Eur J Epidemiol (2013) 28:25–33
DOI 10.1007/s10654-012-9744-0

METHODS

Assessing improvement in disease prediction using net reclassification improvement: impact of risk cut-offs and number of risk categories

Kristin Mühlenbruch · Alexandros Heraclides · Ewout W. Steyerberg · Hans-Georg Joost · Heiner Boeing · Matthias B. Schulze

Received: 2 August 2012 / Accepted: 26 October 2012 / Published online: 20 November 2012

Abstract Net reclassification improvement (NRI) has received much attention for comparing risk prediction models, and might be preferable over the area under the receiver operating characteristics (ROC) curve to indicate changes in predictive ability. We investigated the influence of the choice of risk cut-offs and number of risk categories on the NRI. Using data of the European Prospective Investigation into Cancer and Nutrition-Potsdam study, three diabetes prediction models were compared according to ROC area and NRI with varying cut-offs for two and three risk categories and varying numbers of risk categories. When compared with a basic model, including age, anthropometry, and hypertension status, a model extension by waist circumference improved discrimination from 0.720 to 0.831 (0.111 [0.097–0.125]) while increase in ROC-AUC from 0.831 to 0.836 (0.006 [0.002–0.009]) indicated moderate improvement when additionally considering diet and physical activity. However, NRI based on these two model comparisons varied with varying cut-offs for two (range: 5.59–23.20 %; −0.79 to 4.09 %) and three risk categories (20.37–40.15 %; 1.22–4.34 %). This variation was more pronounced in the model extension showing a larger difference in ROC-AUC. NRI increased with increasing numbers of categories from minimum NRIs of 18.41 and 0.46 % to approximately category-free NRIs of 79.61 and 19.22 %, but not monotonically. There was a similar pattern for this increase in both model comparisons. In conclusion, the choice of risk cut-offs and number of categories has a substantial impact on NRI. A limited number of categories should only be used if categories have strong clinical importance.

Keywords Diabetes mellitus, type 2 · Models · Statistical · Risk assessment · Risk model · Model comparison · Model discrimination

Background

Models for risk prediction are widely used in clinical practice to stratify risk and assign treatment or prevention strategies [1–5]. Whether new risk factors can contribute to existing prediction equations in terms of clinical utility is a question that has received growing attention in epidemiological research. Traditionally, in particular, the area under the receiver operating characteristics curve (ROC-AUC) has been used as a measure of discrimination for assessing the improvement of prediction models when new risk factors are added. However, ROC analyses have been criticized for being too conservative for a meaningful improvement in prediction, because increases can only be

Electronic supplementary material The online version of this article (doi:10.1007/s10654-012-9744-0) contains supplementary material, which is available to authorized users.

K. Mühlenbruch (✉) · A. Heraclides · 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

E. W. Steyerberg
Department of Public Health, Erasmus MC, Rotterdam, The Netherlands

H.-G. Joost
Department of Pharmacology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany

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

seen for variables that carry a very high relative risk for the disease of interest [6, 7].

The net reclassification improvement (NRI) proposed by Pencina et al. [8, 9] assesses improvement in the classification of participants into categories of absolute risk, if new risk factors are added to a risk prediction model. In other words, the NRI reflects the net proportion of people reclassified into the correct direction among cases and non-cases. While the NRI has increasingly been used to evaluate extensions of risk prediction models, its application is not without limitations. Tzoulaki et al. [10] highlighted in their overview of studies published until 2010 that varying cut-offs, number of categories and follow-up durations were used for evaluation of reclassification and it has been reported that the value of the NRI may depend on the choice of cut-off for identifying the high-risk category in a binary risk stratification [8, 11]. In addition, in the case where the estimated absolute risk is categorized into more than two categories, previous reports have indicated that the choice of the number of risk categories may also influence the NRI [8, 12]. Nevertheless, a systematic investigation of the impact of both varying cut-offs for two and more categories and varying number of risk categories on the NRI is lacking. The aim of this study was to evaluate the influence of the choice of cut-off values and number of risk categories on the NRI using a large prospective cohort study on incidence of type 2 diabetes.

Materials and methods

Study setting

Data from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam study were used for the analyses. The EPIC-Potsdam study is a prospective cohort study comprising 27,548 participants within the age range of 35–65 years from the general population of Potsdam, recruited between 1994 and 1998 [13]. Participants were followed-up every 2–3 years with questionnaires. Response rates for follow-up rounds 1, 2, 3 and 4 were 96, 95, 91 and 90 % (by 31 August 2005). After exclusion of prevalent self-reported diabetes cases, missing follow-up information, and missing information on important covariates, the study sample consisted of 25,167 participants. Within a mean follow-up period of 7 years, 849 incident diabetes cases were detected [14].

Statistical analysis

Diabetes prediction models

The diabetes prediction models used in the current analysis are based on the German Diabetes Risk Score, a prediction equation developed in this cohort and validated in an external cohort [14]. The German Diabetes Risk Score consists of the following risk factors used to estimate the 5-year absolute risk for type 2 diabetes: age (continuous per year), height (cont. per cm), waist circumference (cont. per cm), prevalent hypertension (yes vs. no), physical activity (cont. per h/week), smoking (currently smoking $\geq$20 cig./day, ex-smoking vs. never smoker or currently smoking $<$ 20 cig./day), alcohol intake (moderate consumption [10–40 g/day] vs. low or high consumption), intake of red meat (cont. per 150 g/day), intake of wholegrain bread (cont. per 50 g/day) and coffee consumption (cont. per 150 g/day). Using these variables, three different risk prediction models were created for this study. Model 1 contained age, height and hypertension status. Model 2 included age, height, hypertension status and additionally waist circumference. Model 3 included all variables of model 2 as well as dietary and lifestyle components of the German Diabetes Risk Score (red meat, wholegrain bread, coffee, alcohol consumption, smoking status and physical activity). Model 1 and model 2 were refitted to EPIC-Potsdam study data while for model 3 published coefficients of the German Diabetes Risk Score were used. The 5-year risk for incidence of type 2 diabetes was estimated with Cox proportional hazards regression analysis for all the models.

Measures of discrimination and reclassification

We compared model 2 with model 1 as well as model 3 with model 2 to evaluate model extensions in terms of improvement in discrimination and risk classification. ROC analysis was used for evaluating the ability to discriminate between incident diabetes cases and non-cases with the corresponding estimate and 95 %-confidence interval (95%-CI) for the difference in the ROC-AUCs between models calculated using the method by DeLong et al. [15]. We did not use the method by DeLong to test for significance because a recent publication showed that this test is invalid and overly conservative for nested models [16]. Improvement in risk classification was determined by reclassification analysis with the NRI estimated separately among cases and non-cases [17]. The NRI among cases was calculated as the proportion of cases moving up in a risk category minus that moving down in a risk category. The NRI among non-cases was calculated as the proportion of non-cases moving down in a risk category minus that of non-cases moving up in a risk category [8, 17]. In other words, the NRI among cases determines the net proportion of more correctly reclassified cases (i.e. moving up to a higher risk category) and that among non-cases provides the net proportion of more correctly reclassified non-cases (i.e. moving down to a lower risk category). The sum of the

Table 1 Reclassification table comparing 5-year risk strata for models that include risk factors for type 2 diabetes in the EPIC-Potsdam study with and without waist circumference

	Model 2[b]					
Model 1[a]	low	still low	increased	high	very high	Total
Cases						
low[c]	12 (1.41)	7 (**0.82**)	9 (**1.06**)	2 (**0.24**)	0 (0)	30 (3.53)
still low	17 (**2.00**)	90 (10.60)	92 (**10.84**)	41 (**4.83**)	8 (**0.94**)	248 (29.21)
increased	2 (**0.24**)	57 (**6.71**)	170 (20.02)	142 (**16.73**)	49 (**5.77**)	420 (49.47)
high	0 (0)	3 (**0.35**)	51 (**6.01**)	67 (7.89)	30 (**3.53**)	151 (17.79)
very high	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	31 (3.65)	157 (18.49)	322 (37.93)	252 (29.68)	87 (10.25)	849 (100)
Non-Cases						
low	4,512 (18.55)	515 (**2.12**)	94 (**0.39**)	18 (**0.07**)	3 (**0.01**)	5,142 (21.14)
still low	4,978 (**20.47**)	3,984 (16.38)	1,373 (**5.65**)	201 (**0.83**)	32 (**0.13**)	10,568 (43.46)
increased	814 (**3.35**)	2,695 (**11.08**)	2,598 (10.68)	854 (**3.51**)	120 (**0.49**)	7,081 (29.12)
high	16 (**0.07**)	241 (**0.99**)	689 (**2.83**)	489 (2.01)	92 (**0.38**)	1,527 (6.28)
very high	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Total	10,320 (42.44)	7,435 (30.57)	4,754 (19.55)	1,562 (6.42)	247 (1.02)	24,318 (100)

[a] Model 1 includes age (years), height (cm) and prevalent hypertension (no/yes)

[b] Model 2 includes age (years), height (cm), prevalent hypertension (no/yes) and waist circumference (cm)

[c] 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%).

two individual components is the overall NRI. We initially evaluated NRI using as an example five risk categories with cut-offs at 0.88, 2.37, 6.30, and 16.21 % 5-year risk, which is according to the current use of the German Diabetes Risk Score. These five risk categories include specific practical implications but have not been based on cost-benefit-analyses.

Evaluation of risk cut-offs

The simplest way of classifying individuals based on their predicted risks from a risk prediction model is to use a single cut-off creating two risk categories (i.e. low risk vs. high risk). The impact of the choice of cut-off values on the NRI was calculated for risk cut-offs varying from 1 to 0.20 % increments. Furthermore, we evaluated the impact of varying cut-offs on NRI using three risk categories Cut-offs were based on deciles of the distribution of absolute risks.

Evaluation of the number of risk categories

To investigate the influence of the number of risk categories of the risk prediction model [12] on the value of the NRI, we used varying numbers of categories, ranging from 2 up to 50. Category cut-offs were based on quantiles of the distribution of absolute risk, so that the study population was equally distributed across the risk categories. We calculated the "category-free NRI" (or continuous NRI) as a reference, which takes into account each upward or downward movement in the estimated risks from the shorter prediction model to the extended model based on the case/non-case status [9].

All statistical analyses were performed with SAS (Version 9.2, Enterprise Guide 4.3, SAS Institute Inc., Cary, NC, USA). For the calculation of the NRI, we used a published SAS macro [18] modified for our purposes. The significance level was defined with a two-tailed p value of <0.05.

Results

We first evaluated the impact of adding waist circumference to age, height and hypertension status (model 2 vs. model 1) using five categories of risk. This led to 9433 (38.79 %) non-cases being reclassified into a lower risk category and 3302 (13.58 %) being reclassified into a higher risk category (Table 1), yielding an NRI of 25.21 % among the non-cases. Among the cases, 380 (44.76 %) moved up a risk

Table 2 Reclassification table comparing 5-year risk strata for models that include risk factors for type 2 diabetes in the EPIC-Potsdam study with and without lifestyle risk factors

	Model 3[b]					
Model 2[a]	low	still low	increased	high	very high	Total
Cases						
low[c]	25 (2.94)	6 (**0.71**)	0 (0)	0 (0)	0 (0)	31 (3.65)
still low	8 (**0.94**)	124 (14.61)	25 (**2.94**)	0 (0)	0 (0)	157 (18.49)
increased	0 (0)	26 (**3.06**)	258 (30.39)	38 (**4.48**)	0 (0)	322 (37.93)
high	0 (0)	0 (0)	28 (**3.30**)	203 (23.91)	21 (**2.47**)	252 (29.68)
very high	0 (0)	0 (0)	0 (0)	16 (**1.88**)	71 (8.36)	87 (10.25)
Total	33 (3.89)	156 (18.37)	311 (36.63)	257 (30.27)	92 (10.84)	849 (100)
Non-Cases						
low	9,723 (39.98)	596 (**2.45**)	1 (**<0.01**)	0 (0)	0 (0)	10,320 (42.44)
still low	924 (**3.80**)	5,943 (24.44)	568 (**2.34**)	0 (0)	0 (0)	7,435 (30.57)
increased	0 (0)	718 (**2.95**)	3,761 (15.47)	275 (**1.13**)	0 (0)	4,754 (19.55)
high	0 (0)	0 (0)	321 (**1.32**)	1,180 (4.85)	61 (**0.25**)	1,562 (6.42)
very high	0 (0)	0 (0)	0 (0)	44 (**0.18**)	203 (0.83)	247 (1.02)
Total	10,647 (43.8)	7,257 (29.8)	4,651 (19.1)	1,499 (6.2)	264 (1.1)	24,318 (100)

[a] Model 2 includes age (years), height (cm), prevalent hypertension (no/yes) and waist circumference (cm)

[b] Model 3 includes age (years), height (cm), prevalent hypertension (no/yes), waist circumference (cm), physical activity (h/week), current smoker (≥20 cig./d) (no/yes), ex-smoker, moderate alcohol consumption (10-40 g/d,)(yes/no), intake of red meat (150 g/d), intake of wholegrain bread (50 g/d) and coffee consumption (150 g/d)

[c] 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%).

category (correct assignment) and 130 (15.31 %) moved down a risk category, resulting in an NRI of 29.45 %. The overall NRI was thus 54.66 % ($p < 0.0001$) and ROC-AUC (95 % CI) increased from 0.720 (0.704–0.735) to 0.831 (0.819–0.843) by 0.111 (0.009–0.125).

Table 2 shows the reclassification comparing model 3 with model 2. After the addition of diet, physical activity, alcohol consumption, and smoking to age, height, hypertension status and waist circumference, 2007 (8.25 %) and 1501 (6.17 %) non-cases were reclassified to a lower risk and higher risk category, respectively, yielding an NRI of 2.08 % for the non-cases. Furthermore, 90 (10.60 %) and 78 (9.18 %) cases were reclassified to a higher risk and lower risk category, respectively. This resulted in an NRI of 1.42 % for the cases. The overall NRI was 3.50 % (p: 0.024). The ROC-AUCs (95 % CI) were 0.831 (0.819–0.843) for model 2 and 0.836 (0.824–0.848) for model 3 with increase by 0.006 (0.002–0.009).

To illustrate the location of the analyzed cut-off values in the risk distribution of this study population, Fig. 1 shows absolute risks derived from Cox proportional hazards regression for model 3.

Risk cut-offs and NRI

Figure 2 shows the NRI (%) comparing model 2 with model 1 and model 3 with model 2 based on two risk categories for cut-off values of absolute risk ranging from 1 to 20 %. The NRI varied with varying cut-off values, however, the two model comparisons resulted in NRI distributions with different patterns. NRI values were higher and their variation was more pronounced for the comparison of model 2 with model 1 (minimum NRI: 5.59 %, maximum NRI: 23.2 %) compared to the comparison of model 3 with model 2 (−0.79 and 4.09 %, respectively). Also, the minimum and maximum NRI values were at different cut-offs. While there was no clear trend for the value of the NRI with cut-offs below or above those with maximum NRI for both model comparisons, the NRI steadily decreased above the cut-off of 6.8 %-risk for the comparison of model 2 with model 1.

We subsequently evaluated the dependence of the NRI on cut-offs using three categories of absolute risk (Fig. 3) using different decile combinations to define risk categories. Again, the patterns of NRI for the two model comparisons were different and variation was stronger for the comparison of model 2 with model 1. Widening the middle risk category around, below or above the median population risk, all resulted in considerably different NRI values here (min 20.37 %, max 40.15 %). NRI tended to increase with a wider middle risk category, except when the middle risk category spanned into the second risk decile. Variation seemed also more pronounced when the middle risk category was located below the median population risk (NRI when middle risk category spanned 1, 2, 3, or 4 deciles: 25.9, 33.4, 36.5, and 20.4 %). In contrast, the comparison of model 3 with model 2 revealed only little difference in the NRI values, ranging between 1.22 and 4.34 %.

Number of risk categories and NRI

Figure 4 shows the NRI values (comparing model 2 with model 1 and model 3 with model 2) according to increasing numbers of equally distributed risk categories from 2 up to 50. The minimum NRI values were observed for the two-and three-category risk classifications (18.41 % for comparison of model 2 with model 1, 0.46 % for comparison of model 3 with model 2). Although the NRI increased with the increasing number of risk categories converging to the category-free NRI (79.61 and 19.22 %) for both model comparisons, this increase was not monotonic. Also, the increase in NRI was steeper and became stable at the category-free NRI much earlier for the comparison of model 2 with model 1 than for the comparison of model 3 with model 2.

Discussion

We have demonstrated that the improvement in risk classification assessed with the NRI may strongly depend on the choice of cut-offs to categorize risk as well as on the number of risk categories. We observed substantial improvement in the prediction of type 2 diabetes risk when waist circumference was added to a model including age, height and hypertension. A weak improvement was observed when lifestyle variables (smoking, alcohol, physical activity, diet) were added to the model including waist circumference, both based on NRI and ROC-AUC. However, the impact of cut-off choice on NRI was stronger for the model extension resulting in larger differences in ROC-AUC.

Our finding that the NRI strongly depended on the choice of cut-offs is in agreement with previous observations. Steyerberg et al. compared two different cut-offs for the evaluation of model extensions with the NRI based on two risk categories. One cut-off led to the maximum of the sum of sensitivity and specificity (a 5.6 % threshold for 10-year risk) and the other was a commonly used cut-off for cardiovascular risk prediction models (20 % for ten-year risk) [19]. The difference in NRI between the two cut-offs was 1.2 % points. In our study, the range of NRI values observed with varying cut-offs was larger, especially up to the cut-off of 20 % 5-year risk—a range containing the vast majority of participants in our study (~90 %). Similarly, Mihaescu et al. [11] reported that the NRI for two categories varied substantially within the range of most frequently predicted risk. Furthermore, we observed that the variation in NRI was related to the difference in ROC-AUC. This has also been observed in a simulation study [11]. In our study, even for the second model extension from model 2 to model 3 which led to a difference in ROC-AUC of only 0.006, we could observe a significant NRI based on five risk categories (3.5 %) or category-free (around 20 %). This might raise the question whether lifestyle variables are useful to be included in the prediction model. However, there is no general agreement on how to judge improvements reflected by delta ROC-AUC or the NRI in terms of their clinical or public health relevance. In case of the German Diabetes Risk Score, only statistically significant risk factors have been included in the model including lifestyle factors [14]. Because its application is relatively simple, the information gained by including lifestyle risk factors likely outweighs the additional effort needed to collect these data. While the NRI might be useful to reflect the benefits of reclassification at the population level and may thus be more helpful for the interpretation of relevant or meaningful improvements, NRI values varied with varying cut-offs. We could even observe negative values indicating worsening in reclassification by model extension. Steyerberg et al. [19] stated that the maximum NRI when two risk classes are used could be expected at the population incidence as a cut-off.

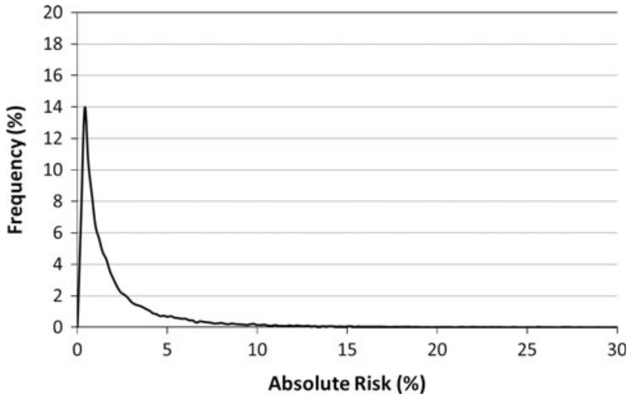

Fig. 1 Distribution of absolute risks derived from Cox-regression for the German Diabetes Risk Score from 1 to 30 % of absolute risk for, EPIC-Potsdam study. Absolute risk: 5-year risk for type 2 diabetes ranging from 0.05 to 91.26 % in the EPIC-Potsdam population; median risk is 1.26 %; and the 99% percentile is 20.42 % absolute risk

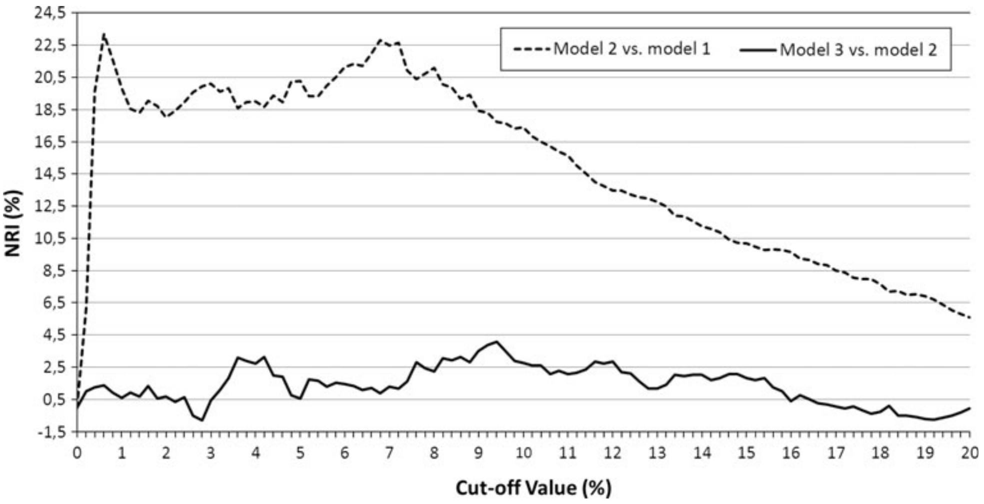

Fig. 2 Net reclassification improvement (NRI) according to cut-off values from 1 to 20 % of absolute risk for two risk categories, EPIC-Potsdam study. NRI was calculated by comparing (1) a model including age, height, prevalent hypertension and waist circumference (model 2) (*dashed line*) with a model including age, height and prevalent hypertension (model 1), and (2) a model including age, height, prevalent hypertension, waist circumference and lifestyle factors (alcohol consumption, smoking, diet, and physical activity) (model 3) with a model that including age, height, prevalent hypertension and waist circumference (model 2) (*solid line*); absolute risk: 5-year risk for type 2 diabetes ranging from 0.05 to 92 % in the EPIC-Potsdam population; median risk is 1.26 %; and the 99 % percentile is 20.42 %

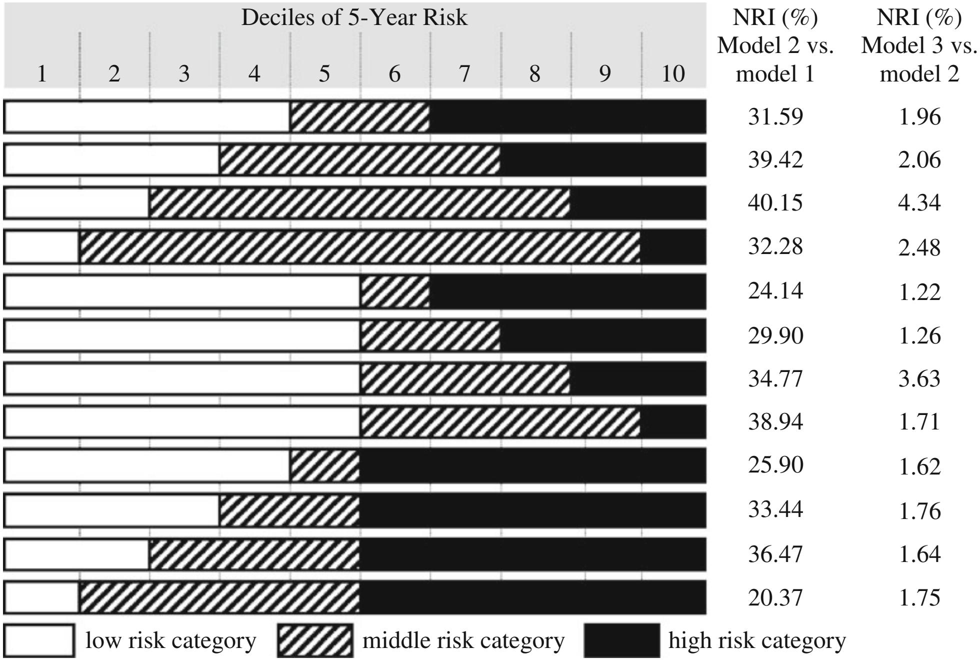

Fig. 3 Net reclassification improvement (NRI) according to varying decile cut-off values of absolute risk for three risk categories determined with two model extensions. NRI was calculated by comparing (1) a model including age, height, prevalent hypertension and waist circumference (model 2) with a model including age, height and prevalent hypertension (model 1), and (2) a model including age, height, prevalent hypertension, waist circumference and lifestyle factors (alcohol consumption, smoking, diet, and physical activity) (model 3) with a model that including age, height, prevalent hypertension and waist circumference (model 2); absolute risk: 5-year risk for type 2 diabetes with decile cut-offs: 0.27, 0.43, 0.64, 0.91, 1.26 (median), 1.73, 2.44, 3.69, and 6.26 %

However, this was not the case in our data, which might be explained by the low median risk and small range of the risk distribution. Also, the calibration of the GDRS was not perfect when analyzing the five predefined risk categories (*p* Hosmer–Lemeshow-Test: 0.0016, Supplementary Fig. 1). Especially for the highest risk group, the average

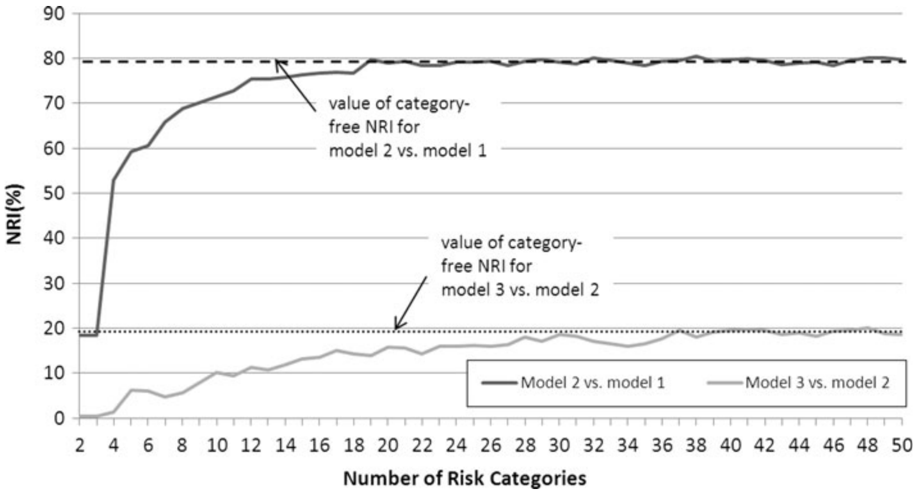

Fig. 4 Net reclassification improvement (NRI) according to different numbers of risk categories (2–50), EPIC-Potsdam study. NRI for 2 up to 50 risk categories was calculated comparing model 2 (age, height, prevalent hypertension and waist circumference) with model 1 (age, height and prevalent hypertension) (*black solid line*) and comparing model 3 (age, height, prevalent hypertension, waist circumference, alcohol consumption, smoking, diet and physical activity) with model 2 (age, height, prevalent hypertension and waist circumference) (*grey solid line*). As reference line the category-free (continuous) NRI was calculated comparing model 2 (age, height, prevalent hypertension and waist circumference) with model 1 (age, height and prevalent hypertension) (*dashed line*) and model 3 (age, height, prevalent hypertension, waist circumference, alcohol consumption, smoking, diet and physical activity) with model 2 (age, height, prevalent hypertension and waist circumference) (*dotted line*)

predicted risk overestimated the observed incidence. This might be due at least in part to outliers and the overall much larger variability of predicted risks in this category as well as the small number of individuals. While we cannot rule out that lack of calibration might have affected our reclassification evaluation for risk categories using cut-offs from the upper end of the risk distribution (e.g. >16 %), it remains unclear if it also affected NRI analyses with lower risk cut-offs given that calibration over a large range of absolute risk was quite acceptable.

We also observed substantial differences in NRI depending on the cut-offs when we analyzed the reclassification based on three risk categories. Mealiffe et al. [20] reported that the NRI increased with increasing distance between two cut-offs. This was not observable in our study for the model extension resulting in only a small difference in ROC-AUC. Also, even with considerable difference in ROC-AUC (comparison of model 2 with model 1) a larger distance between two cutoffs resulted in decreasing NRI in case that the middle risk category was extended to lower risk deciles. In comparison to Mealiffe et al., we only analyzed a limited set of cut-off values and our results may not reflect a general trend.

The number of risk categories also had substantial influence on the NRI, with higher NRI values observed with higher numbers of risk categories. The trend was not monotonic, especially for smaller delta AUC, which supports the hypothesis of Pencina et al. [9] that not each increase in risk categories should result in an increase in NRI in empirical data. In two recent reports, a higher number of risk categories resulted in increasing NRI values [12, 20]. However, the authors investigated only a limited number of possible risk categories, which might have masked the full variation in NRI. For both model comparisons the pattern of NRI changes with increasing numbers of risk categories was comparable, but at a higher level for the model comparison with larger difference in ROC-AUC.

Another interesting aspect of our results was the link between the NRI with categories and the category-free NRI. The category-free NRI includes each change in the predicted risks from the basic to the extended model, which is more objective [9], but may take on substantially larger values than category-based NRIs [12]. In our study, NRI values converged to the category-free NRI for high numbers of categories and this was achieved at a lower number of risk categories for the model extension resulting in a larger difference in ROC-AUC. Furthermore, the category-free NRI did not always show a higher value than the NRI for categories and differences to the category-free NRI were relatively small when using a large number of risk categories. This observation is somewhat in contrast to previous reports [12, 19, 21]; however, these studies evaluated only a limited number of risk strata. Moreover, Cook and Paynter [22] reported even negative values for the category-free NRI and higher values for the NRI based on limited numbers of risk categories.

It should be noted that the derivation of prediction models would usually require a validation of model performance, preferably in independent cohorts. Such data have previously been reported for the full model used in our analyses [14]. However, the focus in the current study was to evaluate the performance of the NRI under different assumptions based on an empirical sample, and not to develop a prediction model.

While our results highlight the dependence of the NRI on risk cut-offs, numbers of categories, and also the difference in ROC-AUC at the same time in empirical data, the NRI has also been shown to be influenced by the ROC-AUC of the reference model [11] and to be related to the incidence of the disease of interest [11, 21]. This strongly supports the recommendation by Pencina et al. that the evaluation of NRI should be based on a priori and clinically meaningful risk categories. For treatment and prevention strategies in cardiovascular diseases such meaningful risk categories have already been established with the Framingham Risk Score. The Adult Treatment Panel III (ATP III) guidelines recommend the use of these categories (<10, 10–20 and >20 % 10-year risk) in addition to main risk factors which results in different consequences for therapy [23, 24]. Unfortunately, categories of risk with different clinical indications have not been established for type 2 diabetes so far and varying cut-offs, number of categories and follow-up durations were used to evaluate reclassification for diabetes prediction models, thus introduces subjectivity in this area [10]. The category-free (continuous) NRI is independent of cut-offs and categories and even of calibration [25] and could therefore be an objective alternative.

Mihaescu et al. [11] observed a bimodal distribution of the NRI with varying cut-offs for a single risk threshold. The two peaks observed corresponded to the maximum net improvement for cases and non-cases. A risk threshold at the disease risk, as suggested before was sub-optimal. Improvements among cases (or non-cases) generally outweigh the worsening of risk classification among the non-cases (or cases, respectively) [26]. Setting the risk threshold should be driven by risk–benefit analysis in the sense that improvement among non-cases is considered more or less important than that among cases [27, 28]. The study by Mihaescu et al. suggests that the risk threshold should be set to a lower level than the disease risk if improvement among non-cases is more important. This approach seems counterintuitive considering that a lower threshold would generally lead to larger proportions of false-positive screens. Higher net improvement for the non-cases is only achievable at the cost of decreasing sensitivity. Other performance measures reflect the sum of the improvements in sensitivity (more true positives) and specificity (less false positives), such as the Net Benefit [28, 29].

The NRI weighs the net improvements in cases and non-cases equally and with that it assigns a weight to the cases with 1/incidence and to non-cases with 1/(1 − incidence) based on the frequency of the outcome observed [27, 30]. The weighted NRI [9, 17, 21] allows to weigh net improvements among cases and non-cases differently for the calculation of the overall NRI. This is a meaningful alternative from a decision-analytic perspective, because improvements are seldom equally important in cases as in non-cases [31].

In conclusion, our study supports that the choice of risk cut-offs and number of risk categories might play a major role in the evaluation of model improvements by reclassification. To avoid subjectivity, evaluation of reclassification should be based on well-established clinical risk categories and cut-offs as has already been recommended by Pencina et al. [8] when introducing the NRI. The category-free (continuous) NRI might be an alternative in the absence of established risk categories.

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 Ellen Kohlsdorf for data management and Olga Kuxhaus for the support in preparing the manuscript.

References

1. Buijsse B, et al. Risk assessment tools for identifying individuals at risk of developing type 2 diabetes. Epidemiol Rev. 2011;33(1):46–62.
2. Wilson PW, et al. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97(18):1837–47.
3. Assmann G, Cullen P, Schulte H. Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Munster (PROCAM) study. Circulation. 2002;105(3):310–5.
4. Conroy RM, et al. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. Eur Heart J. 2003;24(11):987–1003.
5. Hippisley-Cox J, et al. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ. 2007;335(7611):136.
6. Pepe MS, et al. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol. 2004;159(9):882–90.
7. Cook NR, Ridker PM. Advances in measuring the effect of individual predictors of cardiovascular risk: the role of reclassification measures. Ann Intern Med. 2009;150(11):795–802.
8. Pencina MJ, et al. 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. (discussion 207–212).
9. 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.
10. Tzoulaki I, Liberopoulos G, Ioannidis JP. Use of reclassification for assessment of improved prediction: an empirical evaluation. Int J Epidemiol. 2011;40(4):1094–105.
11. Mihaescu R, et al. Improvement of risk prediction by genomic profiling: reclassification measures versus the area under the receiver operating characteristic curve. Am J Epidemiol. 2010; 172(3):353–61.
12. Cook NR, Paynter NP. Performance of reclassification statistics in comparing risk prediction models. Biometric J. 2011;53(2):237–58.

13. 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.
14. Schulze MB, 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.
15. 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(3):837–45.
16. Demler OV, Pencina MJ, D'Agostino Sr. RB. Misuse of DeLong test to compare AUCs for nested models. Stat Med. 2012;31(23):2577–87.
17. Chambless LE, Cummiskey CP, Cui G. Several methods to assess improvement in risk prediction models: extension to survival analysis. Stat Med. 2011;30(1):22–38.
18. Sundstrom J, et al. Useful tests of usefulness of new risk factors: tools for assessing reclassification and discrimination. Scand J Public Health. 2011;39(4):439–41.
19. Steyerberg EW, Van Calster B, Pencina MJ. Performance measures for prediction models and markers: evaluation of predictions and classifications. Rev Esp Cardiol. 2011;64(9):788–94.
20. Mealiffe ME, et al. Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. J Natl Cancer Inst. 2010;102(21):1618–27.
21. Steyerberg EW, et al. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. Eur J Clin Invest. 2011;42(2):216–28.
22. Cook NR, Paynter NP. Comments on 'extensions of net reclassification improvement calculations to measure usefulness of new biomarkers' by M. J. Pencina, R. B. D'Agostino, Sr. and E. W. Steyerberg. Stat Med. 2012;31(1):93–5. (author reply 96–97).
23. 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) final report. Circulation. 2002;106(25):3143–421.
24. Grundy SM, et al. Implications of recent clinical trials for the National Cholesterol Education Program Adult Treatment Panel III guidelines. Circulation. 2004;110(2):227–39.
25. 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. Stat Med. 2012;31(2):101–13.
26. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med. 2006;144(3):201–9.
27. Steyerberg EW, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128–38.
28. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565–74.
29. Peirce CS. The numerical measure of the success of predictions. Science. 1884;4(93):453–4.
30. Pencina MJ. Response to 'Net reclassification improvement and decision theory' by Vickers et al. Stat Med. 2009;28(3):526–8.
31. Greenland P (2008) 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, R. B. D'Agostino Sr., R. B. D'Agostino Jr., R. S. Vasan, Statistics in Medicine. Stat Med. 2008; 27(2): 188–190. doi:10.1002/sim.2929