Practice of Epidemiology

Improvement of Risk Prediction by Genomic Profiling: Reclassification Measures Versus the Area Under the Receiver Operating Characteristic Curve

Raluca Mihaescu, Moniek van Zitteren, Mandy van Hoek, Eric J. G. Sijbrands, André G. Uitterlinden, Jacqueline C. M. Witteman, Albert Hofman, M. G. Myriam Hunink, Cornelia M. van Duijn, and A. Cecile J. W. Janssens *

* Correspondence to Dr. A. Cecile J. W. Janssens, Department of Epidemiology, Erasmus MC, University Medical Center, PO Box 2040, 3000 CA Rotterdam, The Netherlands (e-mail: a.janssens@erasmusmc.nl).

Initially submitted September 18, 2009; accepted for publication April 13, 2010.

Reclassification is observed even when there is no or minimal improvement in the area under the receiver operating characteristic curve (AUC), and it is unclear whether it indicates improved clinical utility. The authors investigated total reclassification, net reclassification improvement, and integrated discrimination improvement for different ΔAUC using empirical and simulated data. Empirical analyses compared prediction of type 2 diabetes risk based on age, sex, and body mass index with prediction updated with 18 established genetic risk factors. Simulated data were used to investigate measures of reclassification against ΔAUCs of 0.005, 0.05, and 0.10. Total reclassification and net reclassification improvement were calculated for all possible cutoff values. The AUC of type 2 diabetes risk prediction improved from 0.63 to 0.66 when 18 polymorphisms were added, whereas total reclassification ranged from 0% to 22.5% depending on the cutoff value chosen. In the simulation study, total reclassification, net reclassification improvement, and integrated discrimination improvement increased with higher ΔAUC. When ΔAUC was low (0.005), net reclassification improvement values were close to zero, integrated discrimination improvement was 0.08% (P > 0.05), but total reclassification ranged from 0 to 6.7%. Reclassification increases with increasing AUC but predominantly varies with the cutoff values chosen. Reclassification observed in the absence of AUC increase is unlikely to improve clinical utility.

Diabetes mellitus, type 2; genetic predisposition to disease; models, statistical; polymorphism, single nucleotide; risk assessment; risk factors

Abbreviations: AUC, area under the receiver operating characteristic curve; CI, confidence interval; IDI, integrated discrimination improvement; NRI, net reclassification improvement.

There is increasing interest in investigating improvement in risk predictions by adding novel genetic variants to traditional risk prediction models for common diseases (1–4). Improvement of prediction of disease risk is commonly assessed by comparing the area under the receiver operating characteristic curve (AUC, or the c statistic) of the prediction models without and with these added variables of interest. AUC assesses the extent to which predicted risks discriminate between individuals who will develop the disease and those who will not (5). Many researchers criticize AUC because this measure lacks an apparent intuitive interpretation and because even statistically significant new biomarkers yield minimal improvement in AUC (1–3). For these reasons, AUC is considered insensitive and not suitable for assessing improvement in prediction, and the use of other measures is encouraged (6, 7).

Currently, much attention is directed toward measures of reclassification (8). Reclassification assesses the extent to which improvements in risk prediction models influence medical decisions. When prediction models are used for making treatment decisions, predicted risks are categorized by using clinically relevant risk cutoff thresholds. Improvement of a prediction model affects medical decisions when individuals are assigned to another category under a new
model compared with the initial model. The percentage of individuals who change risk categories is referred to as the percentage of reclassification.

Several empirical studies have shown that substantial percentages of individuals are reclassified when novel risk indicators are added to prediction models (9, 10). In some studies, reclassification was observed even when there was no or minimal improvement in AUC (6, 11). Yet, more recently, it has been acknowledged that not every reclassification may be correct, which has led to the development of alternative measures such as percentage of correct reclassification, net reclassification improvement (NRI), and integrated discrimination improvement (IDI) (12–14). These measures differ predominantly in their definition of correct reclassification. Reclassification can be considered correct when the observed risk in a reclassified group falls within or is closer to the new risk category (12, 13) but also when individuals who will develop disease change to a higher risk category and those who will not change to a lower risk category (14). Another difference between the measures is that IDI is a global measure of correct reclassification over all possible cutoff values and, in that sense, is comparable to the AUC, whereas the other measures assess correct reclassification for specific cutoff values.

Measures of reclassification are increasingly used in studies investigating the added value of genetic factors beyond traditional risk factors (Table 1) (2, 12, 15, 16). Remarkably, these studies tend to put stronger emphasis on reclassification than on AUC, particularly when reclassification is observed in the absence of improved AUC (2, 15). It is not clear whether reclassification reflects improvement in risk prediction when AUC does not change.

In this study, we investigated the relation between AUC and various measures of reclassification for different ΔAUC and for all possible risk cutoff levels; we used empirical data from the Rotterdam Study and simulated data. In the Rotterdam Study, we compared risk predictions for type 2 diabetes based on age, sex, and body mass index with predictions based on these factors plus 18 genetic polymorphisms. In the simulated data, we additionally investigated measures of reclassification against different improvements in AUC.

MATERIALS AND METHODS

Empirical data

The design of and data collection for the Rotterdam Study have been described elsewhere (17). In short, the Rotterdam Study is a prospective, population-based, cohort study of 7,983 inhabitants of a Rotterdam suburb of the Netherlands designed to investigate determinants of chronic diseases. Participants were aged 55 years or older. Baseline examinations took place from 1990 until 1993, and follow-up examinations were performed in 1993–1994, 1997–1999, and 2002–2004. Between these examinations, continuous surveillance for major disease outcomes was conducted. Details on the genotyped variants and diagnostic criteria for diabetes have been published elsewhere (1, 18, 19). Patients with a recorded diagnosis of type 1 diabetes were excluded. This exclusion is important because the relationship between ΔAUC and reclassification may differ for type 1 and type 2 diabetes. The medical ethics committee of the Erasmus Medical Center approved the study protocol, and all participants gave their written informed consent.

Simulated data

To construct simulated data sets, we adopted a modeling procedure described in detail elsewhere (20). In short, the procedure creates a data set in such a way that the frequencies and odds ratios of the genotypes and the disease risk match specified values. For simplicity, we assumed that each individual polymorphism had only 2 genotypes, one of which was associated with an increased risk of disease and the other being the referent group at baseline risk. Genetic variants were modeled as independent, whereas the joint risk of multiple genotypes followed a multiplicative risk model without any statistical interactions. For a population of 10,000 individuals, we simulated 2 prediction models. The first model was based on 20 polymorphisms, all with an arbitrary odds ratio of 1.2 and a genotype frequency of 20%, which gave an AUC of 0.61. In the second model, we added genetic markers to the first model so the increment in AUC (ΔAUC) was 0.005, 0.05, and 0.10. To do so, we added 1, 25, and 60 markers with the same odds ratios and genotype frequencies as in the first model. The disease risk was arbitrarily set at 10%.

Statistical analyses

In both the empirical and simulated data, predicted risks were obtained by using logistic regression analyses. In the empirical study, we constructed a prediction model for type 2 diabetes (prevalent and incident cases) based on age, sex, and body mass index and one based on age, sex, and body mass index plus 18 genetic polymorphisms, with all polymorphisms entered as categorical variables, enabling effect sizes to differ between heterozygous and homozygous carriers of the risk alleles. The population of the present analyses and the included polymorphisms have been described elsewhere (1). In the simulation study, we constructed risk prediction models based on 20, 21, 45, and 80 polymorphisms.

Cutoff values were varied between 1% and 99%, with increments of 1%. At each possible cutoff value, predicted risks were classified into 2 categories. One included individuals with a risk higher than or equal to the cutoff value; the other included those with a risk lower than the cutoff value. AUC ranges from 0.5 (total lack of discrimination) to 1.0 (perfect discrimination). The percentage of total reclassification was measured as the percentage of individuals, out of the total population, who change risk category. NRI was calculated as the sum of differences in the proportion of individuals moving up minus the proportion moving down for cases and the proportion of individuals moving down minus the proportion moving up for noncases (14). The components of NRI indicate the net benefit of reclassification improvement in cases and noncases. Positive and
Table 1. AUC and Reclassification Measures Reported for Genetic Polymorphisms Added to Prediction Models of Type 2 Diabetes and Cardiovascular Disease Risk

<table>
<thead>
<tr>
<th>Study (Reference No.)</th>
<th>Disease Risk, %</th>
<th>Cutoff Values for Risk Categories, %</th>
<th>Clinical Risk Prediction Model</th>
<th>No. of Added Polymorphisms</th>
<th>AUC for the Initial Model</th>
<th>95% CI</th>
<th>AUC for the Updated Model</th>
<th>95% CI</th>
<th>ΔAUC</th>
<th>95% CI</th>
<th>P Value</th>
<th>Reclassification, %</th>
<th>NRI, %</th>
<th>95% CI</th>
<th>P Value</th>
<th>IDI, %</th>
<th>95% CI</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meigs et al. (16)</td>
<td>10.5</td>
<td>0 to &lt;2; 2 to &lt;8; &gt;8</td>
<td>Sex</td>
<td>18</td>
<td>0.534</td>
<td>0.502, 0.565</td>
<td>0.581</td>
<td>0.546, 0.617</td>
<td>0.047</td>
<td>0.01</td>
<td>2.56c</td>
<td>4.10</td>
<td>0.004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sex and family history</td>
<td>18</td>
<td>0.595</td>
<td>0.560, 0.630</td>
<td>0.615</td>
<td>0.579, 0.652</td>
<td>0.020</td>
<td>0.11</td>
<td>9.54c</td>
<td>2.60</td>
<td>0.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lyssenko et al. (2)</td>
<td>12.8 (Malmö Study)</td>
<td>0 to ≤10; &gt;10 to ≤20; &gt;20</td>
<td>Clinical risk factors</td>
<td>11</td>
<td>0.743</td>
<td>0.753</td>
<td>0.010</td>
<td>0.0001</td>
<td>15.59c</td>
<td>4.5c</td>
<td>2.5 × 10⁻⁵c</td>
<td>NR</td>
<td>3.7 × 10⁻⁵c</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Age, sex, and body mass index</td>
<td>18</td>
<td>0.786</td>
<td>0.801</td>
<td>0.015</td>
<td>0.23</td>
<td>9.44c</td>
<td>8.79c</td>
<td>0.13c</td>
<td>NR</td>
<td>0.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rotterdam Study</td>
<td>20.0</td>
<td>All possible thresholds</td>
<td>Age, sex, and body mass index</td>
<td>18</td>
<td>0.631</td>
<td>0.612, 0.650</td>
<td>0.661</td>
<td>0.643, 0.679</td>
<td>0.030</td>
<td>2 × 10⁻⁶</td>
<td>0 to 22.5a</td>
<td>−0.03 to 9.2a</td>
<td>0.79, 6.8 × 10⁻¹⁴g</td>
<td>2.04</td>
<td>6.5 × 10⁻¹⁰</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paynter et al. (12)</td>
<td>3.2</td>
<td>0 to &lt;5; 5 to &lt;10; 10 to 20; &gt;20</td>
<td>ATP IIIh</td>
<td>1</td>
<td>0.803</td>
<td>0.766, 0.840</td>
<td>0.805</td>
<td>0.768, 0.842</td>
<td>0.002</td>
<td>2.7</td>
<td>2.7</td>
<td>0.02</td>
<td>0.001</td>
<td>0.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Reynolds risk scorei</td>
<td>1</td>
<td>0.807</td>
<td>0.770, 0.844</td>
<td>0.809</td>
<td>0.772, 0.846</td>
<td>0.002</td>
<td>2.6</td>
<td>−0.2</td>
<td>0.59</td>
<td>0.0</td>
<td>0.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kathiresan et al. (15)</td>
<td>5.6</td>
<td>0 to ≤10; &gt;10 to ≤20; &gt;20</td>
<td>ATP IIIh</td>
<td>9</td>
<td>0.800</td>
<td>0.800</td>
<td>0.00</td>
<td>4.39c</td>
<td>6c</td>
<td>0.01c</td>
<td>NR</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: ATP III, Third Report of the National Cholesterol Education Program Expert Panel on Detection, Evaluation, and Treatment of High Cholesterol in Adults; AUC, area under the receiver operating characteristic curve; CI, confidence interval; IDI, integrated discrimination improvement; NR, not reported; NRI, net reclassification improvement.

a In the Rotterdam Study, calculated as the number of prevalent and incident cases divided by the sample size.
b Initial risk prediction model included only clinical factors, whereas the updated risk prediction model included clinical factors and one or more genetic susceptibility variants.
c Calculated from reclassification tables available from the cited papers.
d The model was adjusted for age, sex, family history, body mass index, fasting glucose level, systolic blood pressure, high density lipoprotein cholesterol level, and triglyceride level.
e The initial model included age, sex, family history of diabetes, body mass index, blood pressure, high density lipoprotein cholesterol (available in Botnia only), triglycerides, waist circumference (available in Botnia only), and fasting plasma glucose.
f Cutoff values were used to define 2 risk categories; one included individuals with a risk higher than or equal to the cutoff value, and the other included those with a risk lower than the cutoff value.
g Range of reclassification measures and corresponding P values.
h The ATP III covariates include the natural logarithm of age, systolic blood pressure, total and high density lipoprotein cholesterol, smoking status, antihypertensive medication use, and history of diabetes.
i Reynolds Risk Score covariates include age, smoking status, family history of myocardial infarction, hemoglobin A₁c levels (diabetic patients only), and the natural logarithm of systolic blood pressure and total and high density lipoprotein cholesterol.
negative values represent the net percentage of individuals with improved or worse classification, respectively. Overall, improvement in reclassification is indicated by an NRI significantly greater than 0 (14). IDI was calculated as the difference in mean predicted probabilities between cases and noncases between the 2 models (14). IDI can be seen as an improvement in average sensitivity weighted by average 1 minus specificity (14).

Formulas for the reclassification measures used in this study are listed in the Web Appendix, which is posted on the Journal’s Web site (http://aje.oupjournals.org/). Because both the cutoff levels and most reclassification measures are presented as percentages in this paper, we include zero decimals when percentages refer to the cutoff levels and one decimal when they concern the observed percentages of reclassification.

In the simulation study, reclassification measures and AUCs are presented as averages of 100 simulations. The obtained confidence intervals for AUCs and IDI were extremely small (varying the third digit after the decimal point); therefore, confidence intervals for only total reclassification and NRI are presented in this paper. Differences in AUCs were tested with Analyse-it software, version 2.20 (Analyse-it Software, Ltd., Leeds, United Kingdom (21)), which uses the method of DeLong et al. (22) for receiver operating characteristic curve analyses. Logistic regression analyses in the empirical study were performed by using SPSS software, version 15.0.1 (SPSS Inc., Chicago, Illinois). All other analyses were performed by using the R programming language, version 2.8.0 (23). AUC was obtained as the c statistic by the R function somers2, which is available in the Hmisc library of R software (24).

RESULTS

Prediction of type 2 diabetes

The average risk of type 2 diabetes in the population was 20%. Updating prediction based on age, sex, and body mass index with 18 polymorphisms redistributed predicted risks toward more extreme values (Figure 1A). This redistribution improved the discriminative accuracy, which was also reflected in the AUC values. The AUCs were 0.63 (95% confidence interval [CI]: 0.61, 0.65) for prediction based on age, sex, and body mass index and 0.66 (95% CI: 0.64, 0.68) for prediction based on age, sex, body mass index, and 18 polymorphisms (P < 0.001). The percentages of total reclassification varied with the cutoff values (Figure 1B). The maximum percentage of total reclassification was 22.5% when the cutoff value was 18%. The NRI values ranged from −0.03% (P = 0.79) to 9.2% (P < 0.001) and 8.5% (P < 0.001) when the cutoff values were 4%, 13%, and 31%, respectively (Figure 1C). All reclassification measures were zero for cutoff values lower than 4% and higher than 78%. The IDI was 2.04% (P < 0.001).

Simulation study

Figure 2 shows that risk distributions diverged with increasing AUC. Evidently, when ΔAUC was 0.005, the 2 distributions overlapped, but when ΔAUC increased to 0.05 and 0.10, the risk distribution of the updated models covered more extreme values (Figure 2A). The percentages of total reclassification varied with the cutoff values and were higher with increasing ΔAUC at each cutoff value (Figure 2B). Total reclassification ranged from 0.0 to 6.7% (95% CI: 6.2, 7.2 at a cutoff of 9%) when ΔAUC was 0.005, to 27.2% (95% CI: 26.9, 27.6 at a cutoff of 8%) when ΔAUC was 0.05, and to 35.6% (95% CI: 35.3, 36.0 at a cutoff of 8%) when ΔAUC was 0.10.

NRI also varied with the cutoff value and ΔAUC (Figure 2C). When ΔAUC was 0.005, NRI ranged from −0.7% (95% CI: −0.8, −0.65; P > 0.05) to 1.5% (95% CI: 1.4, 1.7; P > 0.05); when ΔAUC was 0.05, NRI ranged from −0.015% (95% CI: −0.02, −0.01; P > 0.05) to 10.5% (95% CI: 10.2, 10.8; P < 0.001); and when ΔAUC was 0.10, NRI ranged from 0 (95% CI: −0.002, 0.002; P > 0.05) to 19.9% (95% CI: 19.5, 20.2; P < 0.001), depending on the cutoff value chosen.

The individual components of NRI (i.e., net benefit in cases and in noncases) varied with the cutoff value (Figure 3). When ΔAUC was 0.10, the net benefit for cases was negative below and positive above the cutoff value of 10% (range, −9.4% (95% CI: −9.7, −9.2; P < 0.001) to 25.8% (95% CI: 25.2, 26.3; P < 0.001)), whereas the reverse was observed for noncases (range, −7.6% (95% CI: −7.8, −7.5; P < 0.001) to 28.8% (95% CI: 28.3, 29.3; P < 0.001)). IDI was 0.08% (P > 0.05), 2% (P < 0.001), and 5% (P < 0.001) when ΔAUC was 0.005, 0.05, and 0.10, respectively. The distribution of reclassification measures did not change when the odds ratio and genotype frequencies of the added variants were varied (refer to Web Figures 1 and 2, the first 2 of 4 supplementary figures, each referred to as “Web Figure” in the text and posted on the Journal’s Web site [http://aje.oupjournals.org/]). When the AUC of the original model was increased from 0.61 to 0.75, the distribution of reclassification measures changed, but the percentage of reclassification still varied with the cutoff threshold chosen (Web Figures 3 and 4).

DISCUSSION

In this paper, we have demonstrated that the amount of reclassification depends on the increase in AUC achieved by updating a prediction model but even more on the cutoff value chosen. In the empirical study, prediction of type 2 diabetes risk based on age, sex, and body mass index improved by adding 18 polymorphisms, as indicated by the improvement in AUC (0.63 to 0.66). Yet, the percentage of total reclassification ranged from 0% to 22.5% depending on the cutoff threshold, with a maximum at cutoff values around the median of the risk distribution. The simulation study showed that all reclassification measures increased with increasing ΔAUC, at every cutoff value.

Our comparative analyses allowed several inferences to be drawn about the relation between reclassification measures and AUC. First, the cutoff value had a substantial impact on the amount of reclassification. For example, in our simulation study, for an increase in AUC of 0.005, 6.7% of individuals changed their predicted risk categories when
Figure 1. Risk distribution and reclassification measures for prediction of type 2 diabetes risk in the Rotterdam Study, the Netherlands, 1990–1993 to 2002–2004. A) Continuous line: distribution of predicted risks from the model based on age, sex, and body mass index; dashed line: distribution of predicted risks from the model based on age, sex, body mass index, and 18 polymorphisms. B) Percentage of total reclassification per cutoff value for risk stratification. C) Net reclassification improvement (NRI) per cutoff value. Cutoff values were used to define 2 risk categories; one included individuals with a risk higher than or equal to the cutoff value, and the other included those with a risk lower than the cutoff value. NRI was calculated as the sum of differences in the proportion of individuals moving up minus the proportion moving down for cases, and the proportion of individuals moving down minus the proportion moving up for noncases.
Figure 2. Risk distribution and reclassification measures for prediction models in the simulation study. A) Distribution of predicted risks for the model based on 20 polymorphisms (continuous line) and for the updated models when change in the area under the receiver operating characteristic curve ($\Delta$AUC) is 0.005 (dashed line), 0.05 (dotted line), and 0.10 (dashed-and-dotted line). B) Percentage of total reclassification per cutoff value for risk stratification when $\Delta$AUC is 0.005 (dashed line), 0.05 (dotted line), and 0.10 (dashed-and-dotted line). C) Net reclassification improvement (NRI) per cutoff value for risk stratification when $\Delta$AUC is 0.005 (dashed line), 0.05 (dotted line), and 0.10 (dashed-and-dotted line). Each polymorphism has an odds ratio of 1.2 and a frequency of 20%. Disease risk is 10%, and sample size is 10,000. Percentage of total reclassification and NRI are presented as mean values obtained from 100 simulations. Cutoff values are used to define 2 risk categories; one included individuals with a risk higher than or equal to the cutoff value, and the other included those with a risk lower than the cutoff value. NRI was calculated as the sum of differences in the proportion of individuals moving up minus the proportion moving down for cases, and the proportion of individuals moving down minus the proportion moving up for noncases.
the cutoff was 9%, representing the highest amount of reclassification, but fewer than 1.0% of subjects were reclassified when the cutoff was 20%. Similarly, when the increase in AUC of the updated model was 0.10, 35.0% and 10.0% of individuals, respectively, were reclassified. In general, more reclassification was observed for cutoff values around the median of the risk distribution than for cutoff values in the tails of the distribution. Because the amount of reclassification varied with the cutoff value chosen, calculation of reclassification is meaningful only when the cutoff scores defining risk categories with different medical implications, for example, lead to different preventive or therapeutic choices. Only under the condition that cutoff thresholds are clinically meaningful would reclassification become a measure of clinical utility.

Second, when the increase in AUC was minimal (e.g., ΔAUC ≤ 0.005), reclassification was still observed at some cutoff values when the percentage of total reclassification was calculated, but not when NRI or IDI was calculated. When AUC does not change, the initial and the updated prediction model predict equally well or poorly. The percentage of total reclassification indicates that, at the individual level, some cases and noncases may move up and others may move down, yet IDI and NRI show no net benefit of this reclassification at the population level. Thus, when reclassification is observed in the absence of any AUC improvement, reclassification means that an updated model, compared with the initial model, simply makes different errors, not fewer errors.

Third, percentages of reclassification and NRI are calculated for specific cutoff values, whereas IDI is a summary measure over all possible cutoff thresholds. Because IDI assesses improvement in prediction across all possible cutoff values, this measure is interpreted as a weighted AUC, comparable to AUC (14, 25). In our simulation study, changes in AUC were reflected in changes in IDI. When ΔAUC was 0.005, IDI was close to zero, but when ΔAUC increased to 0.10, IDI increased accordingly and became highly statistically significant. Because no specific clinical cutoff thresholds are considered, IDI, similar to AUC, is not a measure of clinical utility but rather of clinical validity.

Fourth, we showed that, across the range of cutoff values, NRI followed a bimodal distribution (Figure 2C), which is explained by the differences in the net benefit for cases and for noncases (Figure 3). When a single threshold is used and this threshold is lower than the disease risk, reclassification is markedly improved for noncases and slightly worsened for cases. At cutoff values higher than the disease risk, reclassification is markedly improved for cases and worsened for noncases. In the above example, where the disease risk was 10%, NRI had 2 peaks corresponding to the maximum net benefit among noncases, when the cutoff was 6%, and the maximum net benefit among cases, when the cutoff was 16% (Figure 3). When it is more important to avoid unnecessary treatment (i.e., because of expensive intervention and serious side effects) of individuals who will not develop the disease, a threshold lower than the disease risk should be considered; when it is more important that all at-risk individuals who might develop disease receive preventive interventions (i.e., because of a significant impact on outcome and safe interventions), a threshold higher than the disease risk may be chosen. Therefore, information on improvement in risk prediction for cases and noncases separately may be more valuable than the global NRI measure.

Finally, reclassification measures did not change when the same ΔAUC was realized by fewer genetic risk factors with higher allele frequencies and stronger odds ratios (Web Figures 1 and 2). However, the distributions of reclassification measures were different when the AUC of the starting model was higher (Web Figures 3 and 4). When the AUC of the starting model was higher, reclassification was observed over a wider range of cutoff values. Therefore, the amount
of reclassification observed for any given ΔAUC and cutoff threshold varied with the magnitude of the AUC because the underlying risk distributions of cases and noncases are more spread out (Web Figures 3A and 4).

Analyzing reclassification for all possible cutoff values suggests that the optimal cutoff value could be based on its impact on medical decisions. However, the choice of the optimal cutoff value will be determined by weighing benefits and harms of false-positive and false-negative decisions. In addition, the number of risk categories should be determined by clinical considerations, namely, the number of available preventive treatment strategies. For our empirical study on type 2 diabetes, there were no guidelines on clinically useful risk categories. Whereas a previous study on reclassification of type 2 diabetes risk arbitrarily considered 3 risk categories, defined by 2 cutoff thresholds (16), we used only 1 risk threshold. It is important to note that percentages of reclassification will be higher when more cutoff values are used (13).

In recent empirical studies, reclassification is often given stronger emphasis than AUC (6, 13), but our findings clearly demonstrate that these measures provide different information. For a prediction model to be a useful tool for clinical practice or public health, the model should have both appreciable clinical validity and clinical utility. Because AUC and IDI are not calculated for specific clinical thresholds, both measures reflect the clinical validity of a test. Percentages of total reclassification and NRI do vary with the choice of cutoff values for given levels of IDI and AUC improvement. AUC and reclassification measures provide complementary information about improvement in risk prediction.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Erasmus MC, University Medical Center, Rotterdam, the Netherlands (Raluca Mihaescu, Moniek van Zitteren, Mandy van Hoek, Jacqueline C. M. Witteman, Albert Hofman, M. G. Myriam Hunink, Cornelia M. van Duijn, A. Cecile J. W. Janssens); Department of Radiology, Erasmus MC, University Medical Center, Rotterdam, the Netherlands (M. G. Myriam Hunink); Department of Health Policy and Management, Harvard School of Public Health, Harvard University, Boston, Massachusetts (M. G. Myriam Hunink); Department of Internal Medicine, Erasmus MC, University Medical Center, Rotterdam, the Netherlands (Mandy van Hoek, Eric J. G. Sijbrands, André G. Uitterlinden); and Netherlands Consortium of Healthy Aging, Rotterdam, the Netherlands (André G. Uitterlinden, Jacqueline C. M. Witteman, Albert Hofman, Cornelia M. van Duijn).

This work was supported by a Vidi grant from the Netherlands Organization for Scientific Research (NWO) and by a grant from the Centre for Medical Systems Biology (CMSB) in the framework of the Netherlands Genomics Initiative (NGI). The genotyping database of the Rotterdam Study was funded through the NWO (grants 175.010.2005.011, 911.03.012) and the Research Institute for Diseases in the Elderly (RIDE). The Rotterdam Study is supported by the Erasmus Medical Center and Erasmus University Rotterdam; NWO; the Netherlands Organization for Health Research and Development (ZonMw); RIDE; the Netherlands Heart Foundation; the Ministry of Education, Culture and Science; the Ministry of Health, Welfare and Sports; the European Commission (DG XII); and the Municipality of Rotterdam.

Conflict of interest: none declared.

REFERENCES