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:

_____ _____

Kunru Ning Date

Value of SNPs With Very Small Effects in The Predictive Ability of Polygenic Risk Score: Illustrated Using Type II Diabetes Mellitus

By

Kunru Ning
Master of Public Health

Department of Epidemiology

_____

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

Value of SNPs With Very Small Effects in The Predictive Ability of Polygenic Risk Score: Illustrated Using Type II Diabetes Mellitus

By

Kunru Ning

B.S.
University of Washington
2018

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 Public Health
in Epidemiology
2021

Abstract

Value of SNPs With Very Small Effects in The Predictive Ability of Polygenic Risk Score: Illustrated Using Type II Diabetes Mellitus

By Kunru Ning

Background: Polygenic risk score (PRS) is an index calculated by summing up multiplications of the number of risk alleles and effects of the single nucleotide polymorphism (SNP) to predict the risk of developing diseases for individuals. PRSs calculated by thousands and millions of SNPs with small effects have become a trend in many studies. We studied the value of including SNPs with very small effects when predicting disease risk using PRS, answering the following research questions: 1. Will the discriminative accuracy change when changing the precision of SNPs with small effects? 2. Will the predicted risks change when changing the precision of SNPs?

Methods: The analysis was conducted using simulated data. Hypothetical populations of 100,000 people, in which we predicted disease using 7502 SNPs associated with type II diabetes at a population risk of 10% and 30%. A genetic profile with predicted disease risk and disease status was generated for each individual. AUC was calculated to quantify the predictive ability of the PRS models using SNPs with small effects that are non-zero in the 6^{th}, 5^{th}, 4^{th}, 3^{rd}, 2^{nd}, and 1^{st} decimal. Predicted risks were calculated 6 times, using SNP effects that are non-zero in the 6 decimal levels. Correlation coefficients were obtained for the predicted risks to measure the strength of association.

Results: When the SNP effects was kept in the 5^{th}, 4^{th} and 3^{rd} decimals, the predicted ability of PRS remained the same (AUC = 0.65, correlation coefficient = 1.0) comparing with when the SNP effects was kept in the 6^{th} decimal, for both population risk of 10% and 30%. When SNP effects were kept in the 2^{nd} decimal, the AUC remained the same at 0.65 with a slightly lower correlation coefficient of 0.97. The AUC and correlation coefficient reduced to 0.64 and 0.74 when SNP effects kept in the 1^{st} decimal.

Conclusion: The study concluded that including SNPs with very small effects in the predictive ability of PRS did not change predicted risks. Therefore, there is no necessity of including SNPs with very small effects when predicting PRS in future studies, keeping SNP effects in the 2^{nd} decimal with an precision of 0.01 should suffice.

Keywords: Polygenic risk score; Single nucleotide polymorphisms; Genetics; Simulation study

Value of SNPs With Very Small Effects in The Predictive Ability of Polygenic Risk Score: Illustrated Using Type II Diabetes Mellitus

By

Kunru Ning

B.S.
University of Washington
2018

Thesis Committee Chair: A. Cecile J.W. Janssens, Ph.D.

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 Public Health
in Epidemiology
2021

Acknowledgements

First, I would like to thank my thesis advisor, Dr. Janssens, for her guidance and patience in helping me with forming the research question and writing the manuscript.

I would also like to thank my field advisor, Qi Zhang, MSPH, for her help in data analysis and coding.

I would further want to thank Emory Rollins School of Public Health, where I have made some lifetime friends, for the most wonderful and meaningful two years in my life. At Rollins, I made my very first baby step towards my career as an epidemiologist, and most importantly, I learned the mission of being a healthcare worker who is responsible for building a better world for humankind.

Lastly, I want to thank my family, for their support emotionally and financially. Without them, I would not be able to make the achievement today. Particularly, I would like to thank my boyfriend Adam, who has become my fiancée now, for his unconditional love and encouragement.

Table of Contents

INTRODUCTION

Polygenic risk score (PRS) is an index calculated by summing up multiplications of number of risk allele and effect size of the single nucleotide polymorphism (SNP), which derived from Genome Wide Associated Studies (GWAS) to predict the risk of developing diseases for individual[1]. PRS has been argued to show its potential in utilization in clinical applications, such as PRS-informed therapeutic intervention, in which PRS may provide information on therapeutic and preventative intervention selection, and PRS-informed disease screening, in which PRS may provide information on disease screening initiation and interpretation.[2,3,4]

PRSs are calculated by tens to hundreds of statistically significant SNPs with effect size to predict disease risk.[3,4] Recent studies get interested in using millions of SNPs when calculating PRS.[5] Study by Khera validated their genome-wide polygenic scores (GPS) conducted from 6.6 million SNPs for five common diseases (coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer). Following Khera, including millions of SNPs when calculating PRS has become a trend.[6,7,8]

Khera et al. found that 0.01% of SNPs assumed to be causal for CAD, indicating the rest 99.9% SNPs have very small effect sizes and have very low impact on disease risk prediction.[9] Bolli et al. recalculated 4 PRSs for the 6.6 million SNPs associated with coronary artery disease from Khera et al. at 100%, 1%, 0.1% and genome-wide significant level and found that AUC for the full PRS panel was 0.011 higher than AUC for top 1% PRS, suggesting that the millions of SNPs with smaller effects do not contribute to disease risk that much.[10] Higher AUC can be obtained with higher effective sizes of SNPs related to the disease.[11] Small effect size with small AUC indicates being indiscriminative between cases and non-cases.[11] Improvement of AUC (Δ

AUC) is used to assess the value of new adding risk factors. New adding risk factor with small Δ AUC should be associated small change in predicted risk.[12]

From 6.6 thousand to 6.6 million SNPs included, Δ AUC showed a change of 0.011 of predictive ability of PRS for CAD.[10] A small Δ AUC of 0.011 is reasonably to be assumed to have a small change in predicted risk of CAD among the study population. Since the change in predicted risk is so low, skepticism arises on the necessity of including millions of SNPs with small effect in PRS when predicting diseases risk in clinical setting.[13] SNPs with small effects reported in PRS, are often kept at the precision level of 0.000001. It is still unclear that when the precision of SNP effects is changed to 0.00001, 0.0001, 0.001, 0.01 and 0.1, whether the predictive ability of PRS will also be changed, characterized by AUC and the predicted risk change.

We investigated the AUC and predicted risk change using a hypothetical population of 100,000, illustrated by SNPs associated with type II diabetes.[14] We are conducting simulated study since predictive ability of SNPs in genetic risk models can be approached by simulated population from other GWAS studies.[15] Empirical studies has some limitation that only simulated study can overcome, as the predictive ability of empirical studies is only warranted when sufficient predictive ability is expected.[15] We hypothesize that SNPs with small effect do not affect predictive ability of PRS.

Therefore, we studied the public health relevance of including SNPs with small effects when predicting disease risk using PRS, answering the following research questions: 1. Will the AUC change when changing precision of SNPs with small effects? 2. Will the risk itself change when changing precision of SNPs with small effects? Keeping large amount of statistically nonsignificant results while reporting PRS in terms of disease risk prediction may generate

confusion when interpretating the results to the general public. Addressing this issue is important in public health, as we don't go way beyond what GWAS needs, when the change of disease risk is so small to be considered negligible.

METHOD

The analysis was conducted using hypothetical data. We assessed the contribution of including SNPs with small effects when calculating PRS, by investigating the change in discriminative ability of PRS and the estimates of the risks. We analyzed 1. AUC change between the original and 5 updated models, which were the hypothetical populations re-constructed using effect sizes with 5 different precision levels; 2. Predicted risk change for the simulated subjects in the original model when assigning different precision levels of effect sizes to the risk alleles; 3. Whether results seen in the previous two research questions could be applied to both common and rare diseases.

Data Collection

We simulated a PRS for type 2 diabetes using SNP data from PGS catalog (www.pgscatalog.org/downloads/#dl_ftp). The dataset contains 7502 SNPs with the risk alleles, effect sizes, and risk allele frequencies.[14]

Data Preparation

The simulated data with genetic profile and disease status for a hypothetical population, the AUC, and predicted risks were prepared using the "PredictABEL" package in R. Details are explained in the "PredictABEL" package.

Simulating Population

To simulate a hypothetical population with individual genetic profile and disease status, the population risk, risk allele ORs, risk allele frequencies, and sample size should be specified as the input parameters. The population risk for type II diabetes in the United States in 2017 was 8911 cases per 100,000 individuals, setting the population risk parameter to 8.91%, which we rounded to 10% for ease of illustration. ORs and risk allele frequencies are obtained from the

PGS catalog dataset.[16] We converted the beta coefficient into odds ratios (ORs) using the following equation: $OR = e^{\beta}$. Sample size for the reference model was set to be 100,000.

Calculating AUC

AUC was calculated as the c-statistic for the model, according to the predicted risk and disease status (explained in the package) of the simulated population, to quantify the predictive ability of the PRS model. To obtain AUC for each model, we regenerated 5 simulated datasets with 5 updated models. Each model has a different precision level of risk allele OR. We rounded ORs to precision level of 0.00001, 0.0001, 0.001, 0.01 and 0.1 for each model. All other parameters were kept the same (population risk, risk allele, and risk allele frequency).

Calculate predicted risk

In a separate analyses, we compared the risk prediction at the individual level. Predicted risk for each individual was calculated as the probability which converted from posterior odds, using the equation *probability=odds/(1+odds).* Posteriors odds were obtained from Bayes' theorem using likelihood ratios and calculation of LR has been described previously somewhere else (Janssens et al., 2006). We calculated predicted risks for each individual within the reference model, using ORs with precisions of 0.000001, 0.00001, 0.0001, 0.001, 0.01 and 0.1. Each individual was calculated with 6 predicted risks.

Statistical Analysis

We analyzed the changes in discriminative ability of PRS by investigating the AUC in different models. The change in predictive values was examined by and change of risk itself within the same model when changing the precision level kept for risk allele effects .

First, we analyzed AUC change when changing the precision level of risk allele ORs. We arbitrarily identified the level of threshold for being considered as a change for the Δ AUC to be 5%.

Next, we analyzed the risk change when changing the precision level of risk allele effects. To show the relationship between risks obtained from different effects precision levels, a set of scatterplots was generated, having the predicted risk calculated from effects precision level of 0.000001 on the x-axes, and predicted risk from effects precision level of 0.00001, 0.0001, 0.001, 0.01 and 0.1 on the y-axes. Correlation coefficients(r) were also obtained for the scatterplots to show the strength of relationship.

In addition, we analyzed the AUC and risk change for common and rare diseases when using effects with different precision levels, by repeating the previous analysis steps by a changing disease population risk to 30%, holding the rest of the parameters constant.

All analyses were performed with R (version 4.0.2) using “PredictABEL” package.

RESULTS

Table 1 presents the general information of the reference model and updated models for the hypothetical populations, including different precision levels of effects, number of SNPs included for calculation in the model, model performance demonstrated by AUC and Δ AUC between the reference and updated models. When precision of ORs changed from 0.000001 to 0.001, over 7000 SNPs in the model had a OR did not equal to 1, whereas the number of SNPs reduced to 2442 when ORs were rounded to 0.01 in model 4 and 60 when ORs were rounded to 0.1 in model 5, saying that 60 of the 7502 SNPs were included in model 5 to calculate the predictive ability of PRS. The discriminative accuracy of the models was presented by AUC. Δ AUC has also shown the change in AUC between the reference and updated models.

When changing the precision of ORs of the SNPs, we observed no change in AUC between the reference and updated models (Table 1). The AUC for the reference model was 0.65. The AUC calculated from SNPs with OR precision of 0.1 was 0.64.

Table 1. General information of models using ORs with different precisions for a simulated population (Sample size=100,000, population risk =10%)

Model [a]	Precision Level [b]	Number of SNPs [c]	AUC [d]	Δ AUC [e]
Reference Model	0.000001	7502	0.653	-
Model 1	0.00001	7502	0.655	0.2%
Model 2	0.0001	7502	0.652	0.1%
Model 3	0.001	7361	0.654	0.1%
Model 4	0.01	2442	0.656	0.3%
Model 5	0.1	60	0.637	1.6%

[a] Models were used to simulate a hypothetical population of 100,000. Odds ratios (ORs) with precision of 0.000001 from the original dataset was set to generate the reference model. For updated models, ORs were rounded to precisions of 0.00001, 0.0001, 0.001, 0.01, and 0.1 for population simulation.
[b] Precisions were rounded to different levels for each updated model.
[c] Number of SNPs remaining in the model that had a OR not equal to 1 after rounding.
[d] Area under the receiver operating characteristic (ROC) curve.
[e] Absolute difference of AUC between the reference model and updated model, shown in %.

The ROC curves for the reference and updated models are shown in Figure 1. The ROC curves overlapped for the reference and updated models, except for the ROC curve using OR with precision of 0.1, which showed a slightly lower discriminative ability by having a lower AUC.

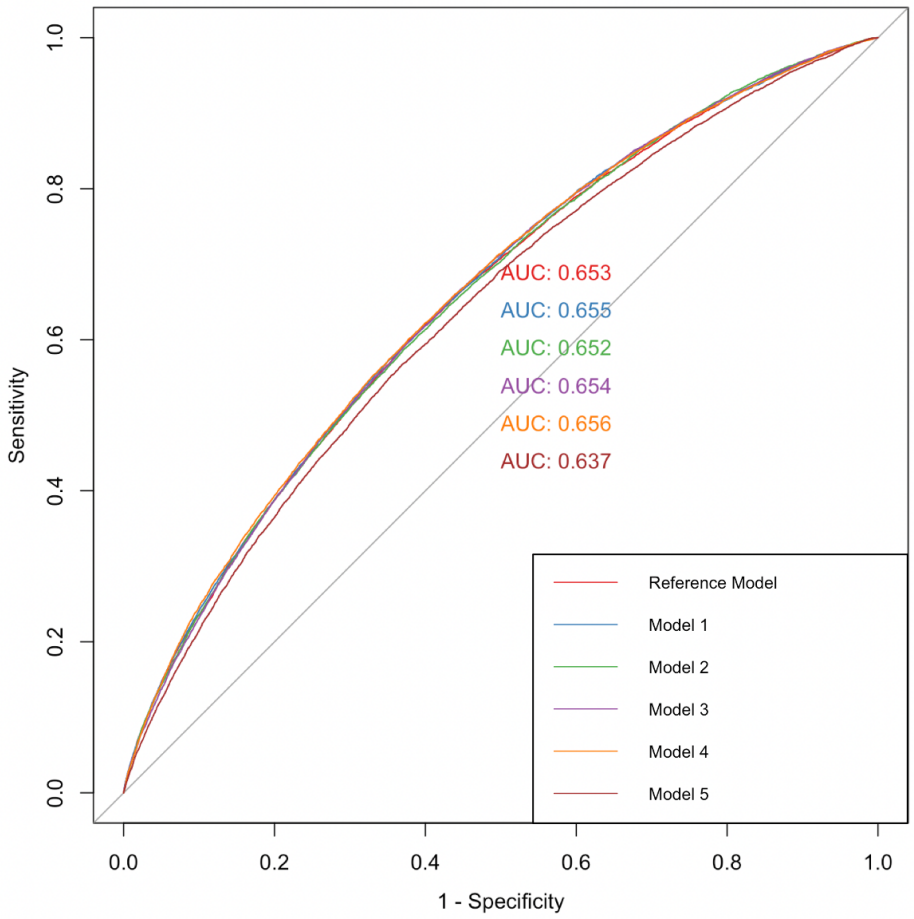

Figure 1. ROC curves for the reference and updated models when simulating a population of 100,000 at population risk of 10%.
Reference model: hypothetical population simulated using ORs with precision level of 0.000001.
Model 1: Hypothetical population simulated using ORs with precision level of 0.00001.
Model 2: Hypothetical population simulated using ORs with precision level of 0.0001.
Model 3: Hypothetical population simulated using ORs with precision level of 0.001.
Model 4: Hypothetical population simulated using ORs with precision level of 0.01.
Model 5: Hypothetical population simulated using ORs with precision level of 0.1.

Figure 2 showed scatterplots of predicted risk change when changing the precision level of ORs for the hypothetical population in the reference model. The predicted risk was calculated 5 additional times, each time using ORs from different precisions. Correlation coefficients between the reference predicted risks and updated predicted risks were also calculated and presented in the scatterplots. When using ORs showing an effect only in the 6^{th}, 5^{th}, 4^{th}, 3^{rd} ,and 2^{nd} decimals, the correlation coefficients of predicted risk between the 6^{th} and 5^{th}, 4^{th}, 3^{rd}, 2^{nd} decimals were all 0.99, indicating strong associations. The correlation coefficient between predicted risk calculated with ORs showing an effect in 6^{rd} and 1^{st} decimal was 0.74, indicating a moderate association.

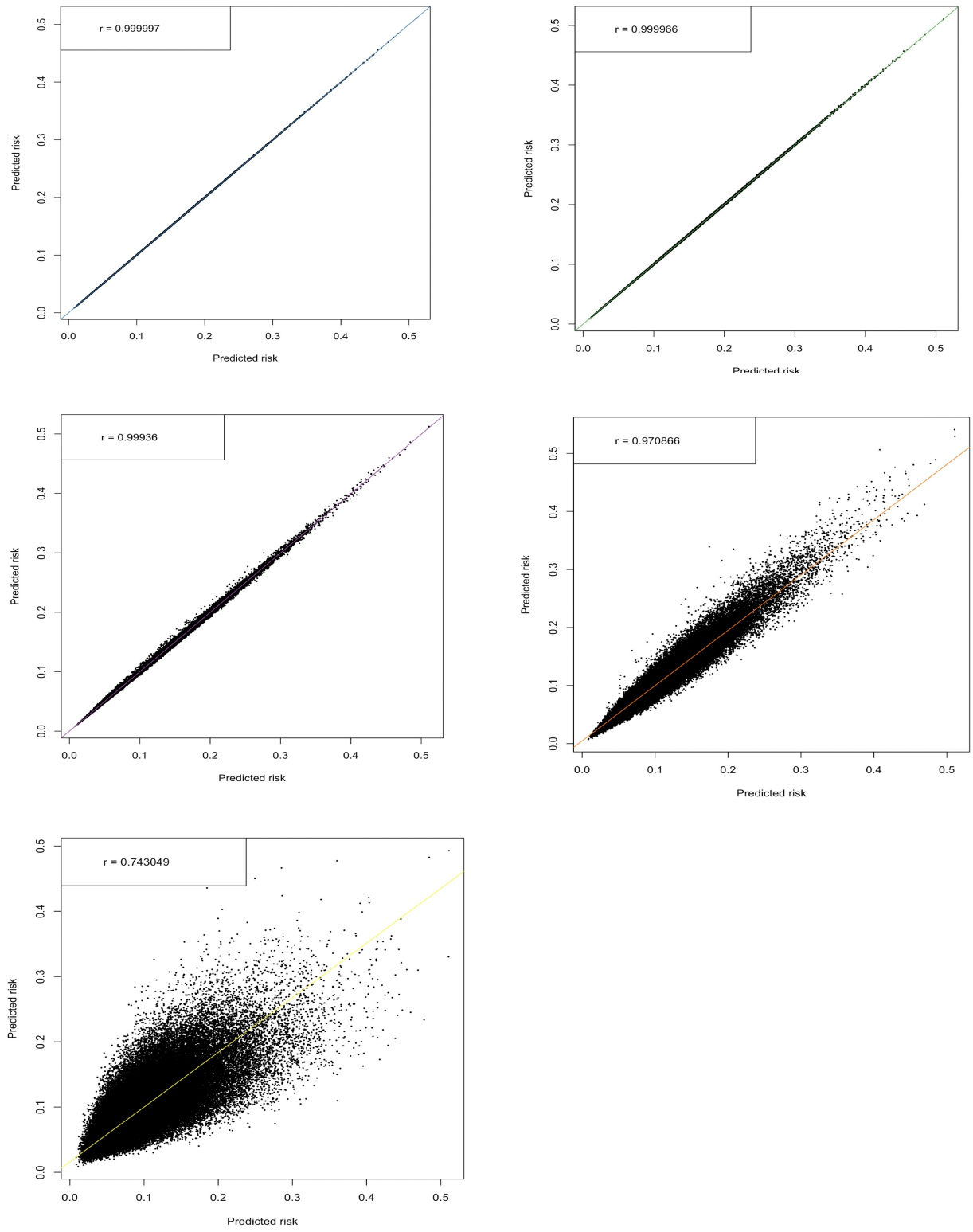

Figure 2. Scatterplots of predicted risks using ORs with different precision levels at population risk of 30%. Each dot represent the predict risk of one person. a) – e), Scatterplots of predicted risks calculated using ORs with precision of 0.000001 vs. predicted risks calculated using ORs with precision of 0.00001, 0.0001, 0.001, 0.01 and 0.1, of the reference model population.

Furthermore, we investigated the predictive ability of PRS at a population risk of 30%, holding the rest of the analysis constant. We observed a similar result for AUC, Δ AUC and predicted risk comparing with population risk at 10%. At population risk of 30%, ROC curves overlapped for the 6th and 5th, 4th, 3rd and 2nd model with OR showing an effect in the 1st decimal (Figure 3).

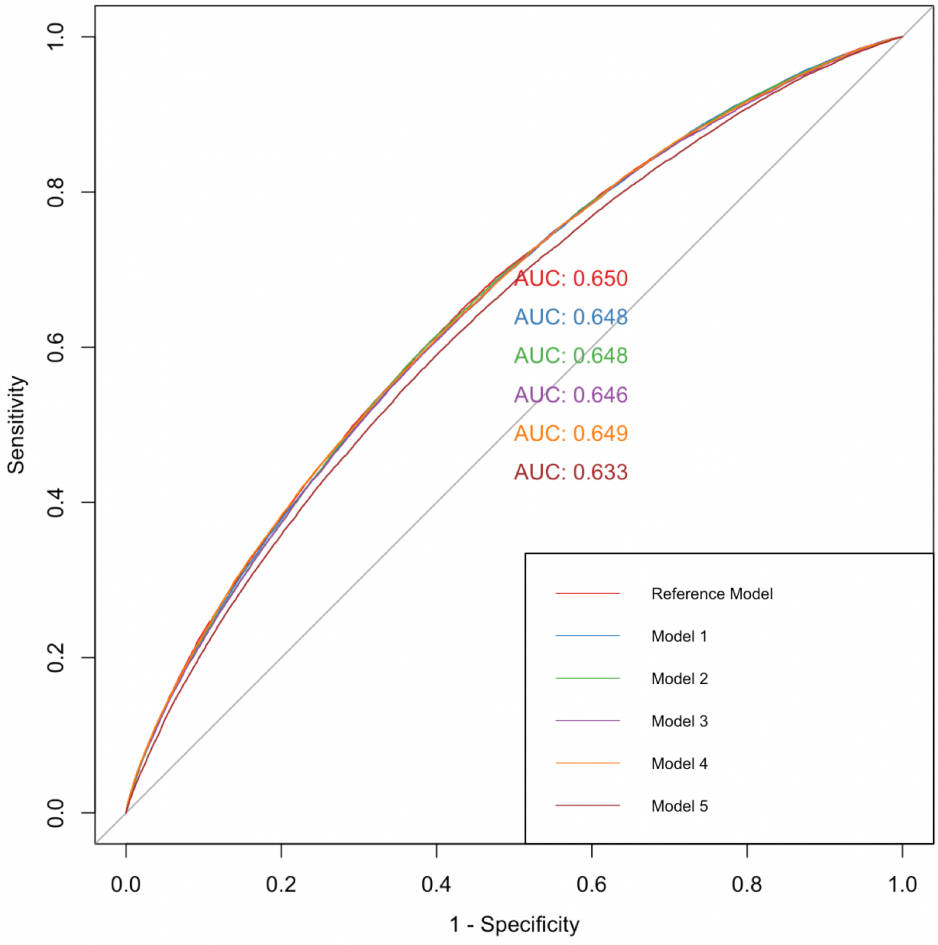

Figure 3. ROC curves for the reference and updated models when simulating a population of 100,000 at population risk of 30%.

Predicted risk calculated with reference and updated precisions of ORs all had correlation coefficients of 0.99, except for the predicted risk using OR with precision of 0.1(correlation coefficient = 0.75) (Figure4.)

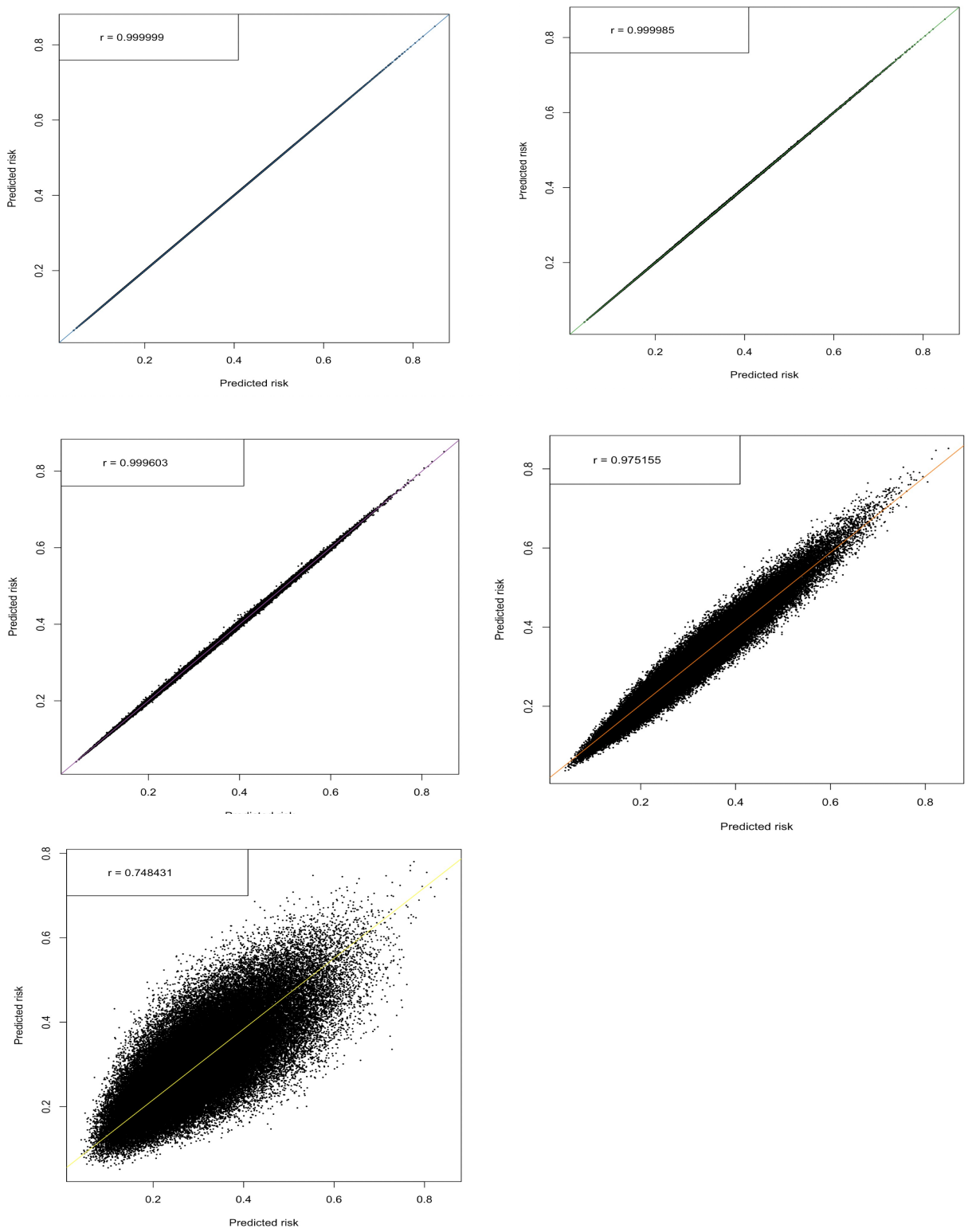

Figure 4. Scatterplots of predicted risks using ORs with different precision levels, at population risk of 30%. Each dot represent the predict risk of one person. a) – e), Scatterplots of predicted risks calculated using ORs with precision of 0.000001 vs. predicted risks calculated using ORs with precision of 0.00001, 0.0001, 0.001, 0.01 and 0.1, of the reference model population.

DISCUSSION

We investigated the value of including SNPs with small effects in the predictive ability of PRS, by investigating the AUC and predicted risk change, when changing the precision of SNP effects. We found that for both of the simulated populations with the disease population risk of 10% and 30%, the AUC and predicted risk showed no difference when the precision of SNP effects were changed from 0.000001 to 0.01 (AUC = 0.65; r = 1.0). At SNP effects precision level of 0.1, the AUC and the correlation coefficient of predicted risk were slightly lower than when using SNP effects with 0.000001 (AUC = 0.64; r = 0.74 for population risk of 10%; r = 0.75 for population risk of 30%). The results showed that the value of including SNPs with every small effects in the predictive ability of PRS is low. There is no necessity of keeping the precision of SNP effects at 0.000001 since the predictive ability of PRS remains the same comparing with when the precision of SNP effects was kept at 0.01.

One key finding of our study was that when modeling with SNP effects of different precision levels, the AUC showed no change for SNP effects with precision of 0.000001 and 0.01 (AUC = 0.65). Previous study also using simulated data for a population of 10,000 found that when AUC remained the same, the predictive ability of the reference and updated model remained the same as well, indicating the discriminative accuracy remained the same.[17] When precision of SNP effects changed to 0.1, we did observe a change in AUC of 0.1% for population risk at 10% and 0.2% for population risk at 30%. Both of the Δ AUCs remained within the threshold we set earlier for Δ AUC to be considered small.

We also investigated the predicted risk change when calculating PRS using SNPs with different precisions for the reference population. With the genetic profile for each individual remaining constant, predicted risk was calculated for 6 times, using SNPs with different

precision of effects. The correlation coefficients between the reference predicted risk (SNP effects precision at 0.000001) and updated predicted risk (SNP effects precision at 0.00001, 0.0001and 0.001) were 1.0 for both population risk of 10% and 30%, and 0.97 for SNP effects precision of 0.01 when population risk was 10%. Correlation coefficient of 0.97 indicates a strong association that predicted risk when using SNP effect precision of 0.01.

Strength and Limitations

The strength of the study was able to perform a simulation study for a population of 100,000. Simulation study has the advantage of including large sample size when empirical data is unavailable. Nevertheless, our study had the limitation of only being able to include 7502 SNPs when modeling the populations. Our purpose was to investigate the value of including SNPs with very small effects in predictive ability of PRS, according to the study with 6.6 millions of SNPs by Khera et al.. But we were not able to model 6.6 million variants due to computer limitations.

Conclusion and Implications

Overall, we quantified the predictive ability change of PRS, with different precisions of SNP effects using simulated data. We found that predictive ability of PRS did not change when the precision of SNP effects changed from 0.000001 to 0.01. Therefore, keeping SNP effects at precision of 0.01 should suffice for future studies when calculating PRS.

REFERENCE

1. National Human Genome Research Institute. Polygenic risk scores. https://www.genome.gov/Health/Genomics-and-Medicine/Polygenic-risk-scores. Published 2020, Aug 11. Accessed.
2. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics.* 2018;19(9):581-590.
3. 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):2091-2101.
4. Mavaddat N, Pharoah PD, Michailidou K, et al. Prediction of breast cancer risk based on profiling with common genetic variants. *J Natl Cancer Inst.* 2015;107(5).
5. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic score to identify a monogenic risk-equivalent for coronary disease. *bioRxiv.* 2017:218388.
6. Inouye M, Abraham G, Nelson CP, et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. *J Am Coll Cardiol.* 2018;72(16):1883-1893.
7. Abraham G, Malik R, Yonova-Doing E, et al. Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. *Nat Commun.* 2019;10(1):5819.
8. Mars N, Koskela JT, Ripatti P, et al. Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat Med.* 2020;26(4):549-557.
9. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics.* 2018;50(9):1219-1224.
10. Bolli A, Di Domenico P, Bottà G. Software as a Service for the Genomic Prediction of Complex Diseases. *bioRxiv.* 2019:763722.
11. Janssens AC, Aulchenko YS, Elefante S, Borsboom GJ, Steyerberg EW, van Duijn CM. Predictive testing for complex diseases using multiple genes: fact or fiction? *Genet Med.* 2006;8(7):395-400.
12. Martens FK, Tonk EC, Kers JG, Janssens AC. Small improvement in the area under the receiver operating characteristic curve indicated small changes in predicted risks. *J Clin Epidemiol.* 2016;79:159-164.
13. Janssens A, Joyner MJ. Polygenic Risk Scores That Predict Common Diseases Using Millions of Single Nucleotide Polymorphisms: Is More, Better? *Clin Chem.* 2019;65(5):609-611.
14. Läll K, Mägi R, Morris A, Metspalu A, Fischer K. Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. *Genet Med.* 2017;19(3):322-329.
15. Kundu S, Mihaescu R, Meijer CMC, Bakker R, Janssens ACJW. Estimating the predictive ability of genetic risk models in simulated data based on published results from genome-wide association studies. *Frontiers in Genetics.* 2014;5(179).
16. Khan MAB, Hashim MJ, King JK, Govender RD, Mustafa H, Al Kaabi J. Epidemiology of Type 2 Diabetes - Global Burden of Disease and Forecasted Trends. *J Epidemiol Glob Health.* 2020;10(1):107-111.

17. 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. *American Journal of Epidemiology.* 2010;172(3):353-361.