Circulation: Genomic and Precision Medicine

ORIGINAL ARTICLE

Population Bias in Polygenic Risk Prediction Models for Coronary Artery Disease

Damian Gola, PhD; Jeanette Erdmann, PhD; Kristi Läll, PhD; Reedik Mägi, PhD; Bertram Müller-Myhsok, MD; Heribert Schunkert, MD; Inke R. König, PhD

BACKGROUND: Individual risk prediction based on genome-wide polygenic risk scores (PRSs) using millions of genetic variants has attracted much attention. It is under debate whether PRS models can be applied—without loss of precision—to populations of similar ethnic but different geographic background than the one the scores were trained on. Here, we examine how PRS trained in population-specific but European data sets perform in other European subpopulations in distinguishing between coronary artery disease patients and healthy individuals.

METHODS: We use data from UK and Estonian biobanks (UKB, EB) as well as case-control data from the German population (DE) to develop and evaluate PRS in the same and different populations.

RESULTS: PRSs have the highest performance in their corresponding population testing data sets, whereas their performance significantly drops if applied to testing data sets from different European populations. Models trained on DE data revealed area under the curves in independent testing sets in DE: 0.6752, EB: 0.6156, and UKB: 0.5989; trained on EB and tested on EB: 0.6565, DE: 0.5407, and UKB: 0.6043; trained on UKB and tested on UKB: 0.6133, DE: 0.5143, and EB: 0.6049.

CONCLUSIONS: This result has a direct impact on the clinical usability of PRS for risk prediction models using PRS: a population effect must be kept in mind when applying risk estimation models, which are based on additional genetic information even for individuals from different European populations of the same ethnicity.

Key Words: association ■ coronary artery disease ■ genetic association studies ■ genetics ■ genome ■ population ■ risk factors

There is growing evidence that polygenic risk scores (PRSs) can be applied clinically to improve prediction of individual disease risks.[1] While earlier work on genetic risk scores (GRSs) was based on few variants with genome-wide significant signals for association,[2,3] more recent models included thousands or even millions of genetic variants, which further improved prediction of the risk of coronary artery disease (CAD) and other conditions.[4–6]

Common to all GRSs is that the model summarizes the number of risk alleles, weighted by the estimated effects of risk alleles derived from genome-wide association studies. The prediction quality of the score, possibly together with sex, age, and other (clinical) variables, is investigated in a training data set, and the optimum significance threshold, and thus the number of genetic variants used, is selected on the basis of the best performance. Through this, the genomic information of thousands or millions of genetic variants distributed throughout the genome may be compressed into a single value, the (P)GRS. It has been argued that thus summarizing the genetic risk is too simple given the complex biological structure of common diseases. However, we recently found that using PRS is more appropriate than using more complex alternatives with common and widely used machine learning algorithms.[7] In any case, the generalizability of any GRS then needs to be verified in another independent test data set.[8]

Correspondence to: Inke R. König, Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany. Email inke.koenig@uni-luebeck.de

The Data Supplement is available at https://www.ahajournals.org/doi/suppl/10.1161/CIRCGEN.120.002932.

For Sources of Funding and Disclosures, see page 575.

Circulation: Genomic and Precision Medicine is available at www.ahajournals.org/journal/circgen

Nonstandard Abbreviations and Acronyms

AUC	area under the curve
AUCPR	area under the precision-recall (PR) curve
CAD	coronary artery disease
EB	Estonian biobank
GRS	genetic risk score
PRS	polygenic risk score
UKB	UK Biobank

Inouye et al[4] proposed a meta-analytic-based approach to construct a PRS to predict the risk of CAD using 1.7 million genetic variants. Similarly, Khera et al[5] introduced a PRS for the risk prediction of 5 common diseases including CAD with 6.6 million genetic variants. Both studies used effect estimates from the CARDIoGRAMplusC4D genome-wide association studies meta-analysis,[9] and the prediction models were trained on a subset of the UK Biobank (UKB).[10] Khera et al[5] additionally added age, sex, the first 4 principal components and an indicator variable for the genotyping array in their model, whereas Inouye et al[4] used the meta PRS by itself. In independent UKB data sets, good predictive performances were replicated (Inouye et al[4]: Harrell C 0.623 [95% CI, 0.615–0.631]; Khera et al[5]: area under the curve [AUC] 0.81 [95% CI, 0.80–0.81]). Risk prediction by PRS was more accurate than that of conventional risk factors, leading the authors to conclude that an individual's genetic risk of common diseases at birth is predictable and would enable effective prevention or detection strategies.

As a caveat, both author groups pointed out that the proposed PRS were studied largely in individuals of European descent and cannot readily be applied to other ethnic groups without taking into account the target population's structure.[11] Moreover, it is as yet unknown whether the performance of a PRS depends not only on ethnicity but also on smaller genomic differences within a population. In this work, we thus studied the discriminative ability of PRS in data from the UKB and Estonian Biobank (EB)[12–14] as well as data from the German population to test how PRS trained in one European data set perform in other European data sets.

METHODS

The detailed methods of this work are available as Data Supplement. All included studies were approved by an institutional review committee, and all subjects gave informed consent. Information on the availability of the data that support the findings of this study is available from the corresponding authors of the respective references given in Table I in the Data Supplement. Restrictions apply to the availability of these data, which were used under license for this study. The result data that support the findings of this study are available from the corresponding author upon reasonable request. The code used and the trained PRS models are available at https://github.com/dagola/GO-3269-1-1_code.

RESULTS

The PRS in the UKB and EB data sets D_{UKB} and D_{EB} were based on the published imputed genotypes.[10,13]

Table 1. Hyperparameter Search Space and Optimal Hyperparameter Settings Found in 10-Fold Cross Validation

Parameter	Type	Possible values	Requires	Model trained on				
				UKB	EB	DE	DE2	Combined
Min. MAF in summary statistics (summary.statistics.maf.thresholds)	Numeric	0–0.1		4.83^{-02}	3.18^{-02}	4.34^{-05}	9.97^{-02}	6.26^{-02}
Nonmissing genotypes in training data set (target.geno)	Numeric	0.9–1		0.99	1.00	0.95	0.93	1.00
Min. MAF in training data set (target.maf)	Numeric	0–0.1		7.57^{-02}	8.52^{-02}	8.82^{-02}	4.77^{-02}	7.74^{-02}
Clumping (clumping)	Logical			True	True	False	False	False
LD information from external data set (ld.external)	Logical		clumping == true	False	False			
Min. MAF in external LD data set (ld.maf)	Numeric	0–0.1	ld.external == true					
Nonmissing genotypes in external LD data set (ld.geno)	Numeric	0.9–1	ld.external == true					
Clumping distance [kb] (clumping.kb)	Integer	$125–5\times10^{+03}$	clumping == true	1366	1058			
Clumping r^2 threshold (clumping.r2)	Numeric	0.1–0.8	clumping == true	0.21	0.45	NA	NA	NA
P value upper bound (pval.level)	Numeric	$5\times10^{-08}–1$		1.23^{-03}	0.39	0.97	0.58	0.27
Handling of missing genotypes (missing.handling)	Discrete	IMPUTE, SET_ZERO, CENTER		CENTER	SET_ZERO	SET_ZERO	SET_ZERO	CENTER

Combinded indicates combined training data set from UKB, EB, and DE; DE, training data set from German case/control data sets; DE2, training data set from German case/control data set with reduced cases faction to match UKB and EB training data sets; EB, training data set from Estonian Biobank; and UKB, training data set from UK Biobank.

Moreover, we combined six imputed CAD genome-wide association studies from the German population D_{DE} (Table I in the Data Supplement). Randomly chosen subsets of 10 000 individuals each were used for training and the remaining samples as corresponding testing data sets. Since D_{DE} is a case/control data set including smaller numbers of controls as compared to the population-based D_{UKB} and D_{EB}, we created an additional data set D_{DE2} in which the number of cases is ≈3% as in D_{UKB} and D_{EB}. Due to the relative low number of controls in D_{DE}, D_{DE2} has a total sample size of 7594 and includes all available controls of D_{DE}. Finally, to test the precision of a PRS trained on a mixed population data set, a combined training data set D_{COMB} was used. This included 10 000 individuals with equal numbers of samples from the different population-specific training data sets D_{DE2}, D_{UKB}, and D_{EB} while maintaining the population-specific prevalences. For every PRS, we optimized the hyperparameters (Table 1) in terms of the area under the precision-recall (PR) curve (AUCPR).

Table 2 (upper part) shows that the area under the ROC curve (AUC) in each testing data set is highest when based on the respective model trained in the corresponding population. On the 5% level, these are also significantly better than the respective second-best models, that is, those trained in another European population (DE test data set: DE versus EB: ΔAUC=0.1345 [95% CI, 0.1108–0.1581], $P<2.2\times10^{-16}$; EB test data set: DE versus EB: ΔAUC=0.0409 [95% CI, 0.0238–0.0579], $P=2.729\times10^{-06}$; UKB test data set: EB versus UKB: ΔAUC=0.009 [95% CI, 0.0047–0.0134], $P=4.522\times10^{-05}$). The PRS model trained on D_{DE2} has slightly but not significantly better performance on the EB and UKB testing data sets compared with the model trained on D_{DE}. The PRS model trained on D_{COMB} performs at least as good as the worst population-specific model with a very consistent AUC of about 0.6 in all testing data sets. The performance of the PRS proposed by Khera et al[5] and Inouye et al[4] on our testing data sets is added for comparison purposes. Their performances are technically the best on the UKB testing data set (ΔAUC=0.6374 and ΔAUC=0.6377). However, the samples in our testing data set might have an overlap with those used to train the model by Khera et al[5] and Inouye et al,[4] and thus are not unbiased estimates. On D_{DE}^{test} the PRS model by Khera et al[5] achieves second-best performance (AUC=0.6699), not significantly worse than that of $M_{h^*}^{DE}$, and better than the performance on D_{UKB}^{test}, whereas on D_{EB}^{test} its performance is worst (AUC=0.5617). However, the PRS model by Inouye et al[4] achieves best performance on D_{EB}^{test} and worst performance on D_{DE}^{test}. Similar results are obtained for the AUCPR. Comparing the distribution of models developed and tested in the UKB and EB data as shown in Figure 1 indicates that there are notable shifts between the different populations. Specifically, many to almost all samples from the UKB data with the highest scores have lower scores than the majority of samples from the EB data. This is also reflected by the estimated CAD prevalence in 100 groups defined by the score percentiles of each population-specific PRS model as

Table 2. Discrimination Performance (95% CI) of Models Trained on Training Data Sets From Different Populations in Population-Specific Testing Data Sets in Terms of AUC and AUCPR

Performance statistic	Model trained on (no. of SNPS)	Model evaluated on		
		DE	EB	UKB
AUC	UKB (1940)	0.5143 (0.4992–0.5294)	0.6049 (0.5857–0.6241)	0.6133 (0.6094–0.6172)*
	EB (375 822)	0.5407 (0.5253–0.5561)	0.6565 (0.6369–0.6760)*	0.6043 (0.6004–0.6082)
	DE (3 423 987)	0.6752 (0.6612–0.6891)*	0.6156 (0.5963–0.6349)	0.5989 (0.5950–0.6028)
	Combined (1 056 021)		0.6112 (0.5919–0.6305)	0.5988 (0.5949–0.6027)
	DE2 (2 490 815)		0.6212 (0.6018–0.6406)	0.6011 (0.5972–0.6050)
	Khera et al[5] (6 630 150)	0.6699 (0.6557–0.6840)	0.5617 (0.5402–0.5833)	0.6374 (0.6335–0.6412)
	Inouye et al[4] (1 745 179)	0.5015 (0.4830–0.5140)	0.6597 (0.6405–0.6789)	0.6377 (0.6339–0.6416)
AUCPR	UKB (1940)	0.5607 (0.5593–0.5621)	0.0460 (0.0454–0.0466)	0.0752 (0.0745–0.0760)*
	EB (375 822)	0.4980 (0.4962–0.4998)	0.0765 (0.0755–0.0774)*	0.0712 (0.0703–0.0721)
	DE (3 423 987)	0.6891 (0.6887–0.6895)*	0.0506 (0.0504–0.0508)	0.0696 (0.0694–0.0698)
	Combined (1 056 021)		0.0480 (0.0473–0.0487)	0.0697 (0.0688–0.0705)
	DE2 (2 490 815)		0.0521 (0.0512–0.0530)	0.0705 (0.0695–0.0716)
	Khera et al[5] (2 490 815)	0.6609 (0.6605–0.6613)	0.0446 (0.0444–0.0448)	0.0837 (0.0835–0.0840)
	Inouye et al[4] (1 745 179)	0.5205 (0.5201–0.5210)	0.0673 (0.0668–0.0679)	0.0832 (0.0830–0.0835)

The AUCPR of a random model equals 0.5230 (DE), 0.0311 (EB), and 0.0487 (UKB). AUC indicates area under the receiver operating characteristic curve; AUCPR, area under the recall-precision curve; Combinded, combined training data set from UKB, EB and DE; DE, training data set from German case/control data sets; DE2, training data set from German case/control data set with reduced cases faction to match UKB and EB training data sets; EB, training data set from Estonian Biobank; and UKB, training data set from UK Biobank.

*The best model developed in this work per testing data set.

shown in Figure 2. Like Khera et al,[5] we binned individuals into 100 groupings according to the percentile of the GRS, and the unadjusted prevalence of disease within each bin was determined. Here, one would expect higher prevalence of CAD with increasing scores. This is generally true for the PRS evaluated on EB and UKB testing data sets. However, applying the population-specific PRS on other population testing data sets results in inconsistent CAD prevalences, especially at the tails. For example, the models $M_{h^*}^{DE}$ and $M_{h^*}^{UKB}$ evaluated on D_{EB}^{test} have too high prevalences in the lower percentile groups and too low prevalences in the high percentile groups compared with those of $M_{h^*}^{EB}$, that is, the extreme scores of nonmatching population-specific PRS do not reflect the subpopulations of very low or high risk. Here, the performance of the PRS models $M_{h^*}^{EB}$ and $M_{h^*}^{UKB}$ on D_{DE}^{test} (Figure 2, left) are of special note as the estimated prevalences are completely inconsistent.

Given the notable difference in number of SNPs used in each PRS (Table II in the Data Supplement), we additionally compared the performances when fixing the number of SNPs at 2213 genome-wide significant SNPs ($P<5\times10^{-8}$), the top 3000, 30 000, and 3 000 000 SNPs, respectively. Again, population-specific PRS yielded the highest performances, with the differences even slightly increasing with increasing numbers of SNPs (Figure 3).

DISCUSSION

We assessed the impact of population-specific data sets of European ancestry on the discriminative performance of PRS and revealed a substantial and clinically relevant drop in performance if training and testing data sets came from different populations. A PRS trained on the combined training data sets performed better than population-specific PRS applied to a different population while being less informative than a population-specific PRS. Importantly, in each of the 3 European populations tested the by far best performance was achieved if the training and testing data set came from the same population. Mimicking the population prevalence in a case/control data set as done for the PRS model trained on D_{DE2} did not substantially improve the performance on the testing data sets from different

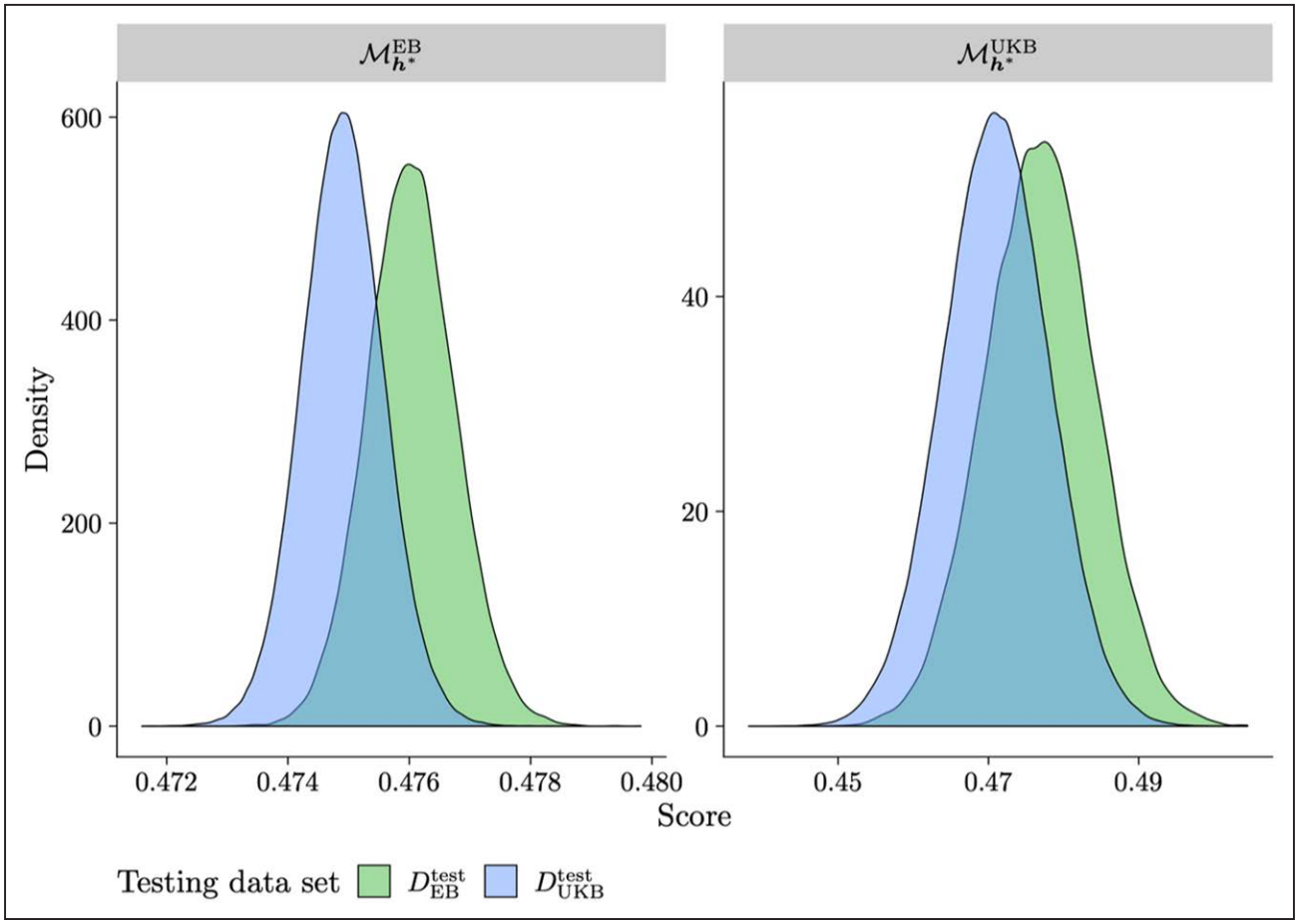

Figure 1. Densities of polygenic risk scores in population-specific testing data sets of models trained on Estonian Biobank (EB) and UK Biobank (UKB) specific data sets.
Scores have been transformed to the interval (0–1), where 0 is the minimum score and 1 is the maximum score to get from each model. Please note the different axis scales.

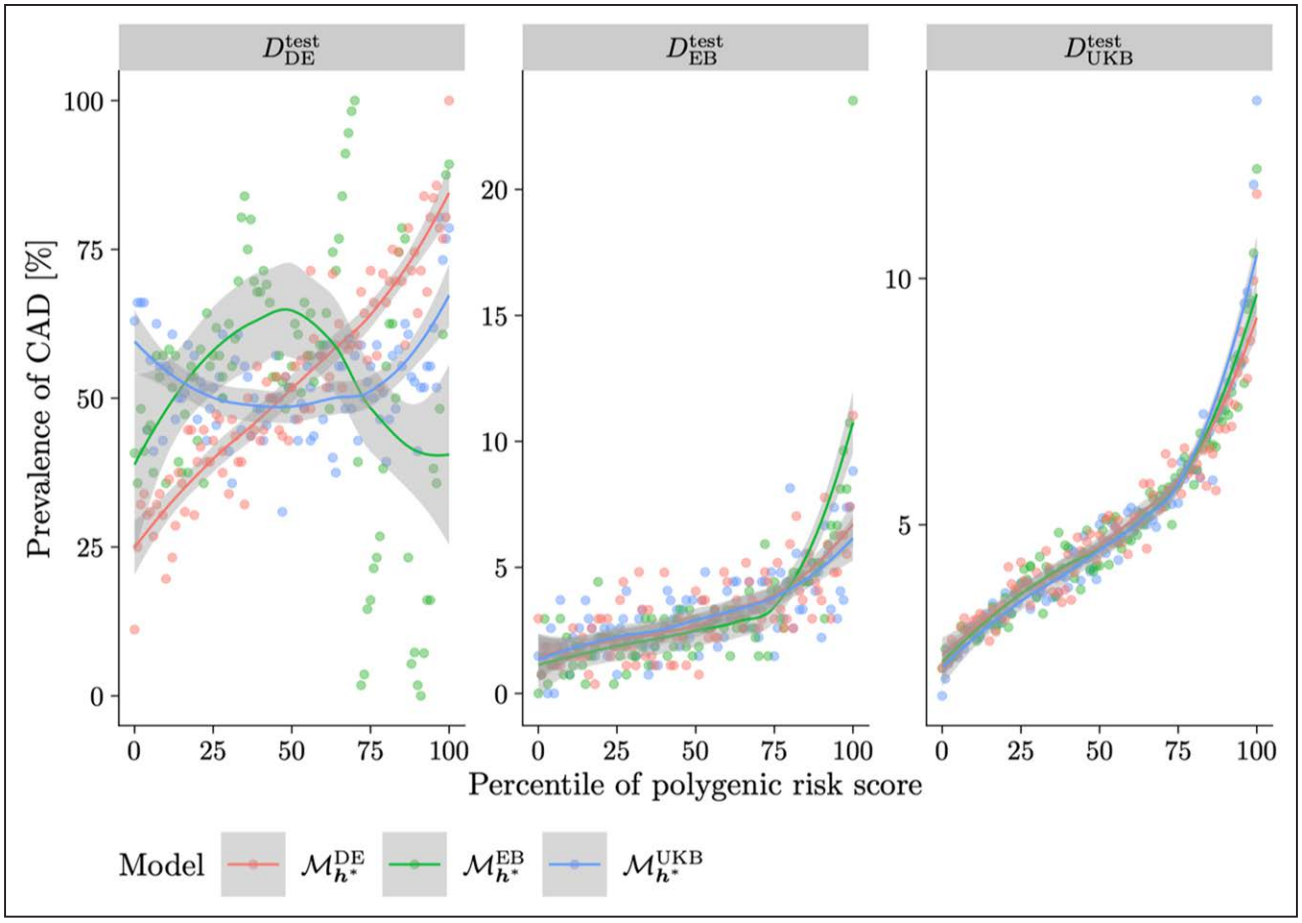

Figure 2. Prevalence of coronary artery disease (CAD) according to 100 groups of the population-specific testing data sets binned according to the percentiles of each population-specific polygenic risk scores (PRS) models.
A loess smoothing function is added for each model to aid the eye. EB indicates Estonian Biobank; and UKB, UK Biobank.

populations. Interestingly, the Khera et al model also performed well on the DE and the Inouye et al[4] model well on the EB testing data sets, whereas vice versa the predictive values were weaker. While we cannot explain this variability in the data, it is interesting to note that both scores proposed by Khera et al[5] and Inouye et al[4] performed best on the UKB testing data, which substantiates our principle findings as it indicates that the population bias effects these scores as well.

Special interest should be payed to models trained on population-based data sets and applied to case/control data sets shown in Figure 2, left. As the estimated prevalences from PRS models $M_{h^*}^{EB}$ and $M_{h^*}^{UKB}$ on D_{DE}^{test} are inconsistent in contrast to those of $M_{h^*}^{DE}$ applied on D_{EB}^{test} and D_{UKB}^{test}. It appears that training of PRS models on case/control data sets and application on population-based data sets is valid in terms of consistently estimated prevalences. This is likely due to the upscaled fraction of cases in the case/control data set allowing for a better discrimination of cases and controls in any data sets. In contrast, models trained on population-based data sets with a comparatively low fraction of cases do not easily generalize to target data sets with higher fractions of cases. In this case, these models may be too sensitive to detect cases.

It should be noted that subsets of the German and Estonian data sets were part of the CARDIoGRAMplusC4D meta-analysis and thus contributed to the summary statistics used to weight the single SNP contributions in the PRS. Therefore, AUC estimates in the testing data sets from these 2 populations might be inflated.[15]

The decreased discrimination performance and shift of scores of population-specific PRS in different populations has direct impact on the clinical utility of risk prediction models by PRS. As scores can be generally lower or higher when applied to samples from other populations than those used for training of the models, estimated risks will also be biased for individuals seeking their personal risk but not matching the population used to derive the PRS and risk prediction models.

Thus, genomic differences between populations must be considered when applying risk estimation models. Importantly, we have shown that this is not only true for individuals from different ethnicities but also for individuals from different populations of the same ethnicity. It is

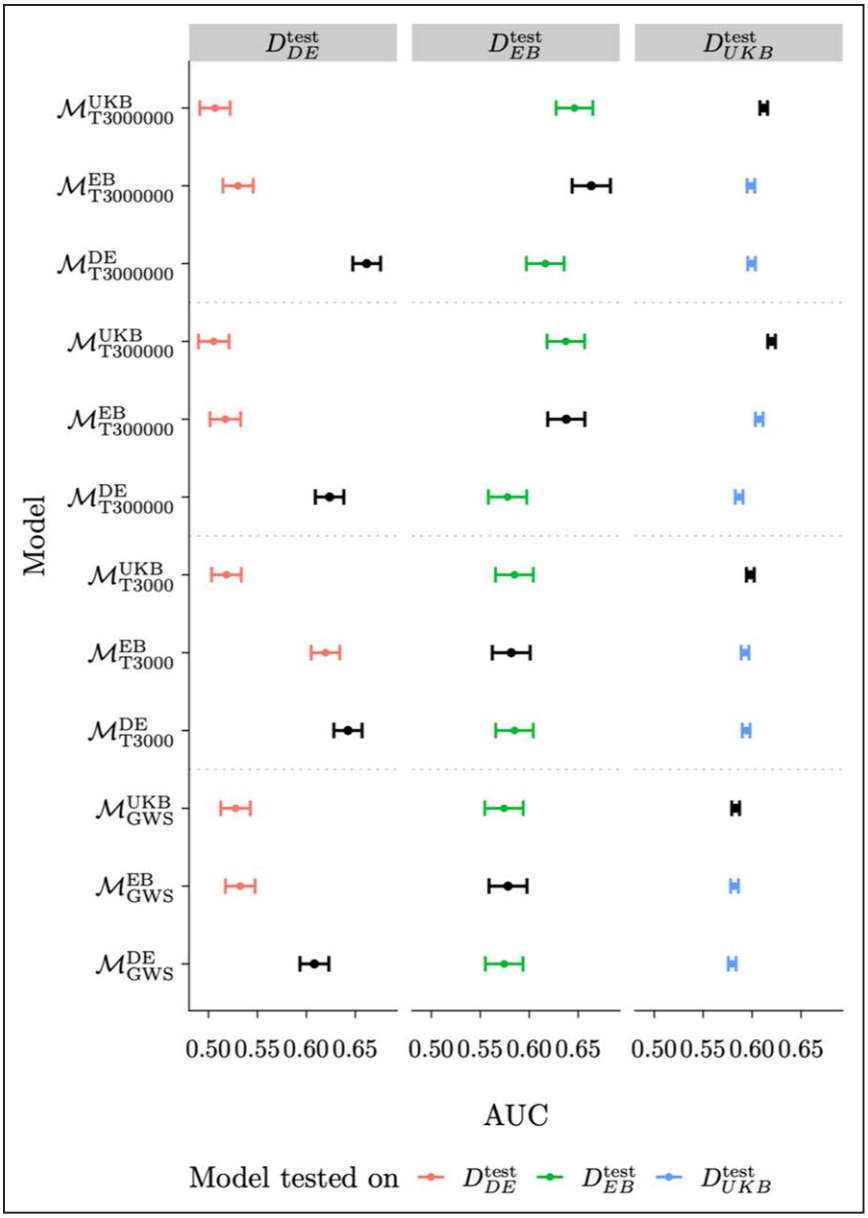

Figure 3. Area under the ROC curve (AUC) in population-specific testing data sets of restricted population-specific polygenic genome-wide risk scores models.
Dots denote the estimated AUC, and error bars denote the 95% CIs. Models were trained on the specific populations (DE, Estonian Biobank [EB], or UK Biobank [UKB]). The number of SNPs was fixed to genome-wide significant ($P<5\times10^{-8}$), or those with lowest 3000, 30 000, or 3 000 000 P values. Black highlights the performance on testing data sets of models trained on concordant training data sets.

in particular important as the advent of huge biobank data sets tempts to use samples from one biobank only to derive PRS, train risk prediction models and test and validate those models. However, it must be kept in mind that these models may be applicable only to those individuals matching the population structure of the samples in these biobank data sets. Using a mixture of different populations may reduce this bias but will simultaneously also reduce the performance for individuals from the same population. Here, more advanced methods will be needed to maximize the benefit for all. Until then, our conclusion is that each and every single population PRS and population-specific risk estimation model enhanced by PRS will have to be derived on their very own training data set or at least verified for application on the target population, even if PRS or risk models trained in other populations of the same ethnicity are available.

ARTICLE INFORMATION

Received January 23, 2020; accepted October 26, 2020.

Affiliations

Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck (D.G., I.R.K.). German Centre for Cardiovascular Research, Partner Site Hamburg/Kiel/Lübeck (D.G., J.E., I.R.K.). Institute for Cardiogenetics, Universität zu Lübeck, Germany (J.E.). Estonian Genome Centre, Institute of Genomics, University of Tartu, Estonia (K.L., R.M.). Department of Translational Research in Psychiatry, Max Planck Institute of Psychiatry, Munich Cluster of Systems Neurology, SyNergy, Germany (B.M.-M.). Institute of Translational Medicine, University of Liverpool, United Kingdom (B.M.-M.). Deutsches Herzzentrum München, Technische Universität München, German Centre for Cardiovascular Research, Partner Site München, Germany (H.S.).

Acknowledgments

This research has been conducted using the UK Biobank Resource under application number 48012 and the Estonian Biobank Resource, and we thank all participants providing their data for research purposes. Data analyses were performed in part using the High-Performance Computing Center of University of Tartu.

Sources of Funding

This work was supported by a grant of the Cluster of Excellence Inflammation at Interfaces, funded by the German Research Foundation to I.R. König, and by a research fellowship grant by the German Research Foundation (GO 3269/1-1) to D. Gola. Drs Mägi and Läll were supported by Estonian Research Council grant PUT PRG687 and institutional grant PP1GI19935 from Institute of Genomics, University of Tartu.

Disclosures

None.

REFERENCES

1. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet*. 2018;19:581–590. doi: 10.1038/s41576-018-0018-x
2. Davies RW, Dandona S, Stewart AF, Chen L, Ellis SG, Tang WH, Hazen SL, Roberts R, McPherson R, Wells GA. Improved prediction of cardiovascular disease based on a panel of single nucleotide polymorphisms identified through genome-wide association studies. *Circ Cardiovasc Genet*. 2010;3:468–474. doi: 10.1161/CIRCGENETICS.110.946269
3. Morrison AC, Bare LA, Chambless LE, Ellis SG, Malloy M, Kane JP, Pankow JS, Devlin JJ, Willerson JT, Boerwinkle E. Prediction of coronary heart disease risk using a genetic risk score: the Atherosclerosis Risk in Communities Study. *Am J Epidemiol*. 2007;166:28–35. doi: 10.1093/aje/kwm060
4. Inouye M, Abraham G, Nelson CP, Wood AM, Sweeting MJ, Dudbridge F, Lai FY, Kaptoge S, Brozynska M, Wang T, et al; UK Biobank CardioMetabolic Consortium CHD Working Group. Genomic risk prediction of coronary artery disease in 480,000 adults: implications for primary prevention. *J Am Coll Cardiol*. 2018;72:1883–1893. doi: 10.1016/j.jacc.2018.07.079
5. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018;50:1219–1224. doi: 10.1038/s41588-018-0183-z
6. Khera AV, Chaffin M, Wade KH, Zahid S, Brancale J, Xia R, Distefano M, Senol-Cosar O, Haas ME, Bick A, et al. Polygenic prediction of weight and obesity trajectories from birth to adulthood. *Cell*. 2019;177:587–596.e9. doi: 10.1016/j.cell.2019.03.028
7. Gola D, Erdmann J, Müller-Myhsok B, Schunkert H, König IR. Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status. *Genet Epidemiol*. 2020;44:125–138. doi: 10.1002/gepi.22279
8. Igl BW, Konig IR, Ziegler A. What do we mean by 'replication' and 'validation' in genome-wide association studies? *Hum Hered*. 2009;67:66–68. doi: 10.1159/000164400
9. Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, Saleheen D, Kyriakou T, Nelson CP, Hopewell JC, et al. A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet*. 2015;47:1121–1130. doi: 10.1038/ng.3396
10. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12:e1001779. doi: 10.1371/journal.pmed.1001779
11. Reisberg S, Iljasenko T, Läll K, Fischer K, Vilo J. Comparing distributions of polygenic risk scores of type 2 diabetes and coronary heart disease within different populations. *PLoS One*. 2017;12:e0179238. doi: 10.1371/journal.pone.0179238
12. Köhler F, Laschinski G, Ganten D. Das estnische Genomprojekt im Kontext der europäischen Genomforschung. *Dtsch Med Wochenschr*. 2004;129:S25–S28.
13. Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, Snieder H, Perola M, Ng PC, Mägi R, Milani L, et al. Cohort profile: estonian biobank of the Estonian Genome Center, University of Tartu. *Int J Epidemiol*. 2015;44:1137–1147. doi: 10.1093/ije/dyt268
14. Metspalu A. The estonian genome project. *Drug Dev Res*. 2004;62:97–101.
15. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet*. 2013;14:507–515. doi: 10.1038/nrg3457