Polygenic Risk Scores for Prediction of Breast Cancer Subtypes

Executive Summary

Polygenic Risk Scores (PRS), a combination of breast cancer susceptibility genetic variants that are common in the population, has been shown to achieve useful level of breast cancer risk stratification in women of Asian ancestry. However, implementation of PRS is hindered by the difference in PRS distribution across ethnic groups. The goal of this project is to calibrate existing breast cancer PRS for Asian population so that a single PRS distribution can be applicable to different ethnic groups for risk stratification. We showed that adjusting for population structure can remove heterogeneity in PRS distribution across ethnicity.

Introduction

In European countries, breast cancer screening is systematic and widely implemented. This improves likelihood of breast cancer survival due to early detection of the disease. Unlike in European countries, breast cancer screening in many Asian countries is opportunistic and has poor uptake. This contributes to delayed detection and hence poor survival rates for cancer patients. With the increase in breast cancer incidence in Asia, there is an urgent need for better breast cancer control strategy to tackle the rising burden [1]. A feasible breast cancer control strategy is risk-stratified approach, where screening is offered to women at higher risk of developing breast cancer to enable early detection of breast cancer.

Polygenic Risk Score (PRS), a measure of the risk a person has to a disease based on their genes, has been shown to achieve useful level of risk stratification [1, 2, 3]. For example, based on a PRS developed for women of European ancestry, women in the top 1% of PRS distribution had 29% risk of developing breast cancer by the age of 80 while women in the bottom 1% had only 3.5% risk [2]. *Ho et al* evaluated and improved the performance of this European-based PRS [1]. The team showed that this PRS was predictive of breast cancer risk and can be useful for risk stratification in Asian women. However, *Ho* & *colleagues* demonstrated that PRS distribution varies across ethnicities in Malaysia, hence complicates its clinical implementation which relies on the distribution of PRS to stratify women into different risk groups. Therefore, the aim of this project is to determine way in which we can calibrate the PRS distribution so that one distribution can be applicable across different ethnicity.

Problem Statements

Although breast cancer PRS has been shown to be predictive of disease risk across Asian ethnic subgroups, the distribution of PRS has shown to be different on different ethnic groups [3].

Objectives

The objective is to identify a method to remove the heterogeneity of the means of PRS distribution across Asian ethnic groups.

Hypothesis

We hypothesise that the heterogeneity in the means of the PRS distribution can be removed by one of the three methods tested. The first method is limiting the construction of PRS using SNPs with low variation in minor allelic frequencies across ethnic groups. The second method is to account for the mean differences in allelic frequencies for all SNPs across ethnic groups. The third method is to account for the population structure using principal components.

Methodology

Study population

Study participants were recruited into the Malaysian Breast Cancer Genetics Study (MyBrCa) and the Singapore Breast Cancer Cohort study (SGBCC). These studies comprised of 5236 cases and 5156 controls of Chinese ancestry, 1084 cases and 1332 controls of Malay ancestry and 580 cases and 1018 controls of Indian ancestry. The genotype data were available for all women in the study. All women provided informed consent [3].

Polygenic risk scores

A PRS score for individual *i* was calculated for all of the participants with the formula:

$$PRS_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}$$

where x_k represents the allele dosages for Single Nucleoid Polymorphism (SNP) k and β_k is the per-allele log odds ratio of SNP *k* on breast cancer risk. We considered two PRSs developed previously for Asian population: [1] one based on 287 SNPs (PRS-287) and one based on 1 million SNPs (PRS-CSX), where β_k reported in *Ho et al 2022* was used to construct PRSs.

Heterogeneity among ethnicities

Heterogeneity of PRS distributions among ethnicities was calculated using. I^2, where

$$I^2 = \frac{Q-(K-1)}{Q}$$

$$Q = \sum_{k=1}^{K} w_k\left(\hat{\theta}_k - \hat{\theta}\right)^2.$$

The variable Q is also known as Cochran's Q [5], K=3 represents the number of groups being compared, $\mathrm{w_k} = \frac{\mathrm{n}}{\mathrm{s}^2}$ represents the inverse of the variance, $\hat{\theta}_k$ is the PRS mean of the kth ethnic group while $\hat{\theta}$ is the global mean of PRS. If the resulting I^2 is negative, it will be taken as 0% meaning there is no evidence of heterogeneity. Higher than 75% means there is evidence of substantial heterogeneity, higher than 50% is moderate heterogeneity and higher than 25% is low heterogeneity [5].

Adjustment of heterogeneity by constructing PRS using SNPs with similar allele frequencies across ethnic groups

The risk allele frequency (RAF) was calculated for each SNP and for each ethnic group with the formula:

$$RAF_{Chinese} = \frac{\Sigma \mathrm{x}_{\mathrm{Chinese}}}{2*10384}$$

$$RAF_{Malay} = \frac{\Sigma \mathrm{x}_{\mathrm{Malay}}}{2*2396}$$

$$\mathrm{R}AF_{Indian} = \frac{\Sigma \mathrm{x}_{\mathrm{Indian}}}{2*1602}$$

Σx represents the sum of the allele dosages for all the individuals in an ethnic group for one SNP with decimals rounded to the nearest whole number. The sum is divided by 2 times the number of participants in the ethnic group to get the RAF for that group in that SNP. A chi-square test for independence is performed on the allele frequency rounded to the nearest whole number for each SNP to determine if the allele frequency is independent of the ethnic group and to attain the p-value.

H_0: The allele frequency is independent on the ethnic group

H_1: The allele frequency is not independent on the ethnic group

SNPs that are not significant at a pre-specified p-value will be used for the calculation of the new PRS and an I^2 test will be used to test the heterogeneity of this new PRS across ethnic groups. The p-value thresholds between 10^{-25} and 0.1 and were considered.

Adjusting for heterogeneity by accounting for the mean difference in RAF across ethnic groups

The difference in RAF between Indian/Malay participants with Chinese (the reference group) was calculated for each SNP. The mean difference in the RAF across all SNPs was subtracted from the mean of PRS distribution constructed using Indian/Malay participants.

Only PRS-287 was considered for adjustment methods that involved RAF. Raw data of PRS-CSX, which involved 1 million SNPs, were not available for this project.

Principal Component Calculation

Principal Components (PC) reported in *Ho et al* was used for this part of the analysis.

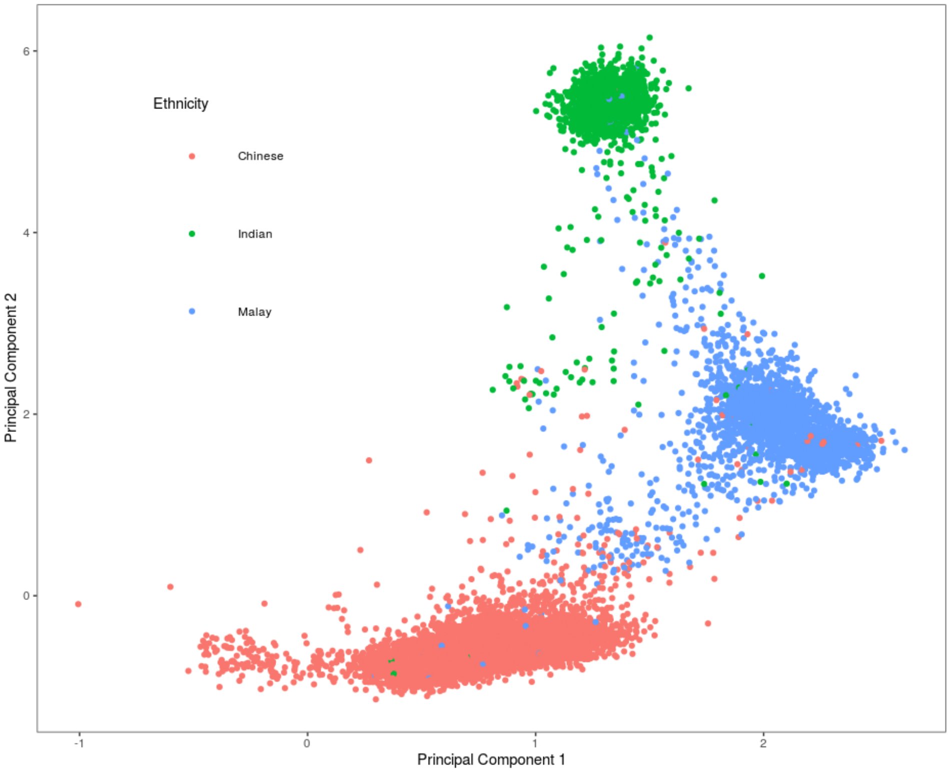

Figure 1: Scatter plot of the first 2 principal components

Based on the PC graph in Figure 1, PC captured the difference in ancestry across Chinese, Malay and Indian ethnicities.

A linear regression model of PRS against principal components is fitted by progressively adding principal components starting from PC1 to PC10. The residual PRS are calculated by taking the difference between the actual PRS value and the predicted PRS values for each data sample using each of the models. The mean PRS residuals are then calculated by ethnic group and tested for heterogeneity using method describe above.

The programming language R was used to perform all the analysis and visualisation of the data and results.

Results & discussion

The data used for the analysis consisted of 14382 participants of Chinese, Malay and Indian ethnicity.

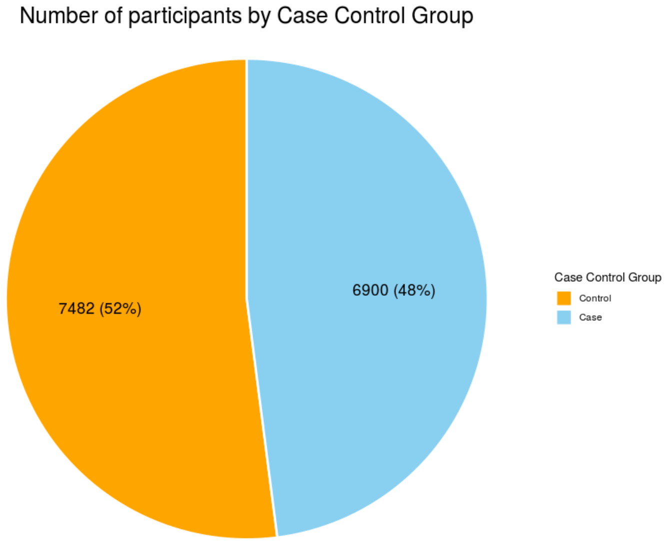

Figure 2: Pie Chart for number of participants by case-control group

As shown in Figure 2, there are 7482 controls and 6900 cases in the study.

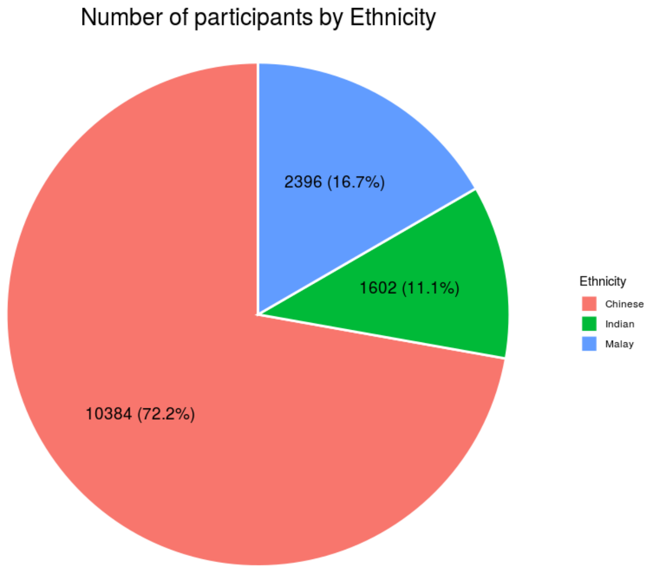

Figure 3: Pie Chart for number of participants by ethnicity

Figure 3 shows there that there are 10384 Chinese, 2396 Malay and 1602 Indian participants. Chinese ethnicity makes up most of the data with 72.2%. Two models of PRS were used for the calculations; PRS287 which uses 287 SNPs and PRSCSX which uses over a million SNPs.

Heterogeneity in mean of PRSs

Figure 4 shows the distribution of PRS287 developed by *Ho et al* across the three major ethnic groups in Malaysia.

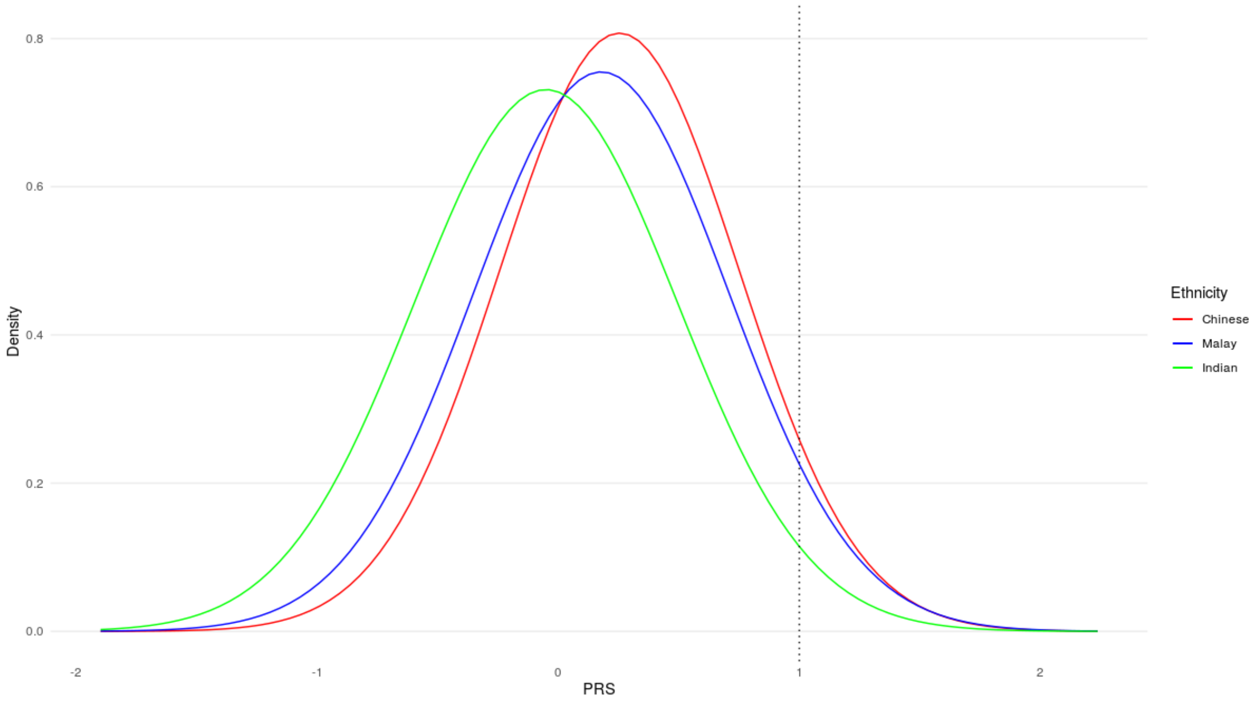

Figure 4: Normal distribution of PRS287 by ethnicity with a dashed line at PRS 1

The differences in the normal distribution curve indicate deviation in the percentile individuals should be assigned to even with the same PRS. For example, as indicated by the dashed line in Figure 4, a score of 1 would be associated with different percentiles, and hence different risk, across different ethnic groups.

Ethnicity	Mean	Standard deviation
Chinese	0.25400	0.494
Malay	0.17900	0.528
Indian	-0.05080	0.546

Table 1: Mean and Standard deviation of PRS287 by ethnicity

Ethnicity	Mean	Standard deviation
Chinese	0.00110	0.493
Malay	-0.19300	0.510
Indian	-0.45900	0.514

Table 2: Mean and Standard deviation of PRSCSX by ethnicity

In Table 1 and 2, both PRS show similar standard deviation across different ethnicities but the means are markedly different, where Chinese has the highest mean while Indian has the lowest.

PRS	I^2 (%)
PRSCSX	99.84239
PRS287	99.56353

Table 3: I^2 value of PRSCSX and PRS287

The I^2 value for both PRS is higher than 75% as shown in Table 1. Therefore there is evidence of strong heterogeneity in PRS distribution among the three ethnic groups.

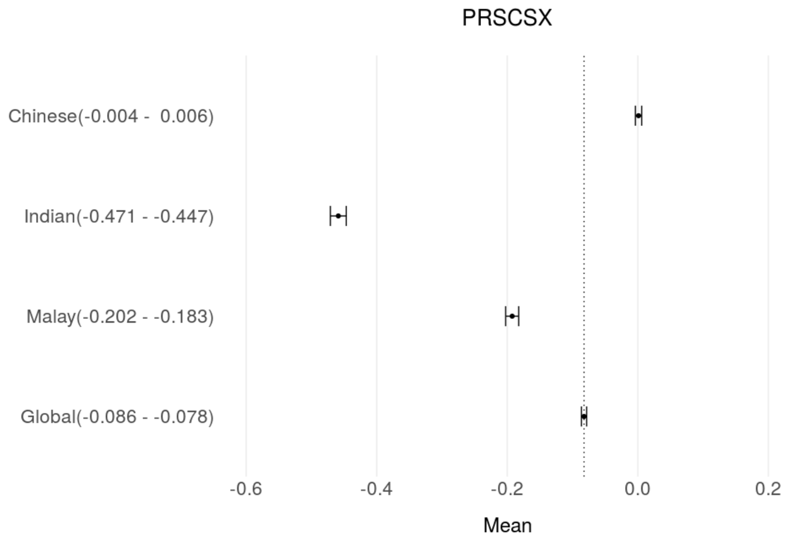

Figure 5: Forest Plot for PRCSX

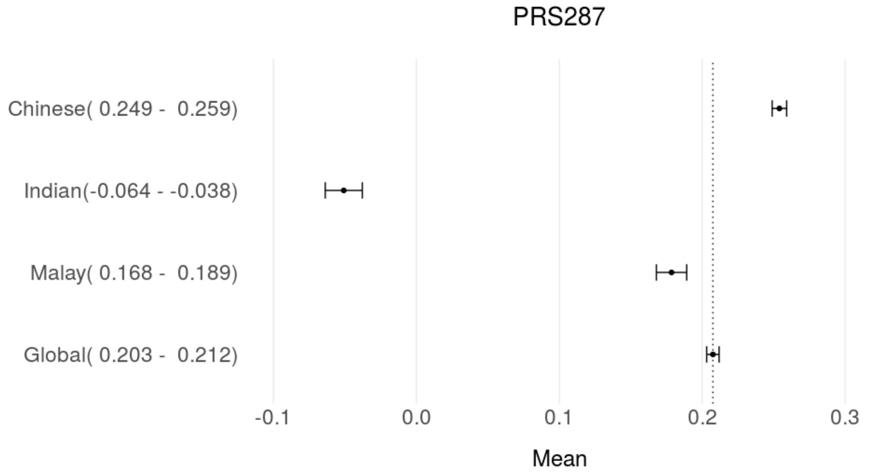

Figure 6: Forest Plot for PRS287

Figure 5 and Figure 6 both show that Malay and Chinese PRS means are closer to the global mean while Indian PRS mean is further away. This observation is expected as the majority of the participants are of Chinese/Malay ethnic groups.

Limiting PRS construction to SNPs with low variation in allele frequency across ethnic groups

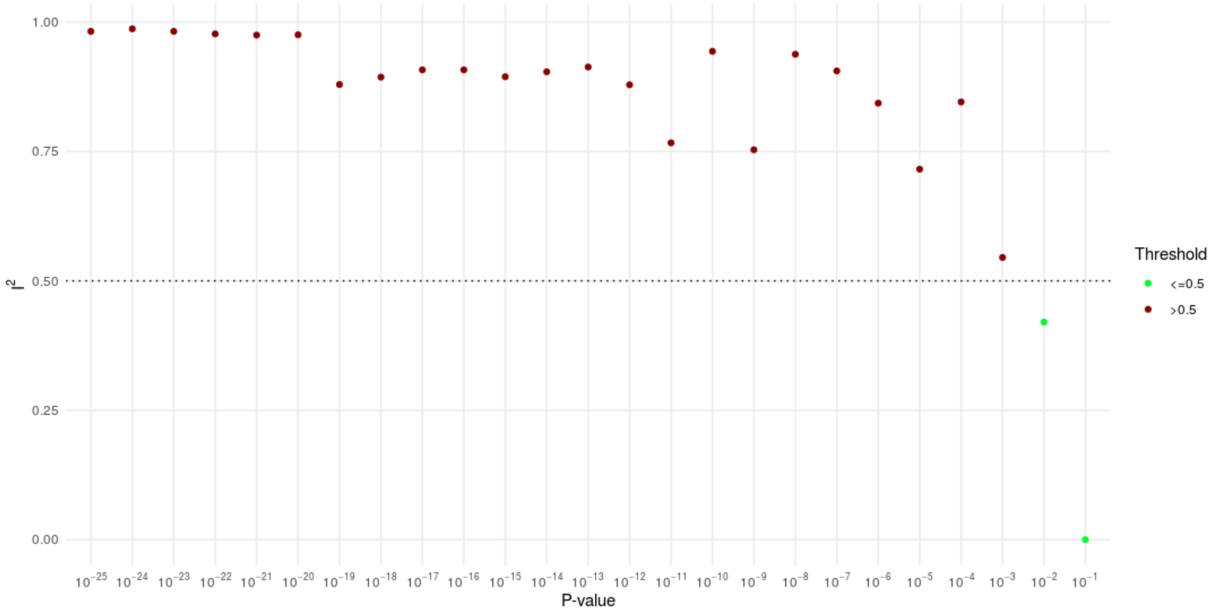

Figure 7: Scatter plot of I^2 against p-value used as threshold

As shown in Figure 7, the I^2 decreases as the p-value increases. The I^2 falls below 50% for p-value of 0.01 and 0.1 which means there is low heterogeneity with the SNPs used for the PRS calculated.

P-value	Number of used SNPs	Number of unused SNPs
0.01	7	280
0.1	3	284

Table 4: The number of SNPs used in the calculation of PRS for the respective p-value

Table 4 shows that the number of SNPs used in the calculation of the PRS is very low which caused the I^2 value to drop significantly. There are too few SNPs used for the PRS to have any predictive power so the modified PRS was not useful despite low I^2 values.

Accounting for the variation in allele frequencies across ethnic groups

Ethnicity	Mean RAF difference
Chinese	0
Malay	-0.00411
Indian	-0.01461

Table 5: Mean RAF difference between each ethnicity and Chinese

Table 5 showed the mean RAF differences across the 287 SNPs for Malay and Indian ethnic groups, using Chinese as reference. As expected, Indian has higher mean difference in RAF than Malay as Indians are genetically more distinct from Chinese compare to Malays.

Ethnicity	Original Mean	Mean after minus mean RAF difference
Chinese	0.25000	-
Malay	-0.05080	-0.03619
Indian	0.17900	0.18311

Table 6: Mean of PRS287 after minus mean RAF difference from the mean of PRS287 of each ethnicity

In Table 6, The Chinese mean of PRS287 is the same. Both Indian and Malay means shifted closer to the Chinese mean after subtracting the mean RAF difference.

PRS	I^2 (%)
Original PRS287	99.564
After minus mean RAF difference	99.518

Table 7: I^2 value between ethnicity for PRS287 and results after minus mean RAF difference

The results of the heterogeneity test in Table 7 show that the improvement in the I^2 result is negligible as it is still highly heterogeneous.

Principal Component Calculation

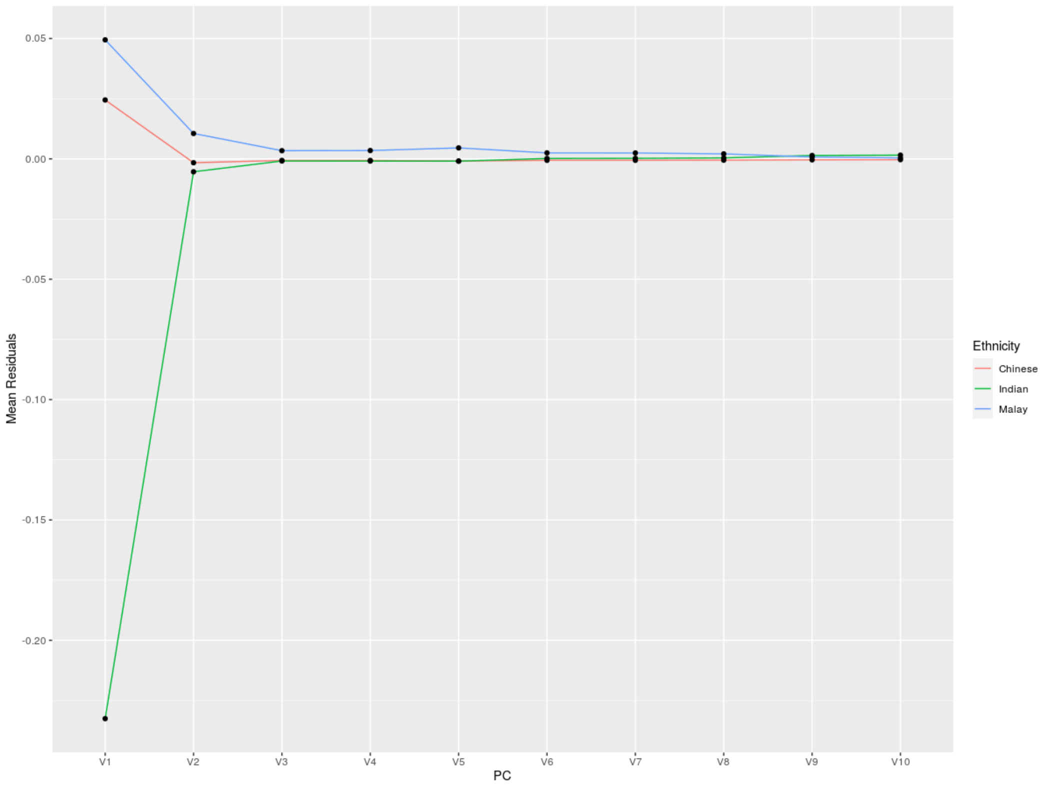

Figure 8: Line chart of mean residuals against number of PC by ethnic group. V1 indicates regression model of PRS~PC1, V2 indicates regression model of PRS~PC1+PC2, and etc. The mean residuals were calculated for each model across each ethnic groups.

Figure 8 shows the difference in mean PRS residuals for each candidate models across the three ethnic groups. The results showed that the heterogeneity in mean PRS across ethnic groups disappear as more PCs were included in the model.

Future Recommendations

The addition of PC has shown to reduce heterogeneity in the mean PRS so it is a method that should be further explored. Future work is needed to explore the minimum number of PCs that can optimally reduce the heterogeneity.

Conclusions

Although using fewer SNPs with similar allele frequencies across ethnicities reduced the heterogeneity, it also lowered the predictive power of the PRS constructed. Adjusting for the mean difference in allele frequency did not substantially improve the heterogeneity of the PRS. The addition of principal components however showed to successfully remove heterogeneity in the mean PRS so this method should be further explored.

Acknowledgement

Data used for the analysis was provided by Cancer Research Malaysia. The analysis was carried out under the supervision and guidance of Dr Ho Weang Kee.

References

[1] Ho, W.-K., Tan, M.-M., Mavaddat, N., Tai, M.-C., Mariapun, S., Li, J., Ho, P.-J., Dennis, J., Tyrer, J. P., Bolla, M. K., Michailidou, K., Wang, Q., Kang, D., Choi, J.-Y., Jamaris, S., Shu, X.-O., Yoon, S.-Y., Park, S. K., Kim, S.-W., … Antoniou, A. C. (2020). European polygenic risk score for prediction of breast cancer shows similar performance in Asian women. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-17680-w

[2] Mavaddat, N., Pharoah, P. D., Michailidou, K., Tyrer, J., Brook, M. N., Bolla, M. K., Wang, Q., Dennis, J., Dunning, A. M., Shah, M., Luben, R., Brown, J., Bojesen, S. E., Nordestgaard, B. G., Nielsen, S. F., Flyger, H., Czene, K., Darabi, H., Eriksson, M., Peto, J., … Garcia-Closas, M. (2015). Prediction of breast cancer risk based on profiling with common genetic variants. Journal of the National Cancer Institute, 107(5), djv036. https://doi.org/10.1093/jnci/djv036

[3] Ho, W.-K., Tai, M.-C., Dennis, J., Shu, X., Li, J., Ho, P. J., Millwood, I. Y., Lin, K., Jee, Y.-H., Lee, S.-H., Mavaddat, N., Bolla, M. K., Wang, Q., Michailidou, K., Long, J., Wijaya, E. A., Hassan, T., Rahmat, K., Tan, V. K., … Teo, S.-H. (2022). Polygenic risk scores for prediction of breast cancer risk in Asian populations. *Genetics in Medicine*, *24*(3), 586–600. https://doi.org/10.1016/j.gim.2021.11.008

[4] *Minor allele frequency*. Minor Allele Frequency - an overview | ScienceDirect Topics. (n.d.). https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/minor-allele-frequency

[5] Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (n.d.). *Chapter 5 between-study heterogeneity: Doing meta-analysis in R*. Chapter 5 Between-Study Heterogeneity | Doing Meta-Analysis in R. https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/heterogeneity.html