Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world-wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Qi Zhang Date

Impact of Predictive Ability on Identifying Higher-Risk Population for Common Diseases in Polygenic Risk Prediction

By

Qi Zhang
Master of Science in Public Health

Department of Epidemiology

_____

A. Cecile J.W. Janssens, Ph.D.

Committee Chair

Impact of Predictive Ability on Identifying Higher-Risk Population for Common Diseases in Polygenic Risk Prediction

By

Qi Zhang

M.Ed.
Tsinghua University
2017

Faculty Thesis Advisor: A. Cecile J.W. Janssens, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Epidemiology
2019

Abstract

Impact of Predictive Ability on Identifying Higher-Risk Population for Common Diseases in Polygenic Risk Prediction

By Qi Zhang

Objective: Polygenic risk scores can identify individuals with increased risk for common diseases. However, the magnitude of increased risk is unclear for different predictive ability, which is assessed by the area under the receiver operating characteristic curve (AUC). Sensitivity, specificity, PPV, NPV and reclassification measures can reflect predictive performance from various aspects. We aimed to investigate how the magnitude of increased risk for higher-risk group and the above parameters for predictive performance will vary by increasing the AUC.

Methods: We simulated hypothetical risk data with genetic variants and disease status for 100,000 individuals and constructed risk prediction models with the AUC ranging from 0.60 to 0.80. Predicted risks were calculated from Bayes' theorem and logistic regression. We first replicated the findings in the study of Khera et al., then examined odds ratio (OR) for higher-risk group, sensitivity, specificity, PPV and NPV when AUC improved against different cut-offs that can define higher-risk group. We also explored the relationship between reclassification measures (IDI, percentage of total reclassification, NRI and reclassification improvement of cases and non-cases) and increment of AUC (ΔAUC).

Results: OR of the higher-risk population (versus the remainder) increased with improving AUC at an increasing rate for a fixed risk threshold; sensitivity and PPV increased with increasing AUC, but specificity and NPV remained almost constant when AUC improved. IDI, percentage of total reclassification and NRI increased with increasing ΔAUC; the reclassification improvement was higher for cases than for non-cases at the same ΔAUC.

Conclusions: We can identify the higher-risk population with increased OR for common diseases across risk thresholds compared with the remainder when the predictive ability of genetic risk model improves. The sensitivity and PPV increase with improving AUC, and this influence varies across different risk thresholds. Reclassification measures favorably increase when the ΔAUC improves, which is achieved mainly by improving the reclassification of individuals with events.

***Keywords*:** Area under curve; Polygenic risk; Reclassification; Prediction

Impact of Predictive Ability on Identifying Higher-Risk Population for Common Diseases in Polygenic Risk Prediction

By

Qi Zhang

M.Ed.
Tsinghua University
2017

Faculty Thesis Advisor: A. Cecile J.W. Janssens, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Epidemiology
2019

Acknowledgements

I would like to thank Dr. Cecile Janssens, my advisor, for all of her guidance, support and patience. Dr. Janssens has inspired me a lot in scientific analytics and writing skills, and her creative thoughts ignited my passion in prediction research. This thesis would not have been possible without her inspiration and encouragement. I would insist on the critical and creative thinking she taught me as I move forward with my career.

I would like to thank Forike Martens for her guidance and help in the statistical analysis and R programming. I appreciate she so generously donated her time and effort in the discussion and meetings on this project. I am very grateful to Jena Black, my ADAP, who provided me with great suggestion and opportunities regard to thesis and career while I met with difficulties. I am also grateful to Katharine Lindquist, who encouraged me and offered me suggestion about writing.

I would like to thank the department of Epidemiology at Emory for their support in the flexible and personalized learning atmosphere in such an energetic community.

Finally, thanks to my parents and my sister, who always support me along the way whenever I feel discouraged. They are the source of my spiritual strength and encouraged me to just keep writing. All of these help motivate me to accomplish this thesis and evolve into a better public health professional.

Table of Contents

Introduction **1**

Method **4**

Data simulation *4*

Statistical analyses *6*

Results **9**

Validation of the Odds ratio for high-risk group in the study of Khera et al. *9*

Investigate how OR of the higher-risk group varied with the predictive ability *12*

Investigate how predictive performance varied with the predictive ability *14*

Investigate how reclassification measures varied with the predictive ability *16*

Correct and incorrect moves for cases and non-cases *20*

Discussion **21**

Strengths and Weaknesses **26**

Conclusions **27**

References **28**

Appendix 1: Definitions and Formulas for Reclassification Measures **30**

Appendix 2: Comparison of the Validation Results and the Study of Khera et al. **32**

Appendix 3: Risk Distributions of Cases and Non-cases for Varied AUCs **33**

Introduction

Identifying individuals with higher genetic risks can predict multifactorial diseases and initiate screening or presymptomatic diagnosis. This could play a key role in prevention and treatment of common complex disorders and even facilitate the development of personalized medicine (1). Polygenic risk scores (PRS) are calculated from genetic risk models using multiple single-nucleotide polymorphisms (SNPs) that are proven to be associated with specific disease among genome-wide association studies (GWAS) (2-4). It is expected that higher-risk individuals for common diseases could be identified using PRS. However, currently only modest predictive performance was indicated for most genetic risk models in empirical studies (5, 6). The predictive ability of genetic risk models can be improved by adding more SNPs to polygenic profiling for common diseases (7). It has long time been debated about whether it is possible that polygenic scores, based on multiple genes, can be applied to the identification of individuals at meaningfully high risk in clinical practice. Some maintain that the application of polygenic risk prediction is limited in the stratification of multifactorial disease risk for individuals because each SNP conveys only a weak effect and substantial environmental factors also affect the disease risk, although many genetic variants are included. Whereas, others believe that most common diseases involve a large number of common SNPs of low risk rather than rare monogenic mutation, and the accumulated polygenic effects can predict the disease risk based on the susceptibility from polygenic inheritance (8-10).

Using the genotype data from UK Biobank cohort, Khera et al. (11) reported that individuals with more than three-fold increased disease risk were identified using genome-wide polygenic scores for five common diseases (coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer) at top 8.0, 6.1, 3.5, 3.2, and 1.5% of the population in the risk distribution respectively, which seems comparable to the increased risk from rare monogenic mutations. The findings of Khera et al. (11) triggers discussion on the possibility of identifying individuals at higher risk of developing complex diseases on the basis of predictive ability of

polygenetic risk models. Whether using multiple genetic variants as predictors can identify individuals at clinically meaningfully higher risk remains to be determined and requires more evidence.

The area under the receiver operating characteristic (ROC) curve (AUC, or the c-statistic) is usually used to assess the predictive ability of risk models. AUC reflects discriminative accuracy between people who will develop diseases and those who will not (12). Apart from AUC, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) are a group of indicators evaluating the discriminative ability of risk models. The improvement of prediction model is commonly evaluated using the increase in AUC (ΔAUC), which could be resulting from inputting more meaningful risk factors. ΔAUC was blamed for being insensitive to detecting improvement in risk prediction, achieved by adding risk factors(13, 14). Recently, Marten et al. (15) illustrated that there was "insensitivity" of ΔAUC only when the baseline model was at a high level of AUC. They explained that AUC, as a rank-order test, evaluates the discriminative ability of risk models by comparing the predicted average ranks for cases and non-cases, and the ranking might not or only slightly be impacted if cases have higher predicted risk or non-cases have lower predicted risks since the discrimination of average ranks is already excellent at higher baseline AUC.

Reclassification measures are gaining more attention in the studies of polygenic risk prediction and reflect clinical utility of risk models (7). When new risk factors are added to the initial prediction model, predicted risks are updated for each individual. Individuals are reassigned to a risk category based on the comparison of predicted risks and risk cutoff thresholds, and it affects clinical decisions when the risk category of an individual is changed compared with the initial risk model. The proportion of individuals that are assigned to a different risk category is referred to as the percentage of reclassification (7). Not all individuals are reclassified correctly, and thus alternative measures were developed, such as percentages of correct reclassification, net reclassification

improvement (NRI), and integrated discrimination improvement (IDI) (16). Reclassification measures after adding more new SNPs to risk model reveals the change of risk categories, which helps explain what happened behind the increased magnitude of risk for higher risk group compared with the reference group at the population level. From the perspective of clinical utility, reclassification measures quantify the effects of the updated prediction model on clinical decisions. The reclassification measures against the improvement in predictive ability of prediction models remains to be investigated.

Identifying higher-risk individuals also depends on risk cutoff thresholds. Higher risk thresholds would identify individuals with high risk of developing diseases but might miss a proportion of individuals with future events; lower risk thresholds would capture most of the cases but might result in many false negatives. Risk thresholds have effects on sensitivity, specificity, PPV and NPV, and also influence the feasibility of preventive strategy focused on higher-risk individuals.

This study is aimed at investigating the impact of improved predictive ability of the polygenic risk model on the identified magnitude of increased risk for the higher-risk group and reclassification of risk categories. We first used simulated data to validate the findings on the identification of population at three-, four- and five-fold increased risk in the empirical study of Khera et al. (11, 17). Next, we investigated the odds ratio of higher-risk group, and predictive performance (i.e., sensitivity, specificity, PPV and NPV) with increasing predictive ability against different risk cut-off thresholds. Finally, we explored the reclassification measures across varied risk thresholds and increases of predictive ability.

Method

Analyses were conducted using simulated data, which allows for varied modeling parameters that determine the predictive ability. We created datasets consisting of risk genetic variants and disease status and constructed prediction models based on the risk genetic variants. In this study, we varied the frequency of added SNPs to change the predictive ability of genetic risk models. We used AUC to quantify the predictive ability of prediction models and ΔAUC to assess the improvement of predictive ability. To answer the research question about the identification of increased risk for the higher-risk population at different predictive ability, we simulated a hypothetical population for each value of AUC ranging from 0.6 to 0.8 for the prediction model; for the research question about reclassification for different improvement in predictive ability, we created a hypothetical population based on the frequency of SNPs for maximum AUC (0.8). For the same simulated population, baseline models were created with ΔAUC of 0.05, 0.10 and 0.15 from maximum AUC. To this end, we changed the number of risk alleles that the prediction model was based on until the specific ΔAUC (± 0.0025) was reached. The model for which the AUC is 0.65 was used as the baseline model for reclassification analyses.

Data simulation

Hypothetical risk data were derived by simulating genotypes and disease status based on prespecified disease prevalence, odds ratios and frequencies of the risk allele and values of the AUC.

Simulation method for a hypothetical population. The modeling procedure used to construct simulated dataset was described in detail elsewhere (18). The simulated risk datasets were modeled to match the prespecified values of disease prevalence, odds ratios of genetic variants and risk allele frequencies. We constructed hypothetical populations by modeling individual genotypes, predicting individual disease risks and assigning disease status. Specifically, frequencies of three genotypes (homozygous risk genotype, heterozygous risk genotype and homozygous wildtype) for each SNP

were calculated based on allele frequencies, assuming Hardy-Weinberg equilibrium, the same risk allele frequencies for all SNPs and no linkage disequlibrium. Genotypes were randomly assigned to 100,000 individuals. Next, individual disease risks were predicted using Bayes' theorem based on disease prevalence and likelihood ratios of genotype profiles, assuming no interaction between SNPs and the same odds ratio for all genetic variants. To assign disease status, we compared the predicted individual disease risks (i.e. posterior risk based on Bayes theorem) with the randomly drawn values from a uniform distribution between 0 and 1. The individual was assigned to the group of patients if the predicted disease risk was higher than the random value and otherwise to the group of non-patients. AUC was calculated from predicted disease risks and disease status. To obtain a risk dataset from a specific AUC, we iterated the procedure of adding SNPs to genetic profiles until the AUC of the genetic risk model reached the prespecified value (19). Then we used the method above to predict individual disease risks and assign disease status.

Calculation of predicted risk. The predicted risk of the same individuals for analyses of reclassification measures were obtained by logistic regression for baseline and updated prediction models, since the posterior risks based on maximum AUC were fixed for the same population. All other analyses used the posterior probability as the predicted risk for individuals, which were calculated on the basis of Bayes' theorem (18). According to Bayes' theorem, the posterior odds of disease were derived through multiplying the prior odds by the likelihood ratio of individuals' genetic profiles, and the posterior odds can be converted to posterior probability ($probability = \frac{odds}{1+odds}$), which is the predicted risk associated with genetic risk factors.

Calculation of AUC. AUC was calculated as the c-statistic by the ROCR package in R. ΔAUC was defined as the difference between the AUC of the baseline and updated risk models.

Model parameters. The sample size for each hypothetical population in different scenarios was

100,000 (except for Atrial fibrillation, the sample size of which was 200,000). For each combination of model parameters, we used an allele OR of 1.1, an allele frequency of 0.15, considering that common complex diseases are affected by many SNPs with weak effects. Average population risk (disease prevalence) were extracted from the study of Khera et al. (20), which was 3.4% for coronary artery disease (CAD) and 1.7% for Atrial fibrillation. Given the range of AUC between 0.6-0.8, the genetic profiles were defined by no more than 700 SNPs.

Statistical analyses

First, we used simulated data to validate the findings on relative risk of higher-risk population in the study of Khera et al. Khera et al. (11, 17) stated that top 8.0, 2.3 and 0.5% of polygenic scores distribution could confer three-, four- and fivefold increased risk with the remainder as the reference population for CAD, and top 6.1, 1.5 and 0.7% of the distribution were identified at three-, four- and fivefold increased risk compared to the remainder for Atrial fibrillation. We tried possible values of AUC to achieve the specific ORs of the higher-risk group compared with the remainder of the population at the fixed risk thresholds for CAD and Atrial fibrillation, and corresponding OR, sensitivity, specificity, PPV and NPV are shown for each possible AUC. The sample size of simulated data for atrial fibrillation was increased to 200,000 so that we could obtain stable relative risks for the higher-risk group based on small average population risk (1.7%).

Next, to investigate how the relative risk (odds ratios) of the higher-risk group for multifactorial diseases will vary with the predictive ability of the risk model, we calculated the OR of higher-risk group compared with the lower-risk group at different cut-off thresholds across AUC values from 0.6 to 0.8. The top percentages of predicted risk distribution were used as risk cut-off thresholds. The main thresholds are set at 5%, 10% and 20% selected from the study of Khera et al (11). We also compared the difference in using top percentages and ORs as risk thresholds to define risk category. Predicted risk distributions of individuals that will and will not develop the disease across risk categories were derived to illustrate the predictive ability indicated by AUC. To evaluate the

impacts of predictive ability on the predictive performance in terms of risk stratification, we analyzed the relationship between the indicators of predictive performance (sensitivity, specificity, PPV and NPV) and varied AUC of the risk model at risk cut-off thresholds of 5, 10, 20%. Sensitivity is the proportion of individuals that are classified into the higher-risk group among cases, and specificity is the proportion of individuals that are classified into the lower-risk group among non-cases. PPV is the probability that individuals classified in the higher-risk group will develop the events, and NPV is the probability that individuals classified in the lower-risk group will remain unaffected.

Finally, to understand the reclassification of risk category when predictive ability of risk model improves, we calculated the percentage of total reclassification, NRI and IDI when AUC improved from baseline 0.65 to 0.70, 0.75 and 0.80 against varied cut-off thresholds, with the ΔAUC of 0.05, 0.10 and 0.15. To simulate a ΔAUC using the hypothetical population from the maximum AUC (0.80), the frequency of SNPs was changed by decrements of 10, and by 1 when ΔAUC approached the higher end of the range. Correct reclassifications are those moving from the lower-risk group to the higher-risk group for individuals with events and moving from the higher-risk group to the lower-risk group for individuals without events. Percentage of total reclassification was calculated as the proportion of individuals that were reclassified to a different risk category, not only including correct moves. NRI was calculated as the sum of net correct moves after reclassification, i.e., the difference of the percentage of cases that were reclassified from lower-risk group to higher-risk group and the percentage of cases that moved in the opposite direction (which is defined as the reclassification improvement of cases), plus the difference of the percentages of non-cases that moved from higher-risk group to lower-risk group and the percentages of non-cases that were reclassified in the opposite direction (defined as the reclassification improvement of non-cases) (16, 21, 22). IDI was obtained by calculating the improvement of risk difference of average predicted risks for cases and non-cases between baseline model and updated model (22) (see Appendix 1 for formulas). All figures of reclassification measures were presented with the cut-off range of 0-60th

percentile, as we did not have enough power to detect the trends after the 60th percentile, resulting in bizarre curves.

Confidence intervals were not included in our study since we constructed huge simulated data and the confidence intervals were very small. All metrics (AUC, OR, Sensitivity, Specificity, PPV, NPV and reclassification measures) were presented as the average of 10 simulations, repeated for each hypothetical population and derived relevant prediction models. All analyses were performed using R software, version 3.5.2.

Results

Validation of the Odds ratio for high-risk group in the study of Khera et al. (11, 17)

Table 1 shows the OR of higher-risk group ≥ 3, 4 and 5 in the top 0.5, 2.3 and 8% of risk distribution compared with the remainder of population, respectively, for CAD. However, the AUC of the prediction model that can identify such a level of OR at the specified risk thresholds is 0.67, much lower than the AUC (0.81) described in the article of Khera et al.(11), which was determined by a logistic regression model adjusted for age, sex and millions of genetic variants. Similarly, Table 2 illustrates that a prediction model with AUC of 0.67-0.68 can identify the population in the top 0.7, 1.5 and 6.1% of the risk distribution with OR ≥ 3, 4 and 5 respectively for Atrial fibrillation, versus the AUC of 0.77 in the study of Khera et al (11). In both tables (Table 1 and Table 2), the specificity and NPV were very high (>0.9); sensitivity and PPV were relatively low.

Table 1. Odds Ratios and predictive performance for the high-risk population of CAD for prediction models with different AUCs (sample size= 100,000)[c]

AUC	Top% of distribution [a]	OR [b]	Sensitivity	Specificity	PPV	NPV
0.65	0.5	4.50	0.020	0.996	0.134	0.967
	2.3	3.48	0.071	0.979	0.104	0.968
	8	2.85	0.190	0.924	0.08	0.97
0.66	0.5	5.37	0.025	0.995	0.083	0.983
	2.3	3.85	0.080	0.978	0.059	0.984
	8	3.01	0.203	0.922	0.043	0.985
0.67	0.5	5.29	0.024	0.995	0.082	0.983
	2.3	4.00	0.082	0.978	0.061	0.984
	8	3.22	0.214	0.922	0.045	0.985

[a] The predicted risk at the top percentage of the risk distribution was used as the threshold to define the higher-risk population: the individuals with predicted risks higher than or equal to the threshold were classified as the higher-risk group, and those with predicted risks lower than the threshold were classified as the lower-risk group.

[b] OR was the odds ratio of the disease for the higher-risk group compared with the lower-risk group, i.e., the remainder of the distribution.

[c] Each value in the table came from the mean of 10 simulations.

Table 2. Odds Ratios and predictive performance for the high-risk population of Atrial fibrillation for prediction models with different AUCs (sample size= 200,000) *

AUC	Top% of distribution	OR	Sensitivity	Specificity	PPV	NPV
	0.7	4.13	0.027	0.993	0.065	0.983
0.65	1.5	3.75	0.052	0.986	0.058	0.984
	6.1	2.98	0.158	0.941	0.044	0.985
	0.7	4.53	0.029	0.993	0.071	0.983
0.66	1.5	4.00	0.055	0.986	0.062	0.984
	6.1	3.06	0.162	0.941	0.045	0.985
	0.7	4.67	0.030	0.993	0.073	0.983
0.67	1.5	4.16	0.057	0.986	0.064	0.984
	6.1	3.33	0.173	0.941	0.048	0.985
	0.7	5.56	0.035	0.993	0.085	0.984
0.68	1.5	4.87	0.065	0.986	0.074	0.984
	6.1	3.58	0.183	0.941	0.051	0.985

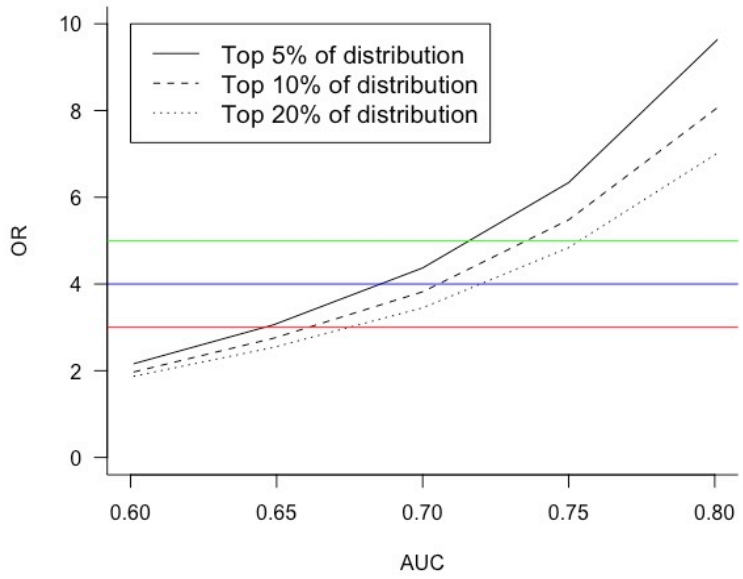
* The threshold of risk stratification and OR were defined in the same way as the Table 1, and each value in the table came from the mean of 10 simulations.

Investigate how OR of the higher-risk group varied with the predictive ability

Figure 1.a demonstrates the OR of population at higher-risk varied with predictive ability and risk cut-off thresholds. For each fixed cut-off threshold, the OR grew with an increasing rate when AUC was improved. When the risk threshold was higher, the same risk model could identify individuals in the top percentage of distribution with a greater OR, and the difference in OR between different risk thresholds ascended with the increment of AUC. For example, the difference of OR were minimal for the higher-risk group when top 5 and 20% were set as the cut-off thresholds using the model with AUC of 0.60; whereas, the OR was estimated around 10 and 8, respectively, for top 5 and 20% of the distribution, when the AUC was increased to 0.80. To detect an OR of 3-5 compared with the remainder of the risk distribution, the required range of AUC for the top 5% of the distribution was roughly 0.64-0.72, for the top 10% of the distribution was around 0.66-0.74, and for the top 20% of the distribution was approximately 0.68-0.75. The predicted risk distribution of individuals who will develop the disease and those who will not overlapped a lot when the AUC was lower, and two risk distributions diverged when AUC improved (refer to Appendix 3).

Figure 1.b compared the options of using OR and top percentages of risk distribution to define the higher-risk group in the reclassification analyses. It shows the values of OR were symmetrically distributed across the percentiles of risk distribution and there was a limited range for the OR with a minimum at the thresholds around top 40%-60% of distribution.

a.



b.

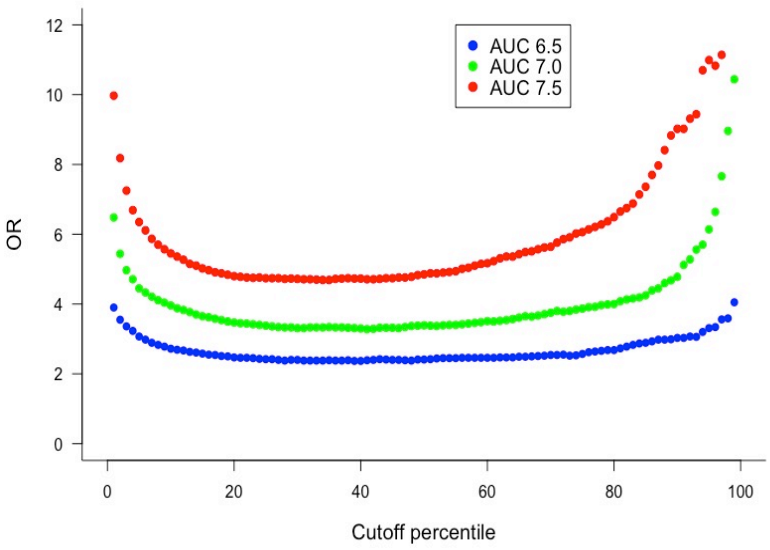


Figure 1. Odds ratios of the individuals at higher risk for different risk thresholds across varied AUCs. a). OR of the higher-risk group compared with the remainder of the distribution across AUCs for thresholds at top 5, 10, 20% of distribution. b). OR versus cut-off percentile (i.e., top percentage of the risk distribution) as the threshold for defining the higher-risk group when the AUC was 6.5, 7.0 and 7.5.

Investigate how predictive performance varied with the predictive ability

Figure 2 presents the change of sensitivity, specificity, PPV and NPV with the improvement of AUC against different risk thresholds. The specificity almost remained unchanged when AUC of the prediction model improved for a fixed risk threshold, and the change of sensitivity was positive with the improvement of AUC. When the threshold for risk stratification decreased from top 5% to top 10% and 20% of the distribution, the specificity of prediction declined from 0.95 to 0.90 and 0.80, but the sensitivity increased when the AUC was fixed, ranging from 0.099 to 0.294 for the risk threshold of top 5% of the distribution, from 0.177 to 0.438 for the threshold of top 10%, and from 0.315 to 0.615 for the threshold of top 20%. The increasing rate of sensitivity appeared to limitedly rise when the risk threshold decreased from top 5% to top 20%. NPV also almost remained constant when AUC changed and even when risk threshold varied. In contrast, PPV increased when AUC was improved, and the increasing rate ascended when the risk threshold was improved.

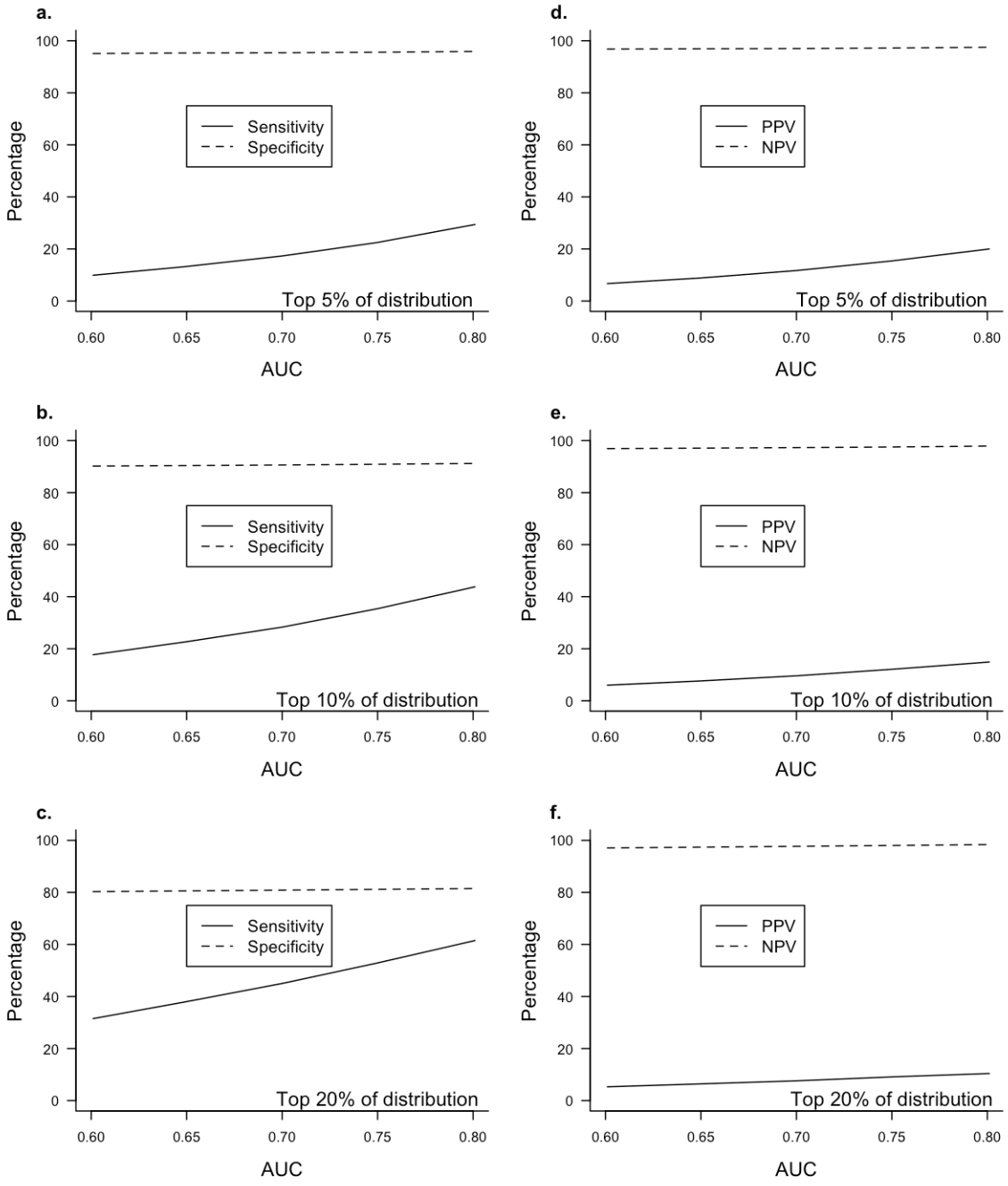


Figure 2. Change of Sensitivity, Specificity, PPV and NPV when AUC improved with different thresholds of the high-risk population. Top % of distribution was used as the threshold to define the two risk categories: the individuals with predicted risks higher than or equal to the threshold were classified as the higher-risk group, and those with predicted risks lower than the threshold were classified as the lower-risk group. Disease prevalence = 3.4%, allele OR = 1.1, allele frequency = 15% and sample size = 100,000. Each value was the mean of 10 simulations, and the accuracy of AUC was 0.0025.

Investigate how reclassification measures varied with the predictive ability

Table 3 reveals how the main reclassification measure varied with different increments of predictive ability. Percentage of total reclassification, NRI and IDI all increased when ΔAUC was altered from 0.05, to 0.10 and 0.15.

Table 3. Reclassification measures with AUC improved by varied increments [c]

ΔAUC[a]	Top % of distribution[b]	Total reclassification (%)	NRI (%)	IDI
0.05	5	7	2.66	0.0108
	10	11	5.55	
	20	18	7.59	
0.10	5	8	9.46	0.0291
	10	14	14.33	
	20	23	15.42	
0.15	5	9	15.8	0.0633
	10	15	21.45	
	20	26	23.85	

[a] ΔAUC was 0.05, 0.10 and 0.15 when the AUC improved to 0.70, 0.75 and 0.80, respectively, from the baseline of 0.60.

[b] The predicted risk at the top percentage of the risk distribution was used as the threshold to define the higher-risk population: the individuals with predicted risks higher than or equal to the threshold were classified as the higher-risk group, and those with predicted risks lower than the threshold were classified as the lower-risk group.

[c] Each value in the table was the mean of 10 simulations.

The percentage of total reclassification, including both correct and incorrect moves across risk categories for individuals, displayed an increase with the decrease of cut-off threshold until it approached around the cut-off threshold of top 40%, then the percentage of total reclassification decreased (Figure 3). The percentage of total reclassification rose with increment of ΔAUC for across fixed thresholds, and it increased with a greater magnitude when the cut-off threshold approached the maximum percentage of total reclassification..

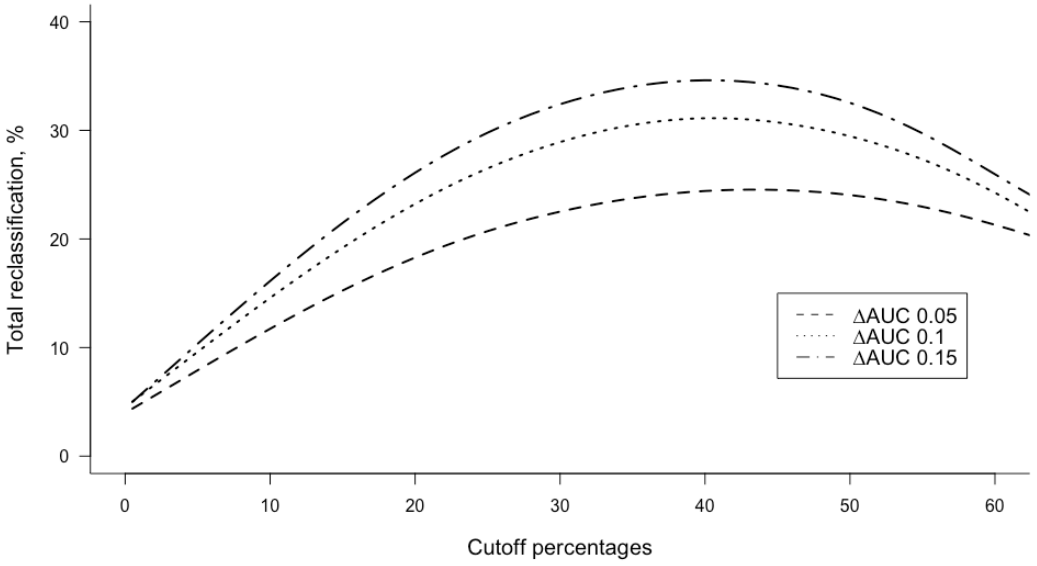


Figure 3. Relationship between total reclassification percentage, ΔAUC and threshold of the high-risk group when AUC varied from 0.65 to 0.70, 0.75 and 0.80. Cut-off percentages (i.e., the percentile of the distribution) were used as the thresholds to define the two risk categories: the individuals with predicted risks higher than or equal to the threshold were classified as the higher-risk group, and those with predicted risks lower than the threshold were classified as the lower-risk group. Disease prevalence = 3.4%, risk allele OR = 1.1, risk allele frequency = 15% and sample size = 100,000. Total reclassification % was presented as the mean of 10 simulations.

The NRI improved first and then dropped with the decrease of cut-off thresholds, with a maximum around top 20-30% of the distribution, and the range of NRI was wider for greater magnitude of ΔAUC. The NRI increased with increment of ΔAUC across cut-off thresholds. (Figure 4).

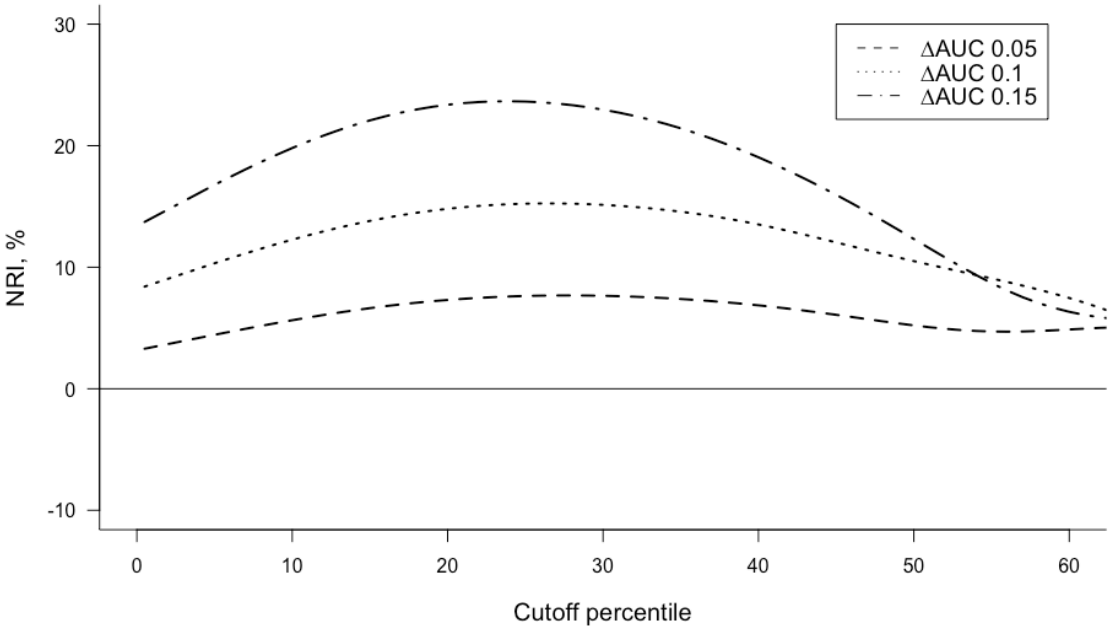


Figure 4. Relationship between total reclassification percentage, ΔAUC and threshold of the high-risk group when AUC varied from 0.65 to 0.70, 0.75 and 0.80. Cut-off percentages (i.e., the percentile of the distribution) were used as the thresholds to define the two risk categories: the individuals with predicted risks higher than or equal to the threshold were classified as the higher-risk group, and those with predicted risks lower than the threshold were classified as the lower-risk group. Disease prevalence = 3.4%, risk allele OR = 1.1, risk allele frequency = 15% and sample size = 100,000. NRI was presented as the mean of 10 simulations.

The improvement of reclassification for cases increased apparently when ΔAUC was greater for a fixed threshold, and the improvement of reclassification for non-cases had only a minor increase (almost remained same) when ΔAUC improved (Figure 5).

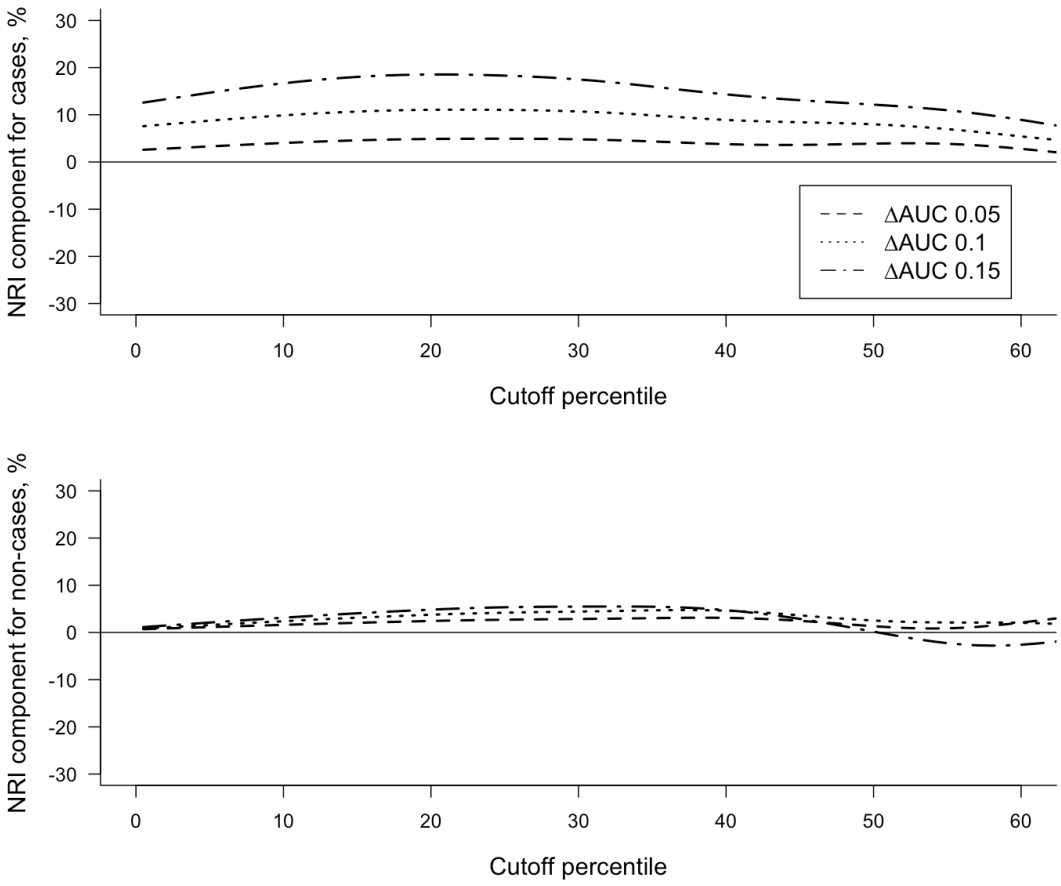


Figure 5. Reclassification improvement for individuals who will develop disease and those who will not, varied with the cutoff threshold and ΔAUC. Cut-off percentages (i.e., the percentile of the distribution) were used as the thresholds to define the two risk categories: the individuals with predicted risks higher than or equal to the threshold were classified as the higher-risk group, and those with predicted risks lower than the threshold were classified as the lower-risk group. Disease prevalence = 3.4%, risk allele OR = 1.1, risk allele frequency = 15% and sample size = 100,000. NRI was presented as the mean of 10 simulations.

Correct and incorrect moves for cases and non-cases

According to Figure 6, the percentage of correct reclassification for cases increased and the percentage of incorrect reclassification for cases decreased with decreasing cut-off thresholds. The opposite trends of the correct and incorrect reclassification with decreasing cut-off apply to cases. The correct moves in risk categories increased for individuals with or without events when ΔAUC improved. The change was minimal for incorrect classification of cases, and also minimal for non-cases when the threshold was close to the higher end of distribution.

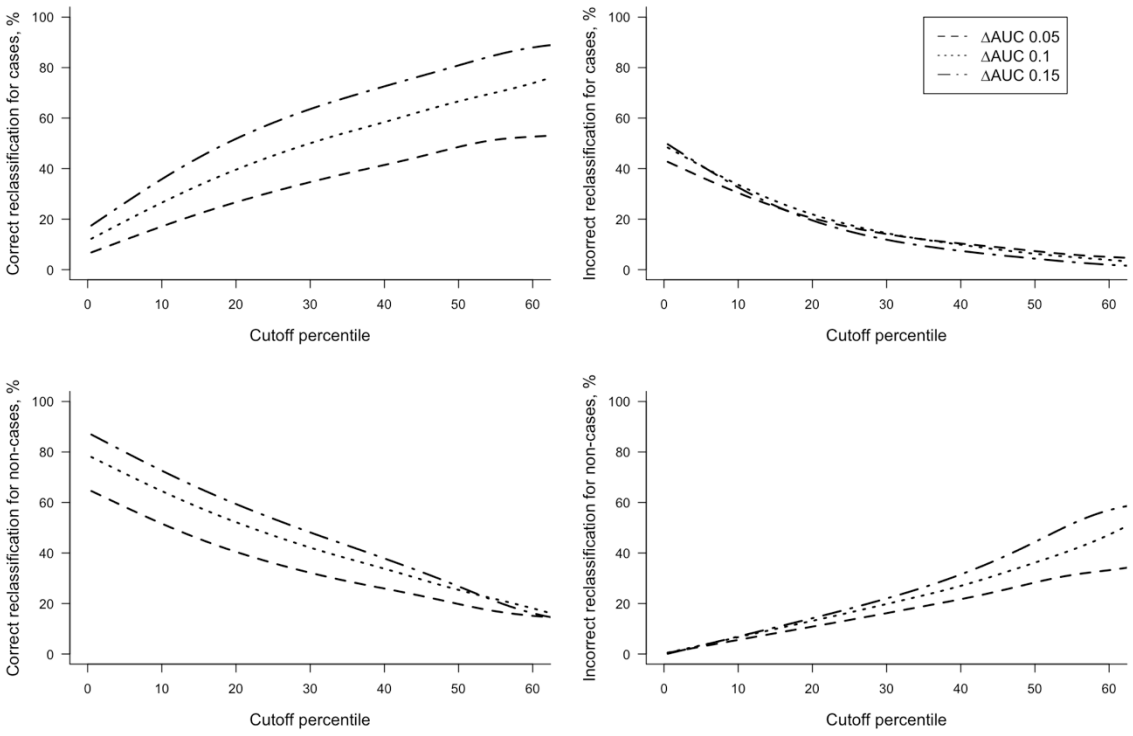


Figure 6. Percentage of Correct and incorrect reclassification for individuals who will develop disease and those who will not, with the cutoff threshold and ΔAUC. Cut-off percentages (i.e., the percentile of the distribution) were used as the thresholds to define the two risk categories: the individuals with predicted risks higher than or equal to the threshold were classified as the higher-risk group, and those with predicted risks lower than the threshold were classified as the lower-risk group. Disease prevalence = 3.4%, risk allele OR = 1.1, risk allele frequency = 15% and sample size = 100,000. Reclassification % was presented as the mean of 10 simulations. Please refer to Appendix 1 for the definition of correct/incorrect reclassification % for cases and non-cases.

Discussion

This study demonstrated that OR of the higher-risk population (versus lower-risk population) increased with improving AUC at an increasing rate for a fixed risk threshold. In line with previous studies (23), sensitivity and PPV of risk prediction increased with increasing AUC for a fixed risk threshold, but specificity and NPV remained almost constant when AUC improved. Moreover, IDI, percentage of total reclassification and NRI increased with increasing ΔAUC. The reclassification improvement is higher for cases than for non-cases at the same ΔAUC, and the same goes for the increase of reclassification improvement when the ΔAUC improved.

Based on the comparative analyses for the identification of higher-risk group, we can draw several inferences about the relationship between the magnitude of increased risk in the higher-risk group and improvement of predictive ability, and the relationship between reclassification measures and increment of predictive ability. First, in line with earlier studies, a higher AUC is needed in order to identify the population at substantially increased risk for a fixed risk threshold. This can be explained using the reclassification measure. When AUC improved, the NRI was positive (Figure 4), which indicates more individuals were correctly reclassified in the risk stratification, resulting in more cases in the higher-risk group and fewer cases in the lower-risk group, followed with higher relative risk of disease in the higher-risk group. Also, the percentage of total reclassification and the NRI increased with improving ΔAUC (Figure 3 and 4), which was in concordance with previous studies (7). This suggests that the greater magnitude of AUC would bring about more correct reclassification, leading to increased OR for the higher-risk group.

Second, the OR of the higher-risk group dropped when the risk threshold decreased until the minimal value of OR at risk thresholds of around 50% of the risk distribution (i.e., the disease prevalence), and the distribution of the OR was symmetrical with respect to the risk cut-off percentiles (Figure 1b), which means the OR was not monotonically increasing against the cut-off percentiles. What is more, the values of OR were insensitive to the percentiles change of

distribution in a limited range, with a minimum at cut-off thresholds around the top 50% of risk distribution (Figure 1b). For example, the minimal OR we could identify was around 2.4 when AUC was 0.65 and the minimum was about 4.7 when AUC was 0.75. Noted that using OR to define risk categories requires considering the minimal value of OR for a specific AUC and the corresponding risk percentile of the distribution. Therefore, we used the top percentage of distribution for risk stratification in this study, allowing for the comparison of different predictive ability across a universal range of cut-off thresholds.

Third, similar to previous results(24), sensitivity, specificity, PPV and NPV were impacted at different magnitudes by AUC and cut-off thresholds, which might be related to the disease prevalence. To begin with, the discrepancy in change of sensitivity/PPV and specificity/NPV with increasing AUC is in line with the reclassification measures for cases and non-cases in this study, which demonstrated that the reclassification improvement for cases was much higher than non-cases with increasing ΔAUC. The reclassification improvement was also higher in cases than non-cases for each ΔAUC (Figure 5). This will be discussed later in this section. In other words, the improvement of AUC would generate more favorable reclassification for cases than for non-cases, which is the reason for greater influence of AUC on sensitivity and PPV compared with specificity and NPV. Additionally, the changes of sensitivity and PPV were different with increasing AUC and cut-off thresholds. The rate of increase for sensitivity with improving AUC decreased with the increasing cut-off threshold; in contrast, the rate of increase for PPV increased with sensitivity had a moderate decrease when the risk cut-off increased even at a relatively low AUC. On the contrary, the increase of PPV held a higher rate when the cut-off threshold increased and there was a minor increase in PPV when the risk cut-off increased at a relatively low AUC. The patterns of sensitivity and PPV we observed is in line with the earlier study by Mihaescu et al. (24), where the researcher reported that only when the cut-off percentile approximated disease prevalence, the improved AUV increased both sensitivity and PPV; otherwise, either sensitivity or PPV would be increased. We also obtained similar results in the finding that sensitivity and PPV were closer when the higher-

risk group frequency approximated the disease frequency.

Fourth, with increasing ΔAUC, NRI improved mainly by increasing the reclassification improvement for cases, which was also indicated in previous studies (25). Reclassification improvement for cases and reclassification improvement for non-cases are the components of NRI. By comparing the two components, we found the improvement in reclassification of cases was much higher than that of non-cases (Figure 5). Figure 6 can validate this distinction between the reclassification improvement of cases and non-cases. For the improvement of ΔAUC at a fixed cut-off threshold, the correct reclassification percentage of cases increased with a higher magnitude than the correct reclassification percentage of non-cases. Meanwhile, the incorrect reclassification % for cases almost remained the same and even decreased for most of the cut-off thresholds with the increase of ΔAUC. By contrast, the incorrect reclassification % for non-cases was also almost unchanged, but for some cut-off thresholds, the incorrect reclassification % for non-cases increased by a small increment with the increase of ΔAUC. In short, the greater increase in correct reclassification % and the more negative change in incorrect reclassification % for cases, compared with the metrics for non-cases, suggests greater improvement for reclassification of cases versus non-cases. From this perspective, the specific reclassification measure for cases and non-cases play a more critical role in demonstrating the mechanism of reclassification. Nevertheless, the distinction between reclassification improvement for cases and reclassification improvement for non-cases we observed in this study conflicted with the demonstration in the study of Mihaescu et al.(7), who stated that for a cut-off value higher than the risk of the disease, reclassification was markedly improved for cases and worsened for non-cases and vice versa for a cut-off value lower than the disease risk. But in our study, the reclassification improvement for non-cases was positive for most cut-off thresholds (Figure 5). This inconsistency might be related to the difference of starting AUC (0.65 in this study versus 0.61 in their study), which would affect the risk distribution of cases and non-cases (Appendix 3) and change the reclassification measures.

Fifth, AUC, parameters of predictive performance (sensitivity, specificity, PPV and NPV) and reclassification measures inform different aspects of risk prediction, as mentioned in earlier studies(7, 24). AUC evaluates the overall predictive performance (26), and reflects the clinical validity. Sensitivity and specificity assess the predictive performance by demonstrating the possibility of four test outcomes (i.e., true positives, true negatives, false positives and false negatives), and they varied with cut-off thresholds (Figure 2); PPV and NPV are not stable parameters to describe the predictive performance as they depend much on disease prevalence. Reclassification measures reflect the clinical utility when the cut-off threshold for risk stratification motivates different preventive strategies or treatment options (25). When ΔAUC improved, there was an evident improvement in total reclassification %, NRI, and correct moves for cases and non-cases. However, NRI has been criticized for being unstable in some circumstances (27), and it is substantially influenced by cut-off thresholds (Figure 4). NRI based on percentile threshold has been believed to be relatively more stable. It was recommended by previous studies that NRI should be used cautiously as an addition to other evaluation tools (28). Evidently, the cut-off thresholds are influential factors to quite a few parameters and measures in the evaluation of risk prediction. The optimal cut-off for defining the higher-risk group would be determined by balancing the benefits and costs as a consequence of false positives and false negatives, which varies in different epidemiologic situation (such as different disease prevalence in the target population) (26). For instance, a high cut-off was better for a deadly disease of high prevalence to increase sensitivity; a lower cut-off was more appropriate for a test aimed to identify patients for an invasive or expensive treatment of a low-prevalence disease, which increases the specificity.

This study started with replicating the findings in the study of Khera et al. (11, 17). Khera et al. used the UK Biobank cohort (n=409,258) to conduct an empirical study on five common diseases, for which they developed, validated and applied polygenic scores to identify individuals with risk similar to monogenic mutations. In the study of Khera et al. (11, 17), the population in the top 8.0, 2.3 and 0.5% of polygenic scores distribution were identified at three-, four- and fivefold increased

risk with the remainder as the reference population for CAD, and the population in the top 6.1, 1.5 and 0.7% of the distribution were identified at three-, four- and fivefold increased risk compared to the remainder for Atrial fibrillation. The AUC of their prediction model based on polygenic scores was 0.81 for CAD and 0.77 for atrial fibrillation, after additional adjustment for two related phenotypic variables (i.e., age and sex). Our validation results show that we were able to derive the values of OR, three-, four- and five-fold respectively for top 8, 2.3 and 0.5% of distribution for CAD (disease prevalence=3.4%) when the AUC was 0.67, and the similar magnitudes of OR for top 6.1, 1.5 and 0.7% for atrial fibrillation (disease prevalence=1.7%) when AUC was 0.67-0.68. The AUC we obtained for CAD was close to the AUC they reported in their preprint version of the article (17), the value of which was 0.64 when they did not adjust for age and sex. We then further validated the similarity in the predictive performance of the prediction model with the AUC of 0.67 from our results and that with the AUC of "0.81" in the study of Khera et al., by comparing our OR, sensitivity, specificity, PPV and NPV for CAD with the corresponding values in their preprint (17), where they announced the CAD risk in both high polygenic-score group (i.e., higher-risk group) and the remainder when the cut-off was at top 2.5% of the distribution. Appendix 2 shows that the OR was 4.0 vs. 3.96, which was almost the same, and all other measures of predictive performance, including sensitivity, specificity, PPV and NPV, were also consistent. In terms of their high AUC (0.81) after adjusting for age and sex (11), it might because that the UK Biobank study involved a cohort with a wide range of age, leading to the inflation of the AUC since age was a traditional risk factor for CAD (29). In addition, researchers argued that it might not be appropriate to use "three-fold increased risk", since the "three-fold" was actually referring to the OR compared with the remainder of distribution rather than the comparison of absolute risks, and the risk ratio was lower than 3 (2.6) for CAD (29).

Strengths and Weaknesses

The key strengths of this study lie in the design of simulation study. When empirical data are not available, simulation studies can be used to create a hypothetical population with genetic profiles and disease status based on Bayes' theorem, construct genetic risk prediction models, and assess the predictive ability. The simulation study not only save time and cost for data collection, but also is more flexible for investigating the mechanisms behind the influence of predictive ability on the identification of higher-risk group and reclassification measures by varying the risk factors. Using large sample size and the average results from 10 simulations, this study could provide stable results and insight on how AUC impacts the risk prediction, which are easier to replicate. Nevertheless, this study only involved a dichotomous genetic test, which simplified the reclassification of risk categories. The number of risk categories should be determined by clinical utility based on the impacts on clinical decisions and medical practice (30, 31) and more risk thresholds would cause a higher percentage of reclassification (22). In addition, the simulate data cannot be used to adjust for other common covariates (e.g. sex, age and other disease related risk factors), leading to difference between the AUC in the simulated data and the AUC in the practice.

Conclusions

Our findings suggest that we can identify the higher-risk population with increased OR for common multifactorial diseases across risk thresholds compared with the remainder when the predictive ability of genetic risk model improves. We have shown that the sensitivity and PPV of the genetic risk prediction are strongly influenced by AUC and the association was positive. In addition, reclassification measures have favorable increase when the ΔAUC improves, and this is achieved mainly by improving the reclassification of individuals with events.

References

1. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* 2018;19(9):581-90.
2. Visscher PM, Wray NR, Zhang Q, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* 2017;101(1):5-22.
3. Tada H, Melander O, Louie JZ, et al. Risk prediction by genetic risk scores for coronary heart disease is independent of self-reported family history. *Eur Heart J* 2016;37(6):561-7.
4. GWAS to the people. *Nat Med* 2018;24(10):1483.
5. Ripatti S, Tikkanen E, Orho-Melander M, et al. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *The Lancet* 2010;376(9750):1393-400.
6. Dunlop MG, Tenesa A, Farrington SM, et al. Cumulative impact of common genetic variants and other risk factors on colorectal cancer risk in 42,103 individuals. *Gut* 2013;62(6):871-81.
7. 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(3):353-61.
8. Golan D, Lander ES, Rosset S. Measuring missing heritability: inferring the contribution of common variants. *Proc Natl Acad Sci U S A* 2014;111(49):E5272-81.
9. Fuchsberger C, Flannick J, Teslovich TM, et al. The genetic architecture of type 2 diabetes. *Nature* 2016;536(7614):41-7.
10. Natarajan P, Young R, Stitziel NO, et al. Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting. *Circulation* 2017;135(22):2091-101.
11. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 2018;50(9):1219-24.
12. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143(1):29-36.
13. 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.
14. Simmons RK, Sharp S, Boekholdt SM, et al. Evaluation of the Framingham risk score in the European Prospective Investigation of Cancer-Norfolk cohort: does adding glycated hemoglobin improve the prediction of coronary heart disease events? *Arch Intern Med* 2008;168(11):1209-16.
15. 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.
16. 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(2):157-72; discussion 207-12.
17. Khera AV, Chaffin M, Aragam K, et al. Genome-wide polygenic score to identify a monogenic risk-equivalent for coronary disease. 2017.
18. Janssens ACJW, Aulchenko YS, Elefante S, et al. Predictive testing for complex diseases using

multiple genes: Fact or fiction? *Genetics in Medicine* 2006;8(7):395-400.

19. Kundu S, Kers JG, Janssens AC. Constructing Hypothetical Risk Data from the Area under the ROC Curve: Modelling Distributions of Polygenic Risk. *PLoS One* 2016;11(3):e0152359.
20. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983;148(3):839-43.
21. Steyerberg EW, Pencina MJ, Van Calster B. [Discriminative ability, reclassification and net benefit: the predictive value of biomarkers]. *Ned Tijdschr Geneeskd* 2012;156(41):A5029.
22. 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.
23. Weng SF, Reps J, Kai J, et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* 2017;12(4):e0174944.
24. Mihaescu R, Moonesinghe R, Khoury MJ, et al. Predictive genetic testing for the identification of high-risk groups: a simulation study on the impact of predictive ability. *Genome Med* 2011;3(7):51.
25. McKearnan SB, Wolfson J, Vock DM, et al. Performance of the Net Reclassification Improvement for Nonnested Models and a Novel Percentile-Based Alternative. *Am J Epidemiol* 2018;187(6):1327-35.
26. Greiner M, Pfeiffer D, Smith RD. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Prev Vet Med* 2000;45(1-2):23-41.
27. 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(19):3405-14.
28. Pencina MJ, D'Agostino RB, Pencina KM, et al. Interpreting incremental value of markers added to risk prediction models. *Am J Epidemiol* 2012;176(6):473-81.
29. Janssens A, Joyner MJ. Polygenic Risk Scores That Predict Common Diseases Using Millions of Single Nucleotide Polymorphisms: Is More, Better? *Clin Chem* 2019.
30. Meigs JB, Shrader P, Sullivan LM, et al. Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N Engl J Med* 2008;359(21):2208-19.
31. Muhlenbruch K, Heraclides A, Steyerberg EW, et al. Assessing improvement in disease prediction using net reclassification improvement: impact of risk cut-offs and number of risk categories. *Eur J Epidemiol* 2013;28(1):25-33.

Appendix 1: Definitions and Formulas for Reclassification Measures

Net Reclassification Improvement (NRI) is defined as the sum of reclassification improvement of cases and non-cases. The **correct reclassification for cases** refer to the upward movements (from the lower-risk group to higher-risk group) in cases, and the **correct reclassification for non-cases** refers to the downward movements (from the higher-risk group to lower-risk group) in non-cases(16). The movements of the opposite direction are the **incorrect reclassification** for cases and non-cases. The difference in proportion of correct reclassification and proportion of incorrect reclassification in cases is the **reclassification improvement of cases**. Similarly, the difference in proportion of correct reclassification and proportion of incorrect reclassification in non-cases is the **reclassification improvement of non-cases** Thus, NRI include two components, reclassification improvement of cases and Reclassification improvement of non–cases.

i.e.,

$$NRI = [Pr(up|case) - Pr(down|case)] + [Pr(down|non-case) - Pr\ (up|non-case)] \quad (22)$$

where

$$Pr(up|case) = \frac{\#\ \text{cases moving up}}{\#\ \text{cases}}$$

$$Pr(down|case) = \frac{\#\ \text{cases moving down}}{\#\ \text{cases}}$$

$$Pr(up|non-case) = \frac{\#\ non-case\ \text{moving up}}{\#\ non-case}$$

$$Pr(down|non-case) = \frac{\#\ non-case\ \text{moving down}}{\#\ non-case}$$

Integrated Discrimination Improvement (IDI) is defined as the improvement in difference of average predicted risks for cases and non-cases between the baseline model and updated model. Noted that risk category is not taken into consideration when we calculate IDI (16).

$$IDI = RD_{updated} - RD_{baseline}$$

i.e.,

$$IDI = \left(\bar{\hat{p}}_{2,1} - \bar{\hat{p}}_{2,0}\right) - \left(\bar{\hat{p}}_{1,1} - \bar{\hat{p}}_{1,0}\right)$$

where $\hat{p}$ is the predicted risk, the first subscript of $\hat{p}$ denotes the model (1 = baseline model; 2 = updated model) and the second subscript denotes the disease status of individuals (0 = non-cases; 1 = cases). Namely, $\bar{\hat{p}}_{2,1}$ denotes the mean of predicted risks in cases based on the updated model, $\bar{\hat{p}}_{2,0}$ denotes the mean of predicted risks in non-cases based on the updated model, $\bar{\hat{p}}_{1,1}$ denotes the mean of predicted risks in cases based on the baseline model, and $\bar{\hat{p}}_{1,0}$ denotes the mean of predicted risks in non-cases based on the baseline model.

Appendix 2: Comparison of the Validation Results and the Study of Khera et al.

Table 1. Validation of OR and Predictive Performance for the Study of Khera et al.

AUC	Top% of distribution[a]	OR	Sensitivity	Specificity	PPV	NPV
0.67[b]	2.5	4.00	0.086	0.977	0.116	0.968
0.81[c]	2.5	3.96	0.076	0.977	0.092	0.972

[a] The top percentage of the risk distribution was used as the threshold to define the higher-risk population.

[b] Results from our validation study, using the parameters: disease prevalence = 3.4%, allele OR = 1.1, allele frequency = 15% and sample size = 100,000. Each value was the mean of 10 simulations, and the accuracy of AUC was 0.0025.

[c] Results from the study of Khera et al. (17), of which the AUC was doubted.

Appendix 3: Risk Distributions of Cases and Non-cases for Varied AUCs

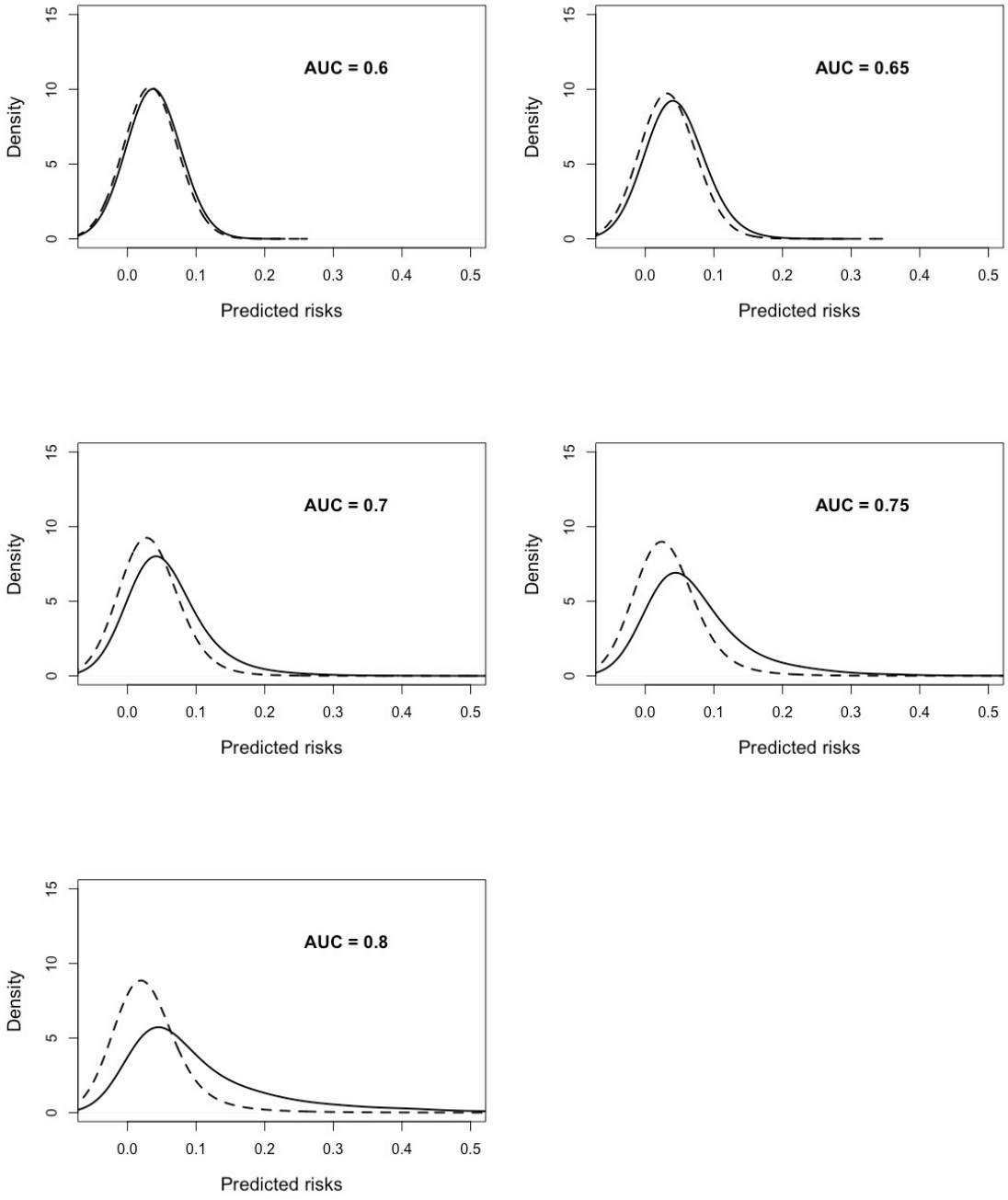


Figure 1. Distribution of predicted risk associated with genetic profiles for individuals who will develop the disease (solid lines) and who will not (dashed lines). Disease prevalence = 3.4%, risk allele OR = 1.1, risk allele frequency = 15% and sample size = 100,000. AUC=0.60-0.80 by the increment of 0.5 for the plots, and the accuracy of AUC was 0.0025.