ARTICLE | Genetics inMedicine

Evaluation of polygenic risk models using multiple performance measures: a critical assessment of discordant results

Forike K. Martens, MSc[1], Elisa C. M. Tonk, PhD[1] and A. Cecile J. W. Janssens, PhD[1,2]

Purpose: The area under the receiver operating characteristic curve (AUC) is commonly used for evaluating the improvement of polygenic risk models and increasingly assessed together with the net reclassification improvement (NRI) and integrated discrimination improvement (IDI). We evaluated how researchers described and interpreted AUC, NRI, and IDI when simultaneously assessed.

Methods: We reviewed how researchers described definitions of AUC, NRI, and IDI and how they computed each metric. Next, we reviewed how the increment in AUC, NRI, and IDI were interpreted, and how the overall conclusion about the improvement of the risk model was reached.

Results: AUC, NRI, and IDI were correctly defined in 63, 70, and 0% of the articles. All statistically significant values and almost half of the nonsignificant were interpreted as indicative of improvement, irrespective of the values of the metrics. Also, small, nonsignificant changes in the AUC were interpreted as indication of improvement when NRI and IDI were statistically significant.

Conclusion: Researchers have insufficient knowledge about how to interpret the various metrics for the assessment of the predictive performance of polygenic risk models and rely on the statistical significance for their interpretation. A better understanding is needed to achieve more meaningful interpretation of polygenic prediction studies.

Genetics in Medicine (2019) 21:391–397; https://doi.org/10.1038/s41436-018-0058-9

Keywords: Area under the curve; Integrated discrimination improvement; Net reclassification improvement; Polygenic; Risk prediction

INTRODUCTION

The area under the receiver operating characteristic (ROC) curve (AUC or c-statistic)[1] is the most commonly used measure for the evaluation of risk prediction models. AUC quantifies the ability to discriminate between individuals who will or will not manifest the outcome of interest (referred to as events and nonevents in this article). When a model is updated with new risk factors, such as genetic factors or polygenic risk scores, the improvement in the discriminative ability is assessed by the increment in AUC (ΔAUC) (Box 1).[2–4]

In recent years, alternative measures for the evaluation of prediction models have been proposed, including reclassification measures such as the net reclassification improvement (NRI) and integrated discrimination improvement (IDI).[2,5,6] NRI quantifies the extent to which the addition of risk factors leads to improved classification of risks, and IDI assesses the improvement of the risk difference between events and nonevents (Box 1).[2] NRI and IDI are increasingly used in addition to AUC, but the rationale and value of adding these metrics remain often unclear. NRI and IDI are frequently described as measures of discrimination[7,8] and IDI is often labeled as measure of reclassification.[9,10] When the purpose and meaning of the metrics are unclear, it is challenging to interpret the findings, especially when these are discordant.

Discordant findings are often attributed to shortcomings of the metrics. AUC is argued to be insensitive as it often fails to detect improvements in prediction that result from adding clinically relevant risk factors.[2,5,11–14] Others argue that NRI and IDI are too sensitive for identifying changes in predicted risks, which may lead to false positive conclusions about the improvement of prediction models.[15–17] We earlier showed that findings might also be discordant because the metrics assess different aspects of the improvement in predictive performance: ΔAUC assesses the gain in discriminative ability, NRI assesses changes in risk classification, and IDI assesses changes in the risk differences.[18] For example, adding genetic factors might increase the risk differences without improving discriminative ability when the AUC of the clinical prediction model is already high.[18]

The aim of this study was to evaluate how researchers describe and interpret the simultaneous use of multiple metrics in the assessment of improvement in predictive performance of polygenic risk models. Following the recommendations given by the Statement on the reporting of genetic risk prediction studies (GRIPS),[19] we reviewed how researchers

[1]Department of Clinical Genetics, Section Community Genetics, Amsterdam Public Health Research Institute, VU University Medical Center, Amsterdam, The Netherlands; [2]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, Georgia, USA. Correspondence: A Cecile J. W. Janssens (cecile.janssens@emory.edu)

Submitted 12 January 2018; accepted: 27 April 2018
Published online: 12 June 2018

Box 1 Evaluating the predictive performance of polygenic models using AUC, NRI, and IDI: a tutorial

Genetic factors are added to clinical prediction models to improve the prediction of disease. If these genetic factors improve the model, these improvements are reflected in the distributions of predicted risks. Figure A shows the distributions of predicted risks using a clinical prediction model for participants in a hypothetical study. The participants who did not develop the disease during the duration of the study (referred to as nonevents) tended to have lower predicted risks than those who did develop the disease (events): the distribution of predicted risks for nonevents is skewed toward lower risk as compared with the distribution of predicted risks for events.

When genetic factors are added to the clinical prediction model, we see that the distribution for nonevents "moves" even more toward lower risk, and the distribution for events moves toward higher risk (Figure B). There are several ways how these changes in the distributions of predicted risks can be quantified. The most commonly known is the area under the receiver operating characteristic curve (AUC),[1] but the net reclassification improvement (NRI) and integrated discrimination improvement (IDI) became popular once introduced.[2] We will explain the measures in reverse order.

IDI: increase in risk difference

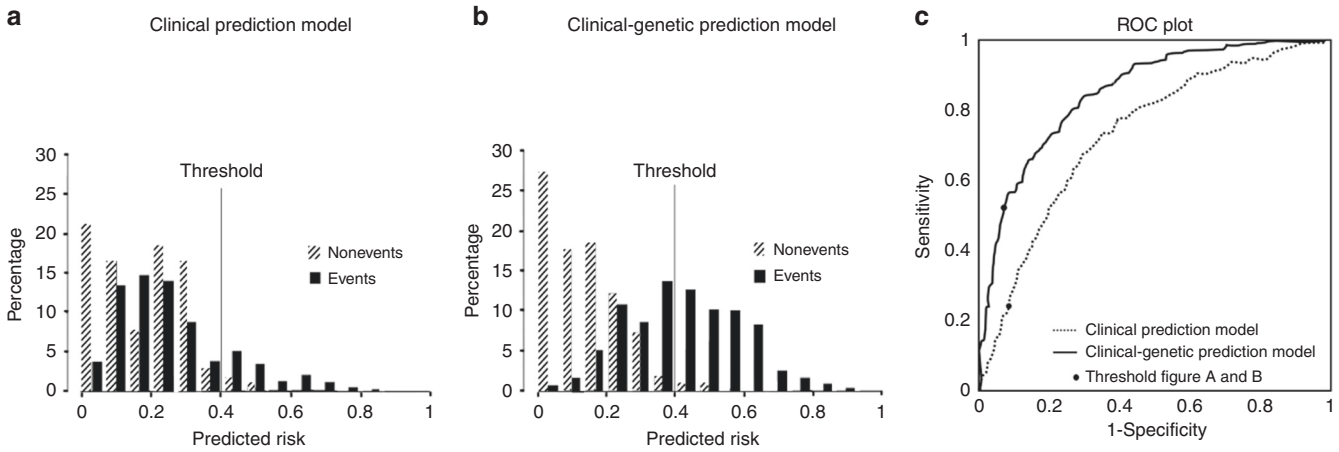
Instead of presenting distributions of predicted risks for events and nonevents, we can calculate the average predicted risks in both groups for each prediction model. When the risk distributions of events and nonevents entirely overlap, the difference between the averages is zero. When the risk distributions "move" further apart—in our example, because genetic factors were added—the difference between the two averages becomes larger. The increase in the risk differences between the clinical and the clinical–genetic prediction model is the IDI.[2]

NRI: reclassification into correct risk category

Prediction models are often used to classify people in risk categories by setting one or more risk thresholds. In our example, we have a single threshold that divides the population into a low- and high-risk group. The proportion of events that have predicted risks above the threshold is the sensitivity and the proportion of nonevents with predicted risks below the threshold is the specificity. The sensitivity and specificity are the proportions of correct classifications. A perfect prediction model would classify all events above the threshold and all nonevents below, and have sensitivity and specificity of 100%. When predicted risks change because genetic factors are added to the clinical model, we want the sensitivity and/or specificity to increase. The increase in sensitivity plus the increase in specificity is the NRI. In general, and if more thresholds are considered, NRI is the sum of the proportion of events that are reclassified to higher risk categories and the proportion of nonevents reclassified to lower categories.[2]

AUC: classification across all risk thresholds

NRI assesses the improvement in discrimination for specific risk thresholds and varies with the number of thresholds and their values.[22] When a clinical prediction model has no known risk thresholds, we can assess the improvement by calculating and comparing sensitivity and specificity across all possible risk thresholds. The lines that connect the sensitivity–specificity of all thresholds of a prediction model is the receiver operating characteristic (ROC) curve and the area underneath is the AUC (Figure C).[1] The figures show that the clinical–genetic prediction model has more favorable combinations of sensitivity and specificity than the clinical model: each sensitivity comes with a higher specificity (or each specificity with a higher sensitivity). The combinations are more favorable, because there is less overlap between the risk distributions of events and nonevents using the clinical–genetic model as compared with the clinical model. This leads to a larger area under the ROC curve and thus a higher AUC. The improvement in discriminative ability between the models is the increment in AUC (ΔAUC).[4]



described what the metrics are assessing, how the metrics were obtained, how their results were interpreted, and how the overall conclusion was reached.

MATERIALS AND METHODS

Literature search

We performed a literature search to find empirical studies that evaluated the improvement in predictive performance of risk models by assessing ΔAUC, NRI, and IDI. Using Thomson Reuters Web of Knowledge (version 5.17) we retrieved all publications that cited the article by Pencina et al. in which the NRI and IDI were introduced (search date 28 December 2016).[2] To limit the number of articles, we focused on studies that investigated the improved predictive performance of adding genetic variants (single-nucleotide polymorphisms, or SNPs) to clinical risk models. For this purpose, we selected publications using the keywords *genetic*, *genomic*, *polygenic*, *polymorphisms*, or *DNA*. We excluded studies on nongermline DNA, such as circulating cell-free DNA or tumor DNA. Full-text articles and Supplementary Materials were obtained for data extraction.

Data extraction

For each study, we recorded sample size, event rate, clinical risk factors in the clinical prediction models as well as the number of SNPs that were added. The event rate is the proportion of individuals with the outcome of interest in the study population, which was the incidence, prevalence, or the size of case population, depending on the design of the study. We extracted AUC values of the baseline and updated models, as well as the values of NRI and IDI along with *P* values and confidence intervals. We recorded whether NRI was used with or without categories: categorical NRI is a metric that is based on the proportions of people that move between risk categories, and continuous NRI is based on the proportions of people that have higher or lower risks after

updating the risk model. When multiple prediction models were investigated in one article, we selected the model that was described in the abstract, the model that had the highest number of risk factors in the clinical prediction model, or the model that had the highest number of SNPs added.

We extracted, verbatim, descriptions of the definitions and calculations of AUC, NRI, and IDI from the methods section of the articles. From the results and discussion sections, we extracted descriptions of the numerical results of the metrics, the interpretation of each measure, and the general conclusions. All descriptions were imported into Microsoft Excel (Microsoft Corporation, Redmond, WA, USA).

Analysis

We evaluated the point estimates and statistical significance of NRI and IDI in relation to ΔAUC. Statistical significance was based on the confidence intervals or the reported *P* values using the threshold of statistical significance mentioned in the articles, which was $P < 0.05$ in all of them.

Using the excerpts of the methods section, we reviewed how the measure and calculation of AUC, NRI, and IDI were described, and evaluated whether these followed common definitions and approaches. For the latter, we required that the definition of AUC should at least have mentioned that it is a measure of discrimination or the concordance between predicted and observed survival, that NRI is a measure of reclassification, and that IDI assesses the improvement in risk differences or discrimination slopes (Box S1). Descriptions of the calculations needed to give insight in the computation. For AUC the description needed to refer to the c-statistic or nonparametric trapezoidal rule. For NRI the description needed to include that it was the sum of the net percentage of correct reclassification in events and nonevents, with reclassification refering to changes between risk categories for categorical NRI and changes in risk for continuous NRI. The description of IDI needed to refer to the difference of the mean increments and mean decrements in estimated probabilities between models or the difference in discrimination slopes of the baseline and updated model (Box S1).

Using the excerpts of the results section, we assessed how the values of AUC, NRI, and IDI were described. We documented whether the results were described by their effect sizes, *P* values or confidence intervals, or both, and whether and how the results were interpreted in terms of model improvement. We documented whether authors reported the presence or absence of improvement, and considered "minimal improvement" when they described the improvement or increase in the estimates as being small or minimal.

Finally, using excerpts from the discussion, we evaluated how the overall improvement of the model was interpreted. In addition to the presence or absence of improvement, we distinguished "minimal improvement" when the reported improvement was considered minimal or marginal, and "inconclusive" when the authors concluded that improvement was demonstrated from some metric(s) but not others. Two researchers independently evaluated the descriptions and disagreements were discussed to reach consensus.

RESULTS

Of the 2509 articles that had cited the article by Pencina et al., 250 articles reported polygenic risk studies of which 32 met the inclusion criteria (Fig. S1). Most excluded articles did not report empirical analyses (such as reviews and commentaries, $n = 94$) or did not report on all three measures ($n = 83$). The majority of the 32 included articles evaluated cardiovascular ($n = 15$) and cancer prediction models ($n = 8$; Table S1).

Definitions of AUC and NRI and IDI were given in 84, 81, and 72% of the articles, of which 63, 70, and 0% were correct (Table 1). IDI was frequently described as a metric of reclassification (30%) and discrimination (22%), and five articles described NRI and IDI together, for example, as measures of "model performance" or "utility." Half of the articles (56%) described how AUC was obtained, of which all mentioned the c-statistic, but only three (9%) explained the calculation of NRI and three others (9%) explained IDI. The three descriptions for the calculation of IDI were correct, but none of the articles described NRI as the *sum* of two *net* percentages.

AUC values of the clinical prediction models ranged from 0.56 to 0.87 (Table S2), and ΔAUC ranged from −0.001 to 0.09 (median 0.01, interquartile range [IQR] 0.002–0.02; Table 2). Most (94%) ΔAUC values were 0.04 or lower. Of the 24 articles that computed the categorical NRI, the values ranged from −0.02 to 0.54 (median 0.044, IQR 0.012–0.142;) and the 7 articles that computed the continuous NRI reported values ranging from 0.07 to 1.24 (median 0.233; IQR 0.137–0.356; Table 2). Of the 24 articles that reported absolute IDI, values ranged from 0.00062 (a 0.062% absolute increase in risk difference between events and nonevents) to 0.128 (median 0.011; IQR 0.002–0.021). NRI and IDI values were, as expected, higher for higher values of ΔAUC (Fig. 1).

ΔAUC was statistically significant in 13 articles, NRI in 21, and IDI in 26 (Table 2). When ΔAUC was higher than 0.01 ($n = 15$ studies), IDI and NRI were both statistically significant in all but 1 of 14 studies (Table 2). Of the 17 studies in which ΔAUC was equal or lower than 0.01, NRI and IDI values were still statistically significant in 7 of 16 of them.

When the value of a metric was statistically significant, the metric was interpreted as indicating improvement of the model in all articles, with several reporting that the improvement was minimal (Table 3). When a metric was not statistically significant, almost half were still described as indicative of model improvement, now with most acknowledging that the improvement was minimal. All ΔAUC values that were not statistically significant and interpreted as no indication of improvement were lower than 0.005, whereas those that were considered to indicate (minimal) improvement were all equal to or higher than 0.005. All statistically

Table 1 Definition and calculation method of AUC, NRI, and IDI as described in included articles

Metric	Definition	% (Articles)	Calculation method	% (Articles)
AUC	*Not reported*	*16 (5)*	*Not described*	*44 (14)*
	Reported	*84 (27)*	*Described*	*56 (18)*
	Discrimination	56 (15)	C-statistic/index	100 (18)
	Probability of concordance between predicted and observed survival	7 (2)		
	Prediction	7 (2)		
	Performance	7 (2)		
	Accuracy, classification, clinical value, incremental value, predictive value, correlation models with outcome	23 (6)		
NRI	*Not reported*	*19 (6)*	*Not described*	*91 (29)*
	Reported	*81 (26)*	*Described*	*9 (3)*
	Reclassification	70 (18)	Comparison of proportions of correct reclassifications to either higher or lower risk	100 (3)
	Classification	7 (2)		
	Discrimination, improvement, model fit, model performance, prediction, utility	23 (6)		
IDI	*Not reported*	*28 (9)*	*Not described*	*91 (29)*
	Reported	*72 (23)*	*Described*	*9 (3)*
	Reclassification	30 (7)	Difference of mean increments and decrements in estimated probabilities between models	67 (2)
	Discrimination	22 (5)		
	Improvement in average sensitivity without sacrificing average specificity	13 (3)		
	Model performance	9 (2)	Differences in discrimination slopes between models	33 (1)
	Classification	9 (2)		
	Model fit, improvement, prediction, utility	17 (4)		

AUC area under the receiver operating characteristic curve, *IDI* integrated discrimination improvement, *NRI* net reclassification improvement
The italic values are used to distinguish them from the other values under "reported" and "described". The total of the italic values is the total of articles included in our study. The total of the non-italic values is the reported/described italic value.

significant ΔAUC values were interpreted as indicating improvement of the model, irrespective of their absolute values.

In 17 of the 27 articles that reported all three values in the results section (Table 2), the authors interpreted that all three metrics showed improvement of the model. Among these were 7 studies in which all three metrics were statistically significant and 7 studies in which NRI and IDI were statistically significant but ΔAUC was not. In 6 of the 27 articles, the authors interpreted that the ΔAUC showed no improvement of the model but that the NRI and IDI did. In all of these, ΔAUC was equal to or lower than 0.003, and NRI was not statistically significant in 2 of them. Only 1 of the 27 articles interpreted that none of the metrics indicated an improvement of the prediction model; in this study, the absolute values of ΔAUC, NRI, and IDI were all lower than 0.001 and not statistically significant.

All but five articles concluded that, overall, the clinical prediction model had improved from the addition of genetic factors (Table 2). Half of them mentioned that the improvement was minimal. All articles in which the individual metrics were evaluated as indicative of improvement, also had a overall positive evaluation, except one in which all three metrics were interpreted as showing minimal improvement leading to an overall conclusion of no improvement. Of the six articles that reported improvement indicated by NRI and IDI but not by ΔAUC, five concluded that the model had improved albeit minimally, and one refrained from making an overall conclusion.

DISCUSSION

AUC, NRI, and IDI are three metrics that are increasingly used together in the assessment of polygenic risk models. Our analysis showed that authors provided minimal information about the purpose and assessment of the three metrics and that they mostly relied on statistical significance when interpreting the results. None of the articles distinguished, in their conclusions, between the different aspects of model performance that the metrics address.

Three observations can be made from this study. First, one-third of the articles did not specify what was measured by IDI and one-fifth did not do so for AUC and NRI. When authors did describe the metrics, only two-thirds were correct about what is measured by AUC and NRI, namely discrimination and reclassification, but were mostly wrong about IDI, which they described as a metric of discrimination, reclassification,

Table 2 Point estimates; interpretations of model improvement based on ΔAUC, NRI, and IDI values; and overall conclusions about improvement of predictive performance

	Point estimates			Model improvement			Overall
First author	ΔAUC	NRI (*P* value or 95% CI)	IDI (P value or 95% CI)	ΔAUC	NRI	IDI	
Park[S1]	−0.001 (0.99)	0.040 (0.32)	0.021 (0.02)	No	No	Yes	No
Eriksson[S2]	0 (0.246)	0.11 (0.005)[a]	0.003 (0.007)	No	[Yes]	[Yes]	[Yes]
Fava[S3]	0 (>0.05)	0.002 (0.39)	0.00449 (0.02)	No	[Yes]	[Yes]	[Yes]
Kathiresan[9]	0 (NR)	NR (0.01)	NR (0.02)	No	Yes	Yes	[No]
Havulinna[S4]	0.0006 (0.16)	−0.0008 (0.92)	0.00062 (0.14)	No	No	No	No
Gränsbo[10]	0.001 (NR)	0.012 (0.043)	0.001 (<0.001)	[Yes]	[Yes]	[Yes]	No
Ripatti[S5]	0.001 (0.19)	0.022 (0.182)	0.004 (0.0006)	No	[Yes]	Yes	[Yes]
Lim[S6]	0.001 (0.1057)	0.019 (0.0495)	0.002 (0.0131)	No	Yes	Yes	[Yes]
Thanassoulis[S7]	0.002 (NR)	−0.01 (−0.052 to 0.033)	0.001 (−0.001 to 0.003)	[Yes]	NR	[Yes]	[Yes]
Fava[S8]	0.003 (>0.05)	0.0659 (0.013)[a]	0.001452 (0.003)	No	Yes	Yes	[Yes]
Brautbar[S9]	0.004 (0.001 to 0.007)	0.008 (0.31)	0.002 (<0.015)	Yes	[Yes]	Yes	[Yes]
Muehlschlegel[S10]	0.005 (NS)	0.195 (0.072)	0.010 (0.053)	NR	Yes	Yes	Yes
Park[S11]	0.005 (0.050)	0.0173 (0.352)	0.0041 (0.007)	[Yes]	No	NR	[Yes]
Juhola[S12]	0.009 (0.015)	0.048 (0.0002)	0.012 (<0.0001)	Yes	Yes	Yes	Yes
Brautbar[S13]	0.009 (0.006 to 0.014)	0.073 (0.019 to 0.12)	0.006 (NR)	Yes	Yes	Yes	Yes
Lyssenko[S14]	0.010 (0.0001)	0.09 (<0.001)	NR (<0.001)	[Yes]	Yes	Yes	[Yes]
Krarup[S15]	0.01 (0.002)	−0.02(NS)	0.001 (NS)	Yes	No	No	No
Butoescu[S16]	0.011 (NR)	0.403 (<0.001)[a]	0.015 (0.035)	[Yes]	Yes	Yes	[Yes]
Yu[S17]	0.011 (>0.050)	0.137 (0.015)	0.0175 (0.041)	[Yes]	Yes	Yes	[Yes]
Lind[S18]	0.012 (0.09)	0.035 (0.047)	0.0072 (0.010)	Yes	Yes	Yes	Yes
Gui[S19]	0.013 (0.17)	0.04850 (<0.001)	0.027 (<0.001)	[Yes]	Yes	Yes	[Yes]
Pitkanen[S20]	0.014 (0.007)	0.163 (0.001)[a]	0.012 (1.8×10^{-5})	Yes	Yes	Yes	Yes
Lobato[S21]	0.015 (NR)	0.194 (0.005)	0.022 (0.01)	[Yes]	Yes	Yes	Yes
Morote[8]	0.02 (0.092)	0.233 (0.003)[a]	NR (<0.001)	[Yes]	Yes	Yes	Yes
Kertai[S22]	0.024 (0.001)	0.308 (0.0003)[a]	0.02 (0.000024)	Yes	Yes	Yes	Yes
Fan[S23]	0.03 (0.0000601)	0.0109 (0.6076)	0.0350 (<0.0001)	[Yes]	No	[Yes]	Yes
Ribeiro[S24]	0.033 (0.0002)	0.095 (<0.0001)	0.021 (<0.0001)	Yes	Yes	Yes	Yes
Borque[S25]	0.034 (0.025)	1.242 (<0.001)[a]	NR (<0.001)	Yes	Yes	Yes	Yes
Ruan[7]	0.04 (<0.001)	0.16 (<0.001)	0.05 (<0.001)	Yes	Yes	Yes	[Yes]
Huesing[S26]	0.041 (NR)	0.158 (NR)	0.0016 (NR)	Yes	Yes	[Yes]	[Yes]
Bolton[S27]	0.069 (0.0001)	0.544 (<0.001)	0.04 (0.02 to 0.06)	Yes	NR	NR	Yes
Chang[S28]	0.088 (0.002)	0.300 (0.005)	0.128 (<0.001)	Yes	Yes	NR	Yes

The point estimates, *P* values, and interpretations of model improvement are as reported in the results section and the overall conclusion as reported in the discussion section of the articles. Square brackets indicate that the authors had expressed hesitancy, e.g., that they considered the improvement of the model to be minimal. References S1–S28 can be found in the Supplementary Data

Δ*AUC* increment in the area under the receiver operating characteristic curve, *CI* confidence interval, *IDI* integrated discrimination improvement, *NR* not reported, *NRI* net reclassification improvement, *NS* not statistically significant

[a]Continuous NRI (see Table S2)

or more generally as a measure of model performance. These findings suggest that researchers may not know what each of the metrics assesses, and that the measures assess different aspects of predictive performance.

Second, only roughly half of the articles reported how AUC ($n = 18$) was obtained and only 9% ($n = 3$) reported how NRI and IDI were calculated. When researchers did provide details, they gave the correct description for the calculation of AUC and IDI, but not of NRI. The three studies that mentioned the calculation of NRI did not describe that NRI is obtained by the sum of the two net proportions. Mentioning the *sum* of the two *net* percentages is important to make clear that NRI is not merely the percentage of reclassified people in a population. These findings confirm that researchers may not know what is measured by NRI and IDI. Whether researchers understand AUC cannot be concluded from this review; evidently, reporting that they obtained the c-statistic may not imply that they understand how the c-statistic is calculated.

And third, inferences about each metric, and hence the overall conclusion about improvement of predictive performance, were largely based on their statistical significance while absolute values of the metrics were small. When the values of the metrics would have been rounded to two decimals, the estimates would be 0.00 for 11 AUC, 2

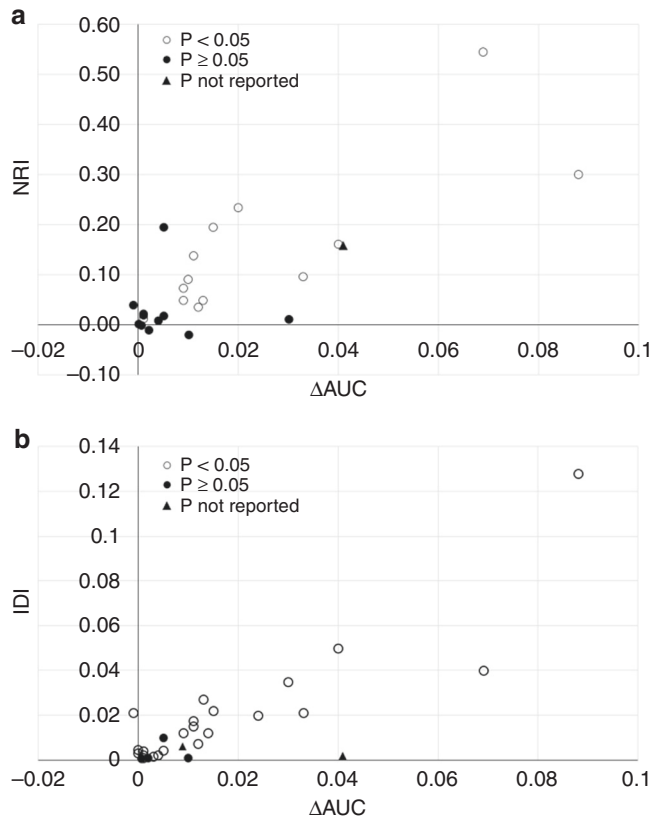


Fig. 1 a Net reclassification improvement (NRI) and b integrated discrimination improvement (IDI) by increments in the area under the receiver operating characteristic curve (Δ AUC). Excluded are studies that **a** used continuous NRI or that did not report the value of the NRI and **b** articles that did not report the value of IDI

Table 3 Inferences about model improvement in the results section of the article in relation to the statistical significance of the metrics

	Model improvement		
	Yes % (articles)	Yes, but minimally % (articles)	No % (articles)
Statistically significant			
ΔAUC	85 (11)	15 (2)	0 (0)
NRI	90 (18)	10 (2)	0 (0)
IDI	83 (19)	17 (4)	0 (0)
Not statistically significant			
ΔAUC	8 (1)	33 (4)	59 (7)
NRI	11 (1)	33 (3)	56 (5)
IDI	25 (1)	25 (1)	50 (2)

Statistical significance was based on reported *P* values and confidence intervals and the criterion of statistical significance in the articles, which was $P < 0.05$ in all of them. Articles that did not report *P* values or confidence intervals for ΔAUC ($n = 6$), NRI ($n = 1$), and IDI ($n = 2$), or did not interpret ΔAUC ($n = 1$), NRI ($n = 2$), and IDI ($n = 3$), are excluded from this table

ΔAUC increment in the area under the receiver operating characteristic curve, *IDI* integrated discrimination improvement, *NRI* net reclassification improvement

NRI, and 12 IDI values. Of these, 3 AUC, 1 NRI, and 9 IDI values were interpreted as showing improvement of the model. Small values of AUC, IDI, and NRI may be statistically significant in large studies, but not clinically relevant. Relying on the statistical significance may lead to false claims about the improvement of prediction. Therefore, the interpretation should focus on the absolute values of the metrics rather than the statistical significance of their estimates.[20,21] What degree of improvement is clinically relevant varies between scenarios and by the answer to the question what is to be gained from the additional information.

The interpretation of polygenic risk studies is straightforward when all measures show the same large and statistically significant improvement in predictive performance. When values are small and inferences are discordant, the question is whether the discordance is due to limitations in the assessment of the metrics or reflecting differential impact on the various aspects of predictive performance. For example, AUC is often criticized for being an insensitive metric to evaluate improvement in predictive performance,[2,5,11–14] but improving discrimination requires a substantial change in the rank order of predicted risks that should not be expected when minor genetic factors are added to the clinical prediction model. In such instances, IDI, which assesses the mean of predicted risks between events and nonevents before and after updating of the clinical prediction model, might still be able to show improvement in risk differentiation. Another example is that changes in risk classification as indicated by NRI may not imply that discrimination is improved as well. NRI has been shown to be too sensitive for identifying minor changes in predicted risks[15–17] and it may be statistically significant, while AUC remains virtually unchanged.[22,23]

All but four studies concluded that the addition of genes to clinical risk models improved the predictive performance of clinical risk models. In most studies, the values of ΔAUC, NRI, and IDI were small and none of them were externally validated. The latter is relevant for the few studies in which the improvement in predictive performance would be of interest if it were replicated in independent data. Judging if clinical risk models improve by the addition of genes is challenging when researchers have limited understanding of the metrics used for evaluation of the models. Our study suggests that this limited understanding leads to false positive conclusions about the value of adding genes to clinical risk models.

Interpretation of polygenic risk studies is straightforward when there is no or substantial improvement in predictive performance, but it is challenging in between. Discordant results from multiple metrics may indicate that there is no improvement but that some metrics are sensitive enough to detect very small effects. Yet, it may also mean that there is improvement in prediction but not on all aspects of predictive performance. A better understanding is needed to achieve more meaningful interpretations of polygenic prediction studies. Overinterpretation of small improvements in

predictive ability will unlikely improve the management of people at risk in public health practice.

ELECTRONIC SUPPLEMENTARY MATERIAL

The online version of this article (https://doi.org/10.1038/s41436-018-0058-9) contains supplementary material, which is available to authorized users.

ACKNOWLEDGEMENTS

This work was supported by a consolidator grant from the European Research Council (GENOMICMEDICINE). Martens and Janssens designed the study. Martens performed all analyses under supervision of Tonk and Janssens. Martens and Janssens drafted the manuscript. All authors contributed to the interpretation of the data and the revisions of the manuscript. All authors approved the final version.

DISCLOSURE

The authors declare no conflicts of interest.

REFERENCES

1. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29–36.
2. Pencina MJ, D'Agostino RB Sr., D'Agostino RB Jr., et al. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27:157–72. discussion 207-112.
3. Steyerberg EW, Pencina MJ, Lingsma HF, et al. Assessing the incremental value of diagnostic and prognostic markers: a review and illustration. *Eur J Clin Invest*. 2012;42:216–28.
4. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983;148:839–43.
5. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. 2007;115:928–35.
6. Pencina MJ, D'Agostino RB Sr., Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med*. 2011;30:11–21.
7. Ruan HL, Qin HD, Shugart YY, et al. Developing genetic epidemiological models to predict risk for nasopharyngeal carcinoma in high-risk population of China. *PLoS ONE*. 2013;8:e56128.
8. Morote J, del Amo J, Borque A, et al. Improved prediction of bochemical recurrence after radical prostatectomy by genetic polymorphisms. *J Urol*. 2010;184:506–11.
9. Kathiresan S, Melander O, Anevski D, et al. Polymorphisms associated with cholesterol and risk of cardiovascular events. *N Engl J Med*. 2008;358:1240–9.
10. Gränsbo K, Almgren P, Sjögren M, et al. Chromosome 9p21 genetic variation explains 13% of cardiovascular disease incidence but does not improve risk prediction. *J Intern Med*. 2013;274:233–40.
11. 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;31:101–13.
12. Pepe MS, Janes HE. Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. *J Natl Cancer Inst*. 2008;100:978–9.
13. Pepe MS. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*. 2004;159:882–90.
14. Ware JH. The limitations of risk factors as prognostic tools. *N Engl J Med*. 2006;355:2615–7.
15. Pepe MS, Janes H, Li CI. Net risk reclassification p values: valid or misleading? *J Natl Cancer Inst*. 2014;106:dju041.
16. Gerds TA, Hilden J. Calibration of models is not sufficient to justify NRI. *Stat Med*. 2014;33:3419–20.
17. Hilden J, Gerds TA. A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Stat Med*. 2014;33:3405–14.
18. Martens FK, Tonk EC, Kers JG, et al. Small improvement in the area under the receiver operating characteristic curve indicated small changes in predicted risks. *J Clin Epidemiol*. 2016;79:159–64.
19. Janssens AC, Ioannidis JP, Bedrosian S, et al. Strengthening the reporting of genetic rIsk prediction studies (GRIPS): explanation and elaboration. J *Clin Epidemiol*. 2011;64:e1–e22.
20. Pepe MS, Kerr KF, Longton G, et al. Testing for improvement in prediction model performance. *Stat Med*. 2013;32:1467–82.
21. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. *BMC Med Res Methodol*. 2011;11:13.
22. Mihaescu R, van Zitteren M, van Hoek M, et al. Improvement of risk prediction by genomic profiling: reclassification measures versus the area under the receiver operating characteristic curve. *Am J Epidemiol*. 2010;172:353–61.
23. Janssens AC, Khoury MJ. Assessment of improved prediction beyond traditional risk factors: when does a difference make a difference? *Circ Cardiovasc Genet*. 2010;3:3–5.