Development and Testing of an Optimal Cardiometabolic Genetic Risk Score to Predict Coronary Artery Disease Risk

Christopher B. Cole[1,2,3], Majid Nikpay[2,3], Ruth McPherson[2,3]

[1]Department of Biology, University of Ottawa, Ottawa, Canada
[2]Atherogenomics Laboratory, University of Ottawa Heart Institute, Ottawa, Canada
[3]Ruddy Canadian Cardiovascular Genetics Centre, University of Ottawa Heart Institute

Introduction

Coronary Artery Disease (CAD) is a major cause of morbidity and mortality globally; much international effort has been expended to detect risk factors, both heritable and environmental. Although there is a well established genetic basis for CAD, it has been difficult to characterize risk to the degree needed for phenotype prediction. Genome wide association studies (GWAS), a statistical methodology used to estimate the genetic effects of every loci in the genome in a hypothesis free manner, have identified just 46 significantly associated common loci explaining only a small fraction (~13%) of the predicted heritability of CAD, estimated by twin studies to be between 40 and 60%. Polygenic Risk Scores (PRS), linear combinations of weighted Single Nucleotide Polymorphisms (SNPs), have been successfully used to predict CAD with low to moderate accuracy; improvements in the methodology and implementation of these PRS are necessary for PRS to realize their full clinical potential.

Methods

Three polygenic risk scores (PRS) were created using summary information from several large consortia including the Coronary ARtery DIsease Genome wide Replication and Meta analysis plus The Coronary Artery Disease Genetics (CARDIOGRAM plus C4D), The Genetic Investigation of ANthropometric Traits (GIANT) consortium, and the Global Lipids Consortium. We used summary effects for LDLc, HDLc, TG, and BMI association.

The first, traditional risk score (TRS), model uses the typical procedure for constructing a PRS, simply summing over all significant variants m weighted by a previously identified risk score. That is, for m-vector of estimated genetic effects β_i and m by n individuals genetic matrix $\boldsymbol{G_{m,n}}$ coded 0, 1, or 2 depending on the number of minor alleles present at the given loci, the TRS score for individual n is given by

$$\hat{S}_{TRS,n} = \sum_{i \in m} \beta_i \boldsymbol{G}_{m,n}$$

The second, novel, risk score incorporates information from several co-morbid traits to re-prioritize the rankings of SNP and introduce new variants into the TRS. We call it a cardiometabolic (CMB) risk score. Thus for each of comorbid conditions $j\,\mathbf{1} \ldots c$, we sum over the significant loci for *each* of the comorbid conditions weighted by the estimated CAD genetic effects. Thus using the same notation as above:

$$\hat{S}_{CMB\ n} = \sum_{j=1}^{c} \sum_{i \in m} \beta_i \boldsymbol{G}_{m,n}$$

The third creates an optimal score which maximizes the P value of association while minimizing environmental noise. This is done by empirically maximizing the P value of association by sliding over thousands of P value thresholds for score inclusion, then finding the optimal one T_o.

$$\hat{S}_{oCMB\ n} = \sum_{j=1}^{c} \sum_{i \in m < T_o} \beta_i \boldsymbol{G}_{m,n}$$

These three scores are constructed in R, Plink, and GCTA, and validated for correctness.

Methods (Cont.)

In order to predict CAD with our PRS, covariate adjusted logistic (binomial) regression models were used. Scores were adjusted for the first two principal components to adjust for population stratification. (Zhang et al 2013)

$$\sigma(t) = \frac{e^{CAD}}{e^{CAD} + 1}$$
$$CAD = \beta_0 + \beta_1 X_{\hat{S}} + \beta_2 X_{PC_1} + \beta_3 X_{PC_2} + \epsilon$$

Nagelkerke's Pseudo R^2 gives an estimation of the scaled explained variance of a logistic model. It is given by

$$R^2 = 1 - \left\{L(M_{Intercept})/L(M_{Full})\right\}^{\frac{2}{N}}$$

Where $L(M)$ is the conditional probability of the outcome variable given each of the independent variables. Area under the receiver operator character curve, also known as the c-statistic, is a well known proxy for predictive accuracy of a binary model. It was calculated in the pROC package in R (Robin et al. (2011)). Differences between ROC model was assessed firstly by the method for correlated ROC curves proposed by DeLong et al. (1988). A secondary, more robust, difference was estimated through 1000 bootstrap permutations of dependent variables in R.

Meta analysis was conducted in the metafor package in R. (Viechtbauer, 2010) Random effects were assumed for study variables to have greater applicable to the population at large.

All analyses were conducted in R, Python, Plink, and GCTA on a large scale computer cluster; computational resources generously provided by the Centre for Advanced Computing.

Traditional Risk Score

The traditional risk score was significantly ($P < 2.2 \times 10^{16}$) associated with case/control status in all cohorts when adjusted for principal components to control for population stratification. (Price et al., 2006; Zhang et al., 2013). On average, scores between cases and control differed by $5.06{\times}10^{-4} \pm 1.73{\times}10^{-4}$.

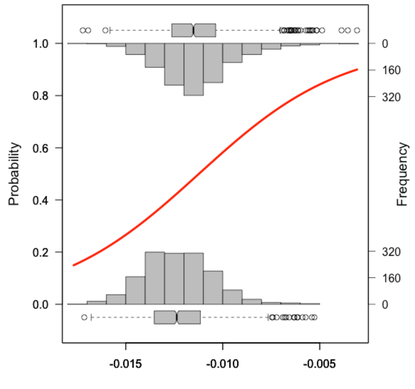

Figure 1: Fit logistic model representing traditional risk score with frequency distribution and boxplot.

Traditional Risk Score (Cont.)

The overall random effects meta analyzed AUC was 0.61 ± 0.03 for $\hat{S}_{TRS}$, with an average NagelKerke's Pseudo-R^2 of 0.047, a moderately well fit model.

As increasing the number of SNP in the score will always increase the fit of the model, we compare the 202 FDR significant loci to scores constructed through randomly selected loci in 1000 bootstraps in **Figure 2**.

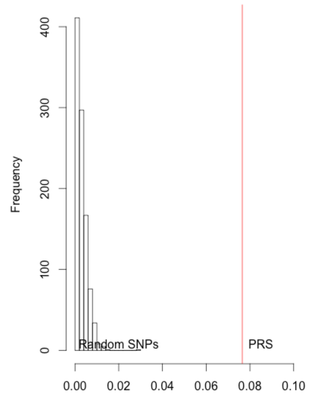

Figure 2: Fit of TRS predicting CAD versus 1000 bootstraps.

Cardiometabolic Risk Score

It was found that the CMB risk score composing only the CAD and BMI risk loci fit the model the best, so it will be examined in detail from here out.

In permutation analyses, each of the scores performed significantly better than an equivalent number of randomly selected SNP in 1000 bootstraps, similar to **Figure 2.**

The overall random effect meta analyzed AUC ROC was found to be 0.65 [0.64, 0.67], a significant ($P < 2.2 \times 10^{-16}$) improvement on the previous model through bootstrap and Delong et al's test for correlated ROC Curves.

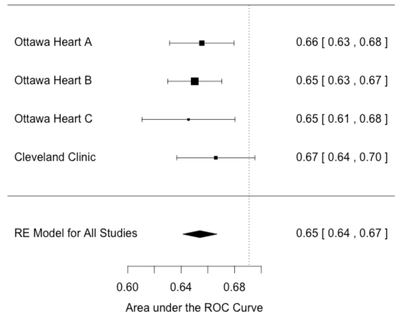

Figure 3: Random Effect Meta Analysis of area under the receiver operator characteristic curve for cardiometabolic risk score predicting CAD in four cohorts.

Cardiometabolic Risk Score (Cont.)

Persons in the upper quintile of this score were 81% more likely (1725 cases vs 953 controls) to have CAD than not (compared to 70 % for the previous score) and people in the bottom quintile were 15.7% less likely (1240 cases vs 1471 controls) to have CAD than having it. There was also a substantive increase in NagelKerke's Pseudo R^2 in the second score compared to any other, especially in the Cleveland cohort.

Optimal Cardiometabolic Risk Score

Optimal P value cutoffs are displayed in **Table 1**.

	LDLc	HDLc	TG	BMI
OHGS A	0.0001	0.0055	0.1743	1
OHGS B	0.2484	0.0999	0.002	1
OHGS C	0.1299	0.0085	0.1528	1
Cleveland	0.1807	0.2039	0.004	1

Table 1: Optimal P value inclusion thresholds by cohort and co-morbid condition.

The optimal cardiometabolic score performed the best out of all scores tested. It was significantly associated with a higher risk of CAD, having an exceptionally high AUC ROC of 0.78 [0.75, 0.81] and R^2 between 0.245 and 0.338 depending on the cohort. The score comprised a large number of SNPs, between 440,000 and 850,000. However, because of the large number of SNPs, the predictive ability of the score was insignificantly different than 1000 randomly constructed scores of equal SNPs.

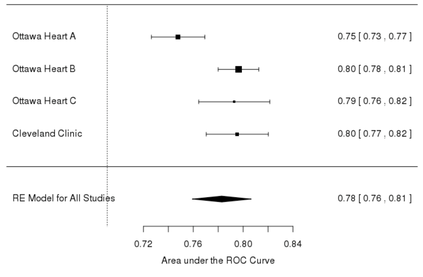

Figure 4: Random Effect Meta Analysis of area under the receiver operator characteristic curve for optimal cardiometabolic risk score predicting CAD in four cohorts.

Additionally, we found that there was no optimal threshold for BMI risk as seen in **Figure 5**.

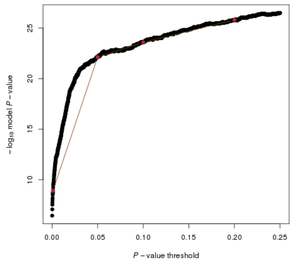

Figure 5: Association of BMI PRS by P value threshold.

Future Directions

Our future work will concentrate on validating these new score's validity mathematically. We will attempt to explain why we have observed the trends which we have seen, as well as making comments on the genetic architecture of traits using our scores following Dudbridge 2013.

Additionally, we will investigate the roll of linkage disequilibrium (LD) on score construction and performance; we theorize that large and complex LD structures may be influencing the trends we observe especially in the oCMB score.

We will also examine cross-over conditions; how duplicate or triplicate variants which co-occur in co-morbid conditions should be treated remains an open and interesting question.

Discussion / Conclusion

In this project we have introduced two novel methodologies for extending traditional polygenic prediction of complex phenotypes using only summary information from co-morbid conditions. Both were found to be significantly superior to traditional PRS in terms of model fit (NagelKerke's Pseudo R^2) and predictive accuracy (AUC ROC).

The CMB model represents a small but significant improvement over traditional methodologies, however information from lipid traits was not additionally predictive once information from BMI was incorporated. There have previously been shown to be substantial gene × environment interactions in lipid levels, and these interactions may be causing the model to perform unpredictably.

The oCMB model showed great potential, as it has displayed a very high predictive accuracy and explains between 24 and 34 percent of variation in CAD. Given that twin studies have estimated genetic heritability for CAD ranges between 40 and 60 percent, the oCMB model categorizes a large portion of the variance in CAD risk attributable to genetics. Additionally, BMI was found to not have an optimal threshold, which is consistent with a genetic model involving many small variants. This lends evidence to many common variants of small effect size influencing BMI.

Our new methodologies improve upon our ability to predict CAD. Polygenic prediction of complex disease is quickly becoming more efficient and accurate; our model represents one step forward towards eventual clinical prediction of complex disease and the advent of truly personalized genetic medicine.

References

1. The CARDIoGRAMplusC4D Consortium. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47,** 1121–1130 (2015).
2. Nagelkerke, N. J. D. A note on a general definition of the coefficient of determination. *Biometrika* **78,** 691–692 (1991).
3. McPherson, R. & Tybjaerg-Hansen, A. Genetics of Coronary Artery Disease. *Circ. Res.* **118,** 564–578 (2016).
4. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31,** 1466–1468 (2014).
5. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet.* **9,** e1003348 (2013).
6. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **44,** 837–845 (1988).

Contact Information: Christopher B. Cole (Ccole019@uOttawa.ca)