Calculation of CIs would be informative not only for the overall NRI but also for the single components. Besides the bootstrapping method, CIs can be calculated with a formula related to the construction of CIs for independent proportions according to Agresti [6]; this approach will be applied further on. The standard errors for the overall *NRI* and its single components were defined before, so that the CIs can be defined as follows:

$$\left[\widehat{NRI} - z_{1-\frac{\alpha}{2}}SE\left(\widehat{NRI}\right), \widehat{NRI} + z_{1-\frac{\alpha}{2}}SE(\widehat{NRI})\right]$$

with $z_{1-\frac{\alpha}{2}}$ as the $\left(1-\frac{\alpha}{2}\right)$-quantile of the standard normal distribution. The CIs for NRI_{cases} and $NRI_{non\text{-}cases}$ can be calculated with the same method. While CIs of the two *NRI* components, NRI_{cases} and $NRI_{non\text{-}cases}$, can be interpreted individually, this again would not allow an easy interpretation in terms of the overall improvement. To overcome this problem, we propose to use a confidence ellipse which allows evaluating the single components NRI_{cases} and $NRI_{non\text{-}cases}$ in combination.

We introduce the following notation: Let $\theta = (\theta_{1,}\theta_2)$ be the parameter consisting of the *NRI* components, i.e.

$\theta_1 = NRI_{cases} = P(up|case) - P(down|case)$ and

$$\theta_2 = NRI_{non-cases} = P(down|non{-}case) - P(up|non{-}case)$$

We define the following probabilities

$p_1 = \mathrm{P}(up \cap case)$, $p_2 = \mathrm{P}(down \cap case)$, $p_3 = \mathrm{P}(up \cap non{-}case)$, $p_4 = \mathrm{P}(down \cap non{-}case)$ and $p_5 = \mathrm{P}(case)$

and can write θ as a function of these probabilities:

$$\theta = (\theta_1, \theta_2) = g(\boldsymbol{p}) = (g_1(\boldsymbol{p}), g_2(\boldsymbol{p})) = \left(\frac{p_1 - p_2}{p_5}, \frac{p_4 - p_3}{1 - p_5}\right).$$

Consequently, the maximum likelihood estimates of θ_1 and θ_2 are given by the relative frequencies $\hat{p}_j = v_j/N$ (with $N = N_{cases} + N_{non-cases}$) as follows:

$$\hat{\theta}_1 = \frac{\hat{p}_1 - \hat{p}_2}{\hat{p}_5} \text{ and } \hat{\theta}_2 = \frac{\hat{p}_4 - \hat{p}_3}{1 - \hat{p}_5}$$

with

	Up	*Down*	Total
Case	v_1	v_2	v_5
Non-case	v_3	v_4	$N - v_5$

Applying the multivariate central limit theorem to the vector of relative frequencies $\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4, \hat{p}_5)^T$ we get, that for a large sample size N the distribution of $\sqrt{N}(\hat{\boldsymbol{p}} - \boldsymbol{p})$ can be approximated by a five-dimensional normal distribution, i.e.,

$$\sqrt{N}(\hat{\boldsymbol{p}} - \boldsymbol{p}) \xrightarrow{D} \mathrm{N}_5(0, A(\boldsymbol{p})). \tag{2}$$

Here $A(\boldsymbol{p})$ is the covariance matrix of the limit distribution. It depends on the underlying probabilities p_j and can be computed as:

$$A(\boldsymbol{p}) = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 & -p_1p_4 & p_1(1-p_5) \\ -p_1p_2 & p_2(1-p_2) & -p_2p_3 & -p_2p_4 & p_2(1-p_5) \\ -p_1p_3 & -p_2p_3 & p_3(1-p_3) & -p_3p_4 & -p_3p_5 \\ -p_1p_4 & -p_2p_4 & -p_3p_4 & p_4(1-p_4) & -p_4p_5 \\ p_1(1-p_5) & p_2(1-p_5) & -p_3p_5 & -p_4p_5 & p_5(1-p_5) \end{pmatrix}.$$

With the help of the so-called delta method we can derive from (2) the asymptotic variance of $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$. Here we use, that $\hat{\theta} = g(\hat{\boldsymbol{p}})$. To derive the asymptotic variance of $\hat{\theta}$ one has to multiply the matrix of partial derivatives of g with $A(\boldsymbol{p})$. This leads to

$$\sqrt{N}\left[\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} - \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}\right] \xrightarrow{D} \mathrm{N}_2(0, W(\boldsymbol{p}))$$

with $W(\boldsymbol{p}) = \begin{pmatrix} w_1 & 0 \\ 0 & w_2 \end{pmatrix}$ and

$$w_1 = \frac{p_1 + p_2}{p_5^2} - \frac{(p_1 - p_2)^2}{p_5^3}, w_2 = \frac{p_3 + p_4}{(1-p_5)^2} - \frac{(p_3 - p_4)^2}{(1-p_5)^3}.$$

The asymptotic normality of $\hat{\theta}$ implies that

$$N\left(\hat{\theta} - \theta\right)^T W^{-1}(\hat{\boldsymbol{p}})\left(\hat{\theta} - \theta\right) \xrightarrow{D} \chi_2^2 \tag{3}$$

with χ_2^2, the Chi squared distribution with two degrees of freedom and $W^{-1}(\hat{\boldsymbol{p}})$ is the inverse of the matrix $W(\boldsymbol{p})$.

Because of the diagonal structure of $W(\boldsymbol{p})$ and with asymptotic result from (3) we can define a $(1-\alpha)$ confidence ellipse for θ as

$$\left\{\theta \in [-1,1]^2 \mid \frac{\left(\hat{\theta}_1 - \theta_1\right)^2}{\hat{w}_1} + \frac{\left(\hat{\theta}_2 - \theta_2\right)^2}{\hat{w}_2} \le \frac{\chi^2_{2;1-\alpha}}{N}\right\}.$$

The determination of the confidence ellipse allows to determine the simultaneous precision of the NRI estimates for cases and non-cases.

Using previous notation and the following relationships $\hat{p}_{up,cases} = \hat{p}_1/\hat{p}_5$, $\hat{p}_{down,cases} = \hat{p}_2/\hat{p}_5$, $\hat{p}_{up,non-cases} = \hat{p}_3/(1-\hat{p}_5)$ and $\hat{p}_{up,cases} = \hat{p}_4/(1-\hat{p}_5)$, the confidence ellipse can also be defined with the following equation.

$$\left\{\theta \in [-1,1]^2 \mid \left(\frac{\widehat{NRI}_{cases} - \theta_1}{SE\left(\widehat{NRI}_{cases}\right)}\right)^2 + \left(\frac{\widehat{NRI}_{non-cases} - \theta_2}{SE\left(\widehat{NRI}_{non-cases}\right)}\right)^2 \le \chi^2_{2;1-\alpha}\right\}.$$

Empirical data

Study population

The European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study is a prospective cohort study initially including 27,548 participants aged 35–65 years. Details of recruitment and follow-up procedures were described previously [7, 8]. Briefly, within a median follow-up time of 7 years, 849 participants out of 25,167 participants free of diabetes at baseline developed incident diabetes. On this basis, the German diabetes risk score (GDRS) was developed using Cox-regression [9]. With the GDRS the 5-year risk for developing future type 2 diabetes can be calculated using information on lifestyle and anthropometric factors, diet and physical activity. It serves as the reference model in this underlying model comparison. We used data from 21,846 participants (727 cases) who had also information on family history of diabetes available. The extended model additionally included family history; this model was compared with the reference model. Table 1 shows the reclassification of cases and non-cases due to model extension based on the use of 5 predefined risk categories.

Calculation of Confidence Intervals and Confidence ellipses

Based on the asymptotic method we determined 95 % CIs for *NRI*$_{cases}$ and *NRI*$_{non\text{-}cases}$ (Fig. 1). Taking into account

Table 1 Reclassification table by cases and non-cases resulting from adding family history of diabetes to the German DRS (GDRS), EPIC-Potsdam cohort (N = 21,846)

N (%)	GDRS + family history					Total
	1: Low	2: Still low	3: Increased	4: High	5: Very high	
Cases						
1: Low[a]	*21 (2.89)*	7 (0.96)	–	–	–	28 (3.85)
2. Still low	13 (1.79)	*102 (14.03)*	30 (4.13)	–	–	145 (19.94)
3. Increased	–	32 (4.40)	*176 (24.21)*	61 (8.39)	–	269 (37.00)
4. High	–	–	29 (3.99)	*146 (20.08)*	36 (4.95)	211 (29.02)
5. Very high	–	–	–	15 (2.06)	*59 (8.12)*	74 (10.18)
Total	34 (4.68)	141 (19.39)	235 (32.32)	222 (30.54)	95 (13.07)	727 (100)
Non-cases						
1. Low	*9001 (42.62)*	625 (2.96)	–	–	–	9626 (45.58)
2. Still low	1415 (6.70)	*4220 (19.98)*	672 (3.18)	–	–	6307 (29.86)
3. Increased	–	858 (4.06)	*2613 (12.37)*	387 (1.83)	–	3858 (18.27)
4. High	–	–	269 (1.27)	*782 (3.70)*	98 (0.46)	1149 (5.44)
5. Very high	–	–	–	40 (0.19)	*139 (0.66)*	179 (0.85)
Total	10,416 (49.32)	5703 (27.0)	3554 (16.83)	1209 (5.72)	237 (1.12)	21,119 (100)

[a] Risk categories were created according to score points of the German Diabetes Risk Score: low risk: <410 points (5-year risk < 0.88 %); still low: 410–<510 (0.88–<2.37 %); increased risk: 510–<610 (2.37–<6.30 %); high risk: 610–<710 (6.30–<16.21 %); very high risk: ≥710 (≥16.21 %)

NRI measures were calculated as follows

$NRI_{cases} = ((0.96 + 4.13 + 8.39 + 4.95) - (1.79 + 4.40 + 3.99 + 2.06))/100 = (18.43 - 12.24)/100 = 0.0619$

$NRI_{non-cases} = ((6.70 + 4.06 + 1.27 + 0.19) - (2.96 + 3.18 + 1.83 + 0.46))/100 = (12.22 - 8.43)/100 = 0.0379$

$NRI = 0.0619 + 0.0379 = 0.0998$

the large number of non-cases it is obvious that estimation of $NRI_{non\text{-}cases}$ was much more precise than of NRI_{cases}. The calculation of CIs for single components does not allow evaluating both components in combination.

Therefore, we computed a confidence ellipse for NRI_{cases} and $NRI_{non\text{-}cases}$ to reflect precision of their estimates in combination and which also allows to evaluate the area of acceptable values. Figure 1 shows CIs for the single components (vertical and horizontal lines) as well as the confidence ellipse, both approaches were based on the five risk categories described before. When constructing CIs for the components separately, NRI_{cases} (0.0619) has a CI of 0.0219–0.1019. Therefore, the value 0.02 lies outside of this interval while the $NRI_{non\text{-}cases}$ (0.0379) had a CI ranging from 0.0318 to 0.0440 thus including a value of 0.035. Using both CIs separately would therefore lead to the conclusion that NRI_{cases} is significantly higher than 0.02 while $NRI_{non\text{-}cases}$ is not significantly higher than 0.035. However, examining the vector (0.02, 0.035) within the confidence ellipse we can see that it is located inside the area of the ellipse. Thus, the confidence ellipse indicates that—when evaluated together—neither is the NRI_{cases} different from 0.02 nor is the $NRI_{non\text{-}cases}$ different from 0.035. This example clearly indicates that evaluating single *NRI* components separately might result in different decisions than evaluating the single components in combination by the use of confidence ellipses.

These results were based on the asymptotic method for both the calculation of CIs and of the confidence ellipse.

Discussion

The use of the *NRI* is informative for the evaluation of improvements of prediction models when taking into account the obvious limitations associated with the use of categories and cut-offs. Given that no established cut-offs for the *NRI* exist which allow interpreting its value as being meaningful from a clinical or public health point of view, reliance solely on significance testing has been frequently adopted in reclassification analyses.

As recommended in a recent review of the NRI methods [3], it is preferable to investigate model improvement separately for cases or non-cases. A general framework for testing the two components of the overall *NRI*, NRI_{cases} and $NRI_{non\text{-}cases}$, has previously been laid out by Pencina et al. [1]. However, a major drawback of examining single components in isolation is that the results cannot be

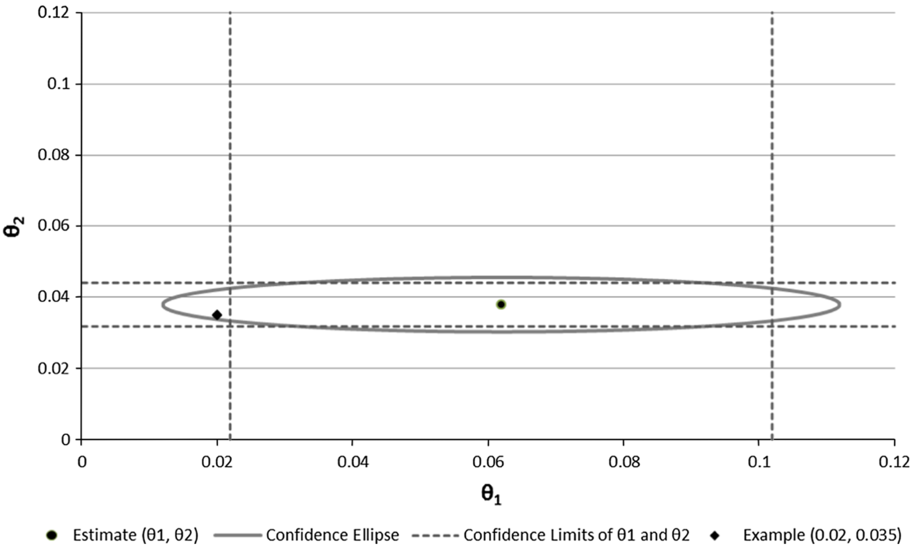

Fig. 1 Confidence ellipse for the two-dimensional estimate $\theta = (\theta_1, \theta_2)$ of the calculated NRI_{cases} and $NRI_{non\text{-}cases}$, EPIC-Potsdam study. *NRI* for cases and *NRI* for non-cases were calculated by comparing the German diabetes risk score extended with family history of diabetes with the initial German diabetes risk score; five predefined risk categories were used for calculation; the confidence ellipse was defined with the accepted θ_1 and θ_2 values for $\alpha = 0.05$. The area within the ellipse defines the area of accepting the null hypotheses $H_0 : NRI_{cases} = \theta_1$ *and* $NRI_{non-cases} = \theta_2$ and the area outside the ellipse is the area of not accepting H_0. The *black dot* is the estimated θ for NRI_{cases} and $NRI_{non\text{-}cases}$. The *horizontal* and *vertical reference lines* display the lower and upper confidence limits of $NRI_{non\text{-}cases}$ (0.0318, 0.0440) and of NRI_{cases} (0.0219, 0.1019) respectively. The *black diamond* displays the discussed example for null hypotheses: $NRI_{cases} = 0.02$ *and* $NRI_{non\text{-}cases} = 0.035$

interpreted in terms of the overall model improvement. We note that recent recommendations suggest not applying statistical testing at all [2, 3]. Likewise, our developments facilitate the use of confidence intervals. A particularly appealing approach is based on using the confidence ellipse which reflects the 2-dimensional nature of the situation. Our empirical example indicates that confidence ellipses can be useful in reflecting both, the precision of the *NRI* estimation as well as putting the results in the context of overall improvement.

Our proposed method of confidence ellipses is also flexible here as it can be applied to evaluating extensions of prediction models using equal or different weights as well as thresholds of acceptable model improvement for cases and non-cases as already discussed by Greenland [10].

In conclusion, confidence ellipses might be particularly useful in the context of evaluating overall or case- versus non-case-specific model improvement as they allow evaluating varying acceptable values of the *NRI* components in combination and also reflect the precision of their estimates.

Acknowledgments This work was supported in part by a grant from the German Federal Ministry of Education and Research (BMBF) to the German Center for Diabetes Research (DZD e.V.). The recruitment phase of the EPIC-Potsdam Study was supported by the Federal Ministry of Science, Germany (01 EA 9401) and the European Union (SOC 95201408 05F02). The follow-up of the EPIC-Potsdam Study was supported by German Cancer Aid (70-2488-Ha I) and the European Community (SOC 98200769 05F02). We thank Dr. Manuela Bergmann who was responsible for the methodological and organisational work of data collections of exposures and outcomes and Wolfgang Fleischhauer for his medical expertise that was employed in case ascertainment and contacts with the physicians and Ellen Kohlsdorf for data management.

References

1. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008;27(2):157–72.
2. Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. Net reclassification indices for evaluating risk prediction instruments: a critical review. Epidemiology. 2014;25(1):114–21.
3. Leening MJG, Vedder MM, Witteman JCM, Pencina MJ, Steyerberg EW. Net Reclassification Improvement: computation, interpretation, and controversies a literature review and clinician's guide. Ann Intern Med. 2014;160(2):122–31.
4. Pencina MJ, D'Agostino RB Sr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med. 2011;30(1):11–21.
5. Pepe MS, Kerr KF, Longton G, Wang Z. Testing for improvement in prediction model performance. Stat Med. 2013;32(9):1467–82.
6. Agresti A. Categorical Data Analysis. Balding DJ, Bloomfield P, Cressie NAC, editors. Hoboken: Wiley; 2002.

7. Bergmann MM, Bussas U, Boeing H. Follow-up procedures in EPIC-Germany—data quality aspects. European prospective investigation into cancer and nutrition. Ann Nutr Metab. 1999;43(4):225–34.
8. Boeing H, Korfmann A, Bergmann MM. Recruitment procedures of EPIC-Germany. European Investigation into Cancer and Nutrition. Ann Nutr Metab. 1999;43(4):205–15.
9. Schulze MB, Hoffmann K, Boeing H, Linseisen J, Rohrmann S, Mohlig M, et al. An accurate risk score based on anthropometric, dietary, and lifestyle factors to predict the development of type 2 diabetes. Diabetes Care. 2007;30(3):510–5.
10. Greenland S. The need for reorientation toward cost-effective prediction: comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina et al., Statistics in Medicine (DOI: 10.1002/sim.2929). Stat Med. 2008;27(2):199–206.