University of Ottawa

Department of Biology

HONOURS B.Sc. BIOMEDICAL SCIENCE, OPTION IN BIOSTATISTICS

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

Honours Dissertation of:

Christopher B. Cole

Thesis Supervisor:

Prof. Ruth McPherson, MD, PhD, FACP, FRCPC, FRSC

Secondary Thesis Supervisor:

Dr. Majid Nikpay, PhD

May 2016

Preface

Fill in later

Christopher B. Cole
Ottawa
May 2016

Abstract

Background and Rationale: Coronary artery disease (CAD) is a major cause of morbidity and mortality and much international effort has been expended to detect risk factors, both heritable and environmental. Although there is a well established genetic basis for CAD, genome wide association studies (GWAS) have identified just 46 common loci, explaining only a small fraction (13%) of the predicted heritability of CAD, estimated by twin studies to be between 40 and 60%. This âĂIJmissing heritabilityâĂİ, may be explained by diverse phenomenon including multiple common variants of very low effect size that may act via multiple causal risk factors for CAD and escape detection in sample sizes investigated to date, rare variants (MAF $< 1\%$) of high effect size, gene ÃŮ gene (GÃŮG) interactions, and gene ÃŮ environment (G × E) interactions. Previous efforts have tested the ability of a genetic risk score based on from 13 to 30 CAD-associated single nucleotide polymorphisms (SNPs) to predict CAD risk. Even this small number of risk alleles was shown to have significant predictive power and recently, to identify individuals who would benefit most from statin therapy to reduce LDL concentrations. However, improvements in genetic risk assessment are necessary and feasible given recent genetic advancements.

Purpose and Specific Objectives: This study hopes to develop an improved genetic risk score for coronary artery disease using a panel of independent risk loci. We address whether or not a panel of 202 independent SNPs with stepwise addition of cardiometabolic condition SNPs significantly predicts CAD.

Materials and Methods: 202 Independent SNPs were identified through GWAS and linear regression with multidimensional scaling in PLINK. The present study will use a stepwise logistic regression model with principal components and additional covariates. The independent variable will be a composite of genetic risk equal to a weighted sum of risk alleles with mean value imputation. The study will also compute NagelkerkeâĂŹs Psuedo-R2 as a proxy measure for goodness of fit of the model. Additionally, we will compute the receiver operator characteristic curve and calculate the area under the curve to determine model predictive accuracy. The net recombination index will also be calculated for each model. Accurate multiple correction will be performed with respect to the correlation matrix between tests. Additionally, the above analysis will be repeated using different FDR thresholds using the R program PRSice.

Results: This study will result in several metrics describing the model's ability to predict CAD in a population. If the predictive ability of our score is meaningful, it will allow clinical researchers to diagnostically determine individual risk to CAD.

Contents

List of Figures vi

List of Tables viii

List of Acronyms xi

1 Introduction 1

1.1 Genetics of Coronary Artery Disease 2

1.2 Genome Wide Association Studies 4

1.2.1 Primer on Genetics 4

1.2.2 Sequencing 4

1.2.3 Statistical Definition 4

1.2.4 Multiple Comparisson Problem 6

1.3 Polygenic Prediction of Complex Disease 7

1.4 Optimal Polygenic Risk Scores 8

1.5 Summary and Study Goals 10

2 Methods 11

2.1 Study Population 11

2.1.1 Population Descriptions 11

2.2 Genotyping and Imputation 11

2.3 Training Populations 11

2.3.1 GIANT Consortium 11

2.3.2 Global Lipids Consortium 11

2.4 Polygenic Prediction of CAD 11

2.4.1 Traditional Risk Score 11
2.4.2 Cardiometabolic Risk Score 11
2.4.3 Optimal Cardiometabolic Risk Score 11

Bibliography **13**

List of Figures

1.1 Progression of the formation of plaque causing *Coronary Artery Disease* (CAD). Adapted from Gretch 2003. 2

1.2 Fine mapping of the genomic interval on chromosome 9 associated with Coronary Heart Disease. Adapted from Mcpherson et al 2006 3

1.3 Example of a Manhattan plot from a *Genome wide Association Study* (GWAS) for CAD performed by Shunkert et al. 2011 6

1.4 *Optimal Polygenic Risk Score* (oPRS) plot for schizophrenia predicting major depresive disorder status. Adapted from Euesden et al. (2014) 9

1.5 Expected $-\log_{10}(P)$ value of linear regression estimate as a function of P-value threshold for selecting markers into Polygenic score. Note that π_0 refers to the proportion of *true null* markers, that is, markers which have no effect on phenotype. (Dudbridge (2013)) 10

List of Tables

1.1 Notation relating to hypothesis testing. Adapted from Sun et al. (2006) 6

List of Acronyms

CAD	Coronary Artery Disease	vi
PRS	Polygenic Risk Score	1
oPRS	Optimal Polygenic Risk Score	vi
CARDIOGRAMC4D	Coronary ARtery DIsease Genome wide Replication and Meta analysis plus The Coronary Artery Disease Genetics	1
GWAS	Genome wide Association Study	vi
GLC	Global Lipids Consortium	1
GIANT	The Genetic Investigation of ANthropometric Traits	1
BMI	Body Mass Index	1
MI	Myocardial Infarction	2
kb	kilobase	2
DNA	Deoxyribonucleic acid	4
A	Adenine	4
C	Cytosine	4
T	Thymine	4
G	Guanine	4
locus	specific genetic location	4

CNV copy number variant 4

InDel insertion/deletion 4

RNA ribonucleic acid 4

SNP single nucleotide polymorphism 4

LD linkage disequilibrium 6

FWER family wise error rate 7

FDR false disovery rate 7

PRDS positive regression dependence on subsets 7

OR odds ratio 5

A note on notation

Throughout this thesis, the following conventions for notation are used.

1. A hat ($\hat{\cdot}$) denotes the estimator of a variable (i.e. $\hat{\beta}$ is the estimator of β).
2. Underlining a variable ($\underline{\cdot}$) implies that it is a n-vector, or $n \times 1$ dimensional matrix. (i.e. $\underline{Y} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$ is a $n = 3$-vector).
3. Bolding indicates a matrrix (i.e. $\mathbf{G} = \begin{bmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{bmatrix}$)

Introduction 1

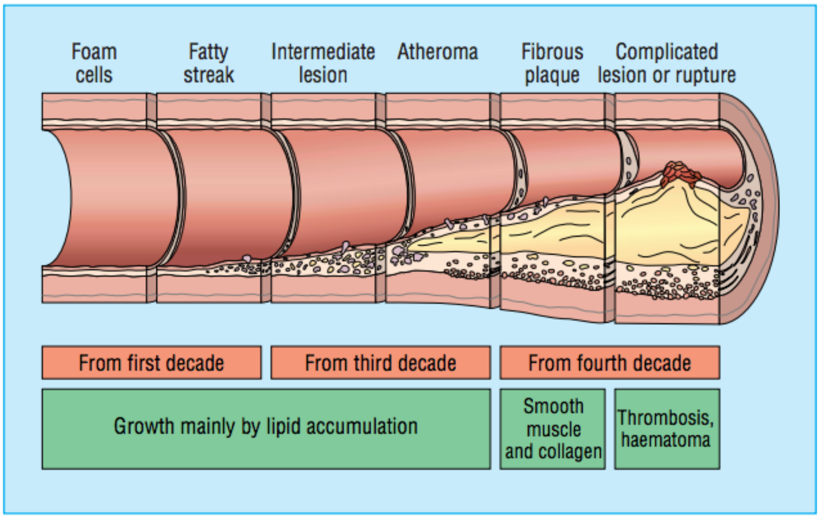
As the efficiency and accuracy of rapid genome sequencing skyrockets, the potential for personalized therapies has made its way from science fiction to scientific reality. Using genetics to understand, diagnose, and eventually to predict illness is not a new idea; in recent years, however, technological ability and scientific understanding have advanced to such a point that researchers may predict risk for several diseases with reasonable confidence. Increasingly, variants in the human genome are being identified as being robustly linked to risk for complex illnesses such as heart disease [cite 9p21], obesity [cite fto], and schizophrenia [cite something]. However, much work remains to be done in order to create tools which may accurately predict individual disease risk from known and unknown genetic risk factors. In this thesis, we propose a novel extension to a well known methodology in order to better characterize disease risk from comorbid conditions using only summary statistics.

In brief, we present preliminary evidence for the use of *Polygenic Risk Score* (PRS)s in predicting CAD. We use recently published summary statistics from a GWAS conducted by the *Coronary ARtery DIsease Genome wide Replication and Meta analysis plus The Coronary Artery Disease Genetics* (CARDIOGRAMC4D) consortium alongside evidence gathered by the *Global Lipids Consortium* (GLC) and the *The Genetic Investigation of ANthropometric Traits* (GIANT) consortium for lipids and *Body Mass Index* (BMI) *per say.* We use this data alongside previously identified variants to construct first a simplistic PRS using only genome wide statistically signficant ($P_{Bonferonni} < 0.05$ or $q_{FDR} < 0.05$) variants, then expand our search to variants which may not be as robustly linked to phenotype. [Cite storey, BH, and dudbridge]. We use an empirical maximation approach and several strategies of mathematical optimization in order to construct an oPRS, then devise a novel technique for integrating information from co-morbid oPRS diseases in order to better predict CAD in four cohorts comprising approximately $n = 12,000$ individuals

1.1 Genetics of Coronary Artery Disease

CAD occurs when the major blood vessels supplying the heart become diseased or damaged, often leading to severe complications such as *Myocardial Infarction* (MI) and death. [cite review articles] CAD is known to be a complex genetic disease with heritability estimated by twin studies between 40 and 60%. [mcPherson 2016, twin studies paper] Several important variants have been indentified which have been shown to robustly increase risk to CAD by altering lipid transporting pathways [cite LDLR], structural collegan bodies [TRIB 1??], and others factors.

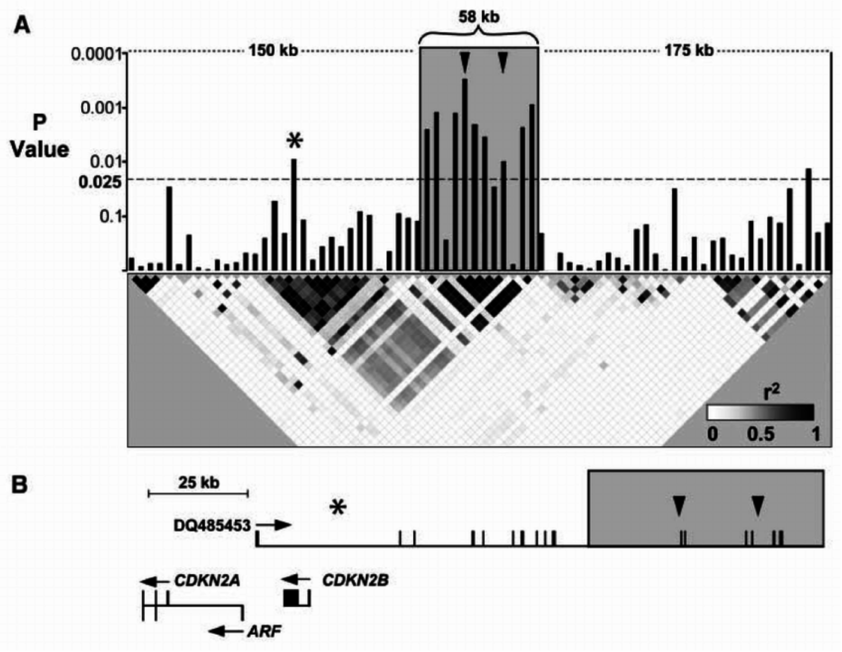
FIGURE 1.1: Progression of the formation of plaque causing CAD. Adapted from Gretch 2003.


Foam cells
Fatty streak
Intermediate lesion
Atheroma
Fibrous plaque
Complicated lesion or rupture
From first decade
From third decade
From fourth decade
Growth mainly by lipid accumulation
Smooth muscle and collagen
Thrombosis, haematoma

With heart disease and stroke the leading cause of perscription drug use in Canada as well as one of the leading causes of death and hospitalization [cite herat and stroke], the need to better understand, diagnose, and prevent this deadly disease is apparent. In order to better understand the need for improved statistical methodologies, it is important to understand the large body of previous attempts to characterize the genetic determinants of CAD

Despite some promising beginings, initial attempts to understand and explain CAD through genetics were largely unsuccessful.[CITE] The first variant to be succesfully and robustly linked to risk for CAD was the 9p21.3 locus. Discovered by a team of researchers at the Univeristy of Ottawa Heart institute, the allele consists of a 58 *kilobase* (kb) region on chromosome 9 which was shown to be associated with CAD in a population of 23,000 caucasion individuals. (McPherson and Tybjaerg-Hansen (2016))

Figure 1.2: Fine mapping of the genomic interval on chromosome 9 associated with Coronary Heart Disease. Adapted from Mcpherson et al 2006



This initial success began the era of the GWAS, explained in more detail in section 1.2. Researchers across the globe began frantically searching for more loci with the hope of understanding and predicting complex disease; in that goal, the GWAS has failed. (Visscher et al. (2012)) A number of important genetic markers for CAD have been discovered, but often in small familial cases or with very low effect sizes. [Cite] As the dust settles and the low hanging fruit have been picked, common variants have been shown to explain approximately 28% of the heritablity of CAD [cite majid], yet a large portion remains to be accounted for. This has become known as the problem of "missing heritabillity" of complex disease; common genetic variants explain a relatively small portion of the total estimated heritability of a disease, therefore researchers must resort to ever more obscure and complex methods to attempt to explain the complex interactions between genetic elements in the human genome. [cite review paper] From pathway analysis to partitioned heritability to all kinds of arcane statistical procedures, researchers from across the globe have tried their hardest to shrink this gap between our knowledge and accurate prediction and understanding of complex disease. To this end, we develop our own methodology incorporating multiple sources of information for the more accurate prediction of clinical end points.

1.2 Genome Wide Association Studies

In order to properly introduce the model, however, the basic underpinnings must be explored and explained. Genome wide association studies seek to indetify associations between individual genotypes and disease phenotypes in a hypothesis free manner. In this section, the statistical model required to understand GWAS is presented and explored.

1.2.1 Primer on Genetics

Deoxyribonucleic acid (DNA) is a double helical molecule which encodes the genetic blueprints for the construction of proteins and other materials that make up every known living organism. DNA is composed of three parts: a negatively charged phosphate group, a five carbon sugar *deoxyribose*, and (usually) one of four nitrogen bases. It is these bases, *Adenine* (A), *Cytosine* (C), *Thymine* (T), and *Guanine* (G) and their combinations which are under investigation in a GWAS. The specific combinations of these four bases in a *specific genetic location* (locus) determine the product produced by the DNA, and even a small change in this order can have large ramifications on the overall health, survival, and proper function of the organism.

1.2.2 Sequencing

DNA sequencing is the process of ascertaining a particular individual's genotype by means of chemical identification of the bases present at predefined sites. [cite] These sites, whether they be a change in a single base called a *single nucleotide polymorphism* (SNP), a variation in the number of tandem repeats of a small sequence named a *copy number variant* (CNV) or an *insertion/deletion* (InDel) of a sequence, may alter amino acid sequence, affect regulatory regions, or impact regulatory *ribonucleic acid* (RNA) sequences.

Definition 1.2.1 (Allele) *A specific form or subtype of a genetic locus. This could be one or more individual variations or a combination therof.*

Remark 1 *Allele frequency is the frequency at which a particular allele occurs in the population. I.e. for locus A having n different alleles, the true population allele frequency of allele* $freqA_m \equiv \frac{A_m}{\sum_{i=1}^{n} A_i}$, *which is estimated in a sample population with a biased ratio estimator* $freq\hat{A}_m \equiv \frac{\hat{A}_m}{\sum_{i=1}^{n} \hat{A}_i}$

1.2.3 Statistical Definition

Consider a simple case control population where 1 defines case and 0 defines control. Define $\underline{\mathbf{Y}}$ as an n-vector where n denotes the number of individuals in a population and $\underline{\mathbf{Y}}_i$ gives the individual's diesease staet. Additionally define G as an $m \times n$ matrix where m is the number of informative genotypic sites available with $\mathbf{G}_{ij}$ being the "state" (allele number) present at site $j, 1 \leq i \leq m, i \in \mathbb{Z}^+$ in individual $i, 1 \leq i \leq n, i \in \mathbb{Z}^+$.

$$\underline{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{G} = \begin{bmatrix} G_{1,1} & \dots & G_{1,n} \\ \vdots & \ddots & \vdots \\ G_{m,1} & \dots & G_{m,n} \end{bmatrix}$$

In an additive genetic model, we define the phenotype $\underline{\mathbf{Y}}$ as a linear combination of $\mathbf{G}$ weighted by a vector of $\underline{\beta}$ coefficicent vectors estimated by regression analysis and $\underline{\epsilon}$ vector of errors. Express $\underline{\mathbf{Y}}$ such that

$$\underline{\mathbf{Y}} = \underline{\beta}'\mathbf{G} + \underline{\epsilon} = \left(\sum_{i=1}^{m} \beta_i \mathbf{G}_{i,n} + \epsilon_n \right)'$$

$\underline{\beta}$ and $\underline{\epsilon}$ are approximated optimally by $\underline{\hat{\beta}}$ and $\underline{\hat{E}}$ in practice.

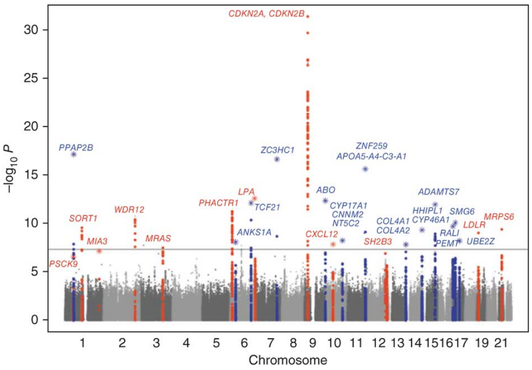
The purpose of a GWAS is not only to estimate these genetic effects $\underline{\beta}$ by $\underline{\hat{\beta}}$ but also to estimate their significance of association with phenotype vector $\underline{\mathbf{Y}}$ through a χ^2 test and corresponding test statistic m-vector $\hat{\chi}^2$. The degrees of freedom of this test statistic will vary between methods and models, and so will be left as futher reading.

GWAS commonly estimate these effects through linear regression. The disease state (or disease level, should it be a continuous variable) is used as the response variable, while the main dependent variable is usually the number of minor alleles (0, 1, or 2) present. The β coefficient (for continous disease state) or *odds ratio* (OR), therefore represents the average increase (for the continuous case) or the OR per additional risk allele present.

Remark 2 *This description assumes an additive genetic model, which states that the effect of possessing one minor allele is exactly the same as half the effect of having two risk alleles. Additional genetic models include the dominant scheme, where the effect of having two minor alleles is the same as having one minor allele, the recessive scheme where only the case of two minor alleles impacts the phenotype, and the general genetic model, where the effect of one allele is* $a\times$ *the effect of two alleles,* $a \in [0, 1]$.

By approximating $\underline{\chi^2}$ with $\underline{\hat{\chi^2}}$ and computing the corresponding P values, reserchers are able to identify and quantify the effects of variants significantly ($P < 0.05$) associated with the phenotype. These results can be summarized in a Manhattan plot, named after the city of Manhattan with it's high rise buildings towering over the scenery. The x axis of this plot is the genomic location (usually coloured by chromosome number) while the y axis is the $\log_{10}$ of the P value of association derived from $\underline{\hat{\chi}^2}$.

FIGURE 1.3: Example of a Manhattan plot from a GWAS for CAD performed by Shunkert et al. 2011



1.2.4 Multiple Comparisson Problem

In such a set up, where m may be in the millions and the threshold of significance is set to $P = \alpha = 0.05$, we encounter a canonical issue in statistical inference. Recall that P is the probability of observing a χ^2 statistic as large or larger than a specific χ^2_m assuming H_0 of no association is correct and α is the threshold at which a significant effect is declared. Table 1.1 introduces relevant notation for this section.

Table 1.1: Notation relating to hypothesis testing. Adapted from Sun et al. (2006)

	True H_0	**True H_1**	**Total**
Declared significant	V	S	R
Declared non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

For the sake of description, we define M as the number of *independant* variants (that is, the effective number of variants which are not in *linkage disequilibrium* (LD) for a given R^2 or D' threshold) for sake of description. Thus, M tests and corresponding M-vector of P values $\underline{\mathbf{P}}$ is constructed. Because in any statistical test, assuming that H_0 is true, there is α chance of falsely rejecting H_0 (type I error), by increasing the number of simultaneous tests conducted, the probability of falsely rejecting H_0 compounds exponentially as a function of the number of independent test conducted. That is, the conditional probability of falsely rejecting H_0 for all M tests may be written as

$$Pr(P \leq \alpha \,|\, H_0) = 1 - (1 - \alpha)^M$$

This may equivalently be described as the probability of making at least one false positive in M tests. This may alternatively be notated

$$Pr(V \geq 1) = 1 - (1 - \alpha)^M$$

Speaking asymptotically, $\lim_{M \to \infty} 1 - (1 - \alpha)^M = 1$ and false positves are guarenteed. It is against this backdrop that we recall in any relevant genetic context, M is large, and false positives are almost guarenteed.

There exist several ways to correct for this issue, chief among them is the widely adopted Bonferroni correction. Put simply, Bonferroni correction adjusts testing such that $Pr(V \geq 1) = \alpha$ rather than $1 - (1 - \alpha)^M$. It does so by rejecting all tests $p_i \in \underline{\mathbf{P}} | i \in 1 \ldots M, i \in \mathbb{Z}^+$ such that

$$p_i \leq \frac{\alpha}{M}$$

The proof is not complex, but shall not be presented here for the sake of brevity. [cite bf] This adjustment (for $Pr(V \geq 1)$) is defined as control of the *family wise error rate* (FWER). This approach does not make any assumptions about the internal dependency strucutre of the tests, and as such, is conservative in the case of all categories of dependency. This is often undesired, as typically researchers will not prune their GWAS data to only independant variants. A more commonly accepted procedure, controlling the *false disovery rate* (FDR) rather than the FWER adjusts $\underline{\mathbf{P}}$ such that the proportion of false disoveries in all discoveries is controlled at α:

$$FDR \equiv E\left[\frac{V}{R}\right] = \alpha$$

This approach has the benefit of being adaptable and more powerful in circumstances of some forms of dependency (most notably *positive regression dependence on subsets* (PRDS) which is common scenario) and is most often applicable to GWAS where researchers are more willing to find more true positives at the cost of a fraction of false positives.

Therefore, in summation, GWAS is a statistical investigation which estimates several parameters given certain assumptions. Concepts presented in this section will be important background knowledge for the following sections, as most of our model builds off of these premeses.

1.3 Polygenic Prediction of Complex Disease

Refering to the defintions proposed in the previous section and recalling that in a general additive model, a phenotype vector $\underline{\mathbf{Y}}$ may be expressed as a linear combination of the $\underline{\beta}$ weighted genetic $n \times m$-matrix $\mathbf{G}$ and $\underline{\epsilon}$ following a standard normal $N(0, 1)$ distrbution:

$$\underline{\mathbf{Y}} = \underline{\beta}'\mathbf{G} + \underline{\epsilon}$$

It has been previously proposed to combine genetic variants in order to crease a score S which encompases estimated genetic effects in order to predict the phenotype vector $\underline{\mathbf{Y}}$. Define S for individual n:

$$S_n = \sum_{i=1}^{m} \beta_i G_{ni}$$

Note that in practice, our true statistics must be estimated. The logical estimator of $\underline{\beta}$ is the ordinary least squares regression estimator $\hat{\beta}$. There are other estimators, but the remainder of this section assumes this estimator. Our score is therefore described as:

$$\hat{S}_n = \sum_{i=1}^{m} \hat{\beta}_i G_{ni} \tag{1.1}$$

This score has several important properties which will be exploited in the below analysis. Note that in practice, $\underline{\beta}$

Notably, the non-centrality parameter of the χ^2 test for association between $\hat{S}$ and $\underline{Y}$ in the test population, assuming that $\hat{\beta}$ has been estimated in a training population of size n_1 and tested in a test population of size n_2, is given by:

$$\lambda = \frac{n_2 R^2_{\hat{S},Y}}{1 - R^2_{\hat{S},Y}}$$

Where $R^2_{\hat{S},Y}$ is the percent explained variance of the phenotype Y with the estimated score $\hat{S}$. Additionally, note that $E[\hat{S}] = 0$ and the second moment in a particular individual is given by

$$\begin{aligned} Var(\hat{S}) &= \sum_{i=1}^{m} Var(\hat{\beta}_{il}, G_i) \\ &= \sum_{i=1}^{m} \hat{\beta}_i \\ &\approx mVar(\hat{\beta}_i) \end{aligned}$$

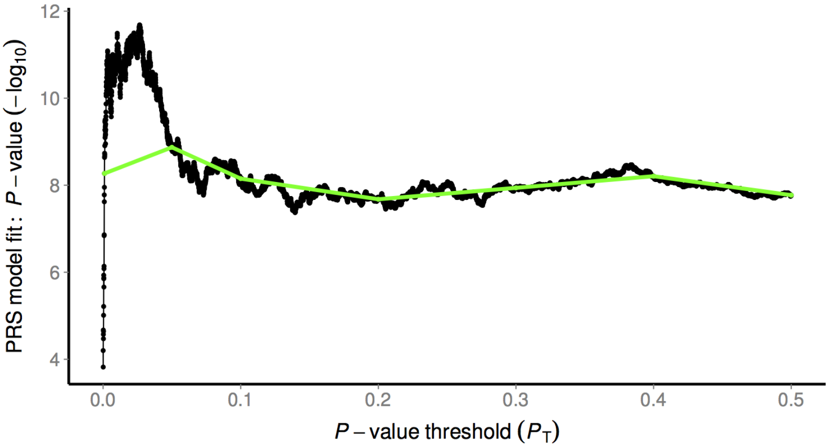
These mathematical properties become important later. These identities have been adapted from Dudbridge (2013).

1.4 Optimal Polygenic Risk Scores

Frequently, not all m variants are used in the construction of the PRS though. Typically, researchers will type the top $m|P_m \leq \alpha_{adj}$ where α_{adj} denotes the shifted acceptance threshold after multiple testing correction. We denote these variants as $m_{P \leq T}$ where T is the P value threshold.

Though these variants have the highest probability of being truly associated with the phenotype, constructing a score with this few SNPs misses the many small and insignificant effects hidden in marginally significant and insignificant hits. Thus, Euesden et al. (2014) have developed a method to find the best-fit PRS, that is, the PRS which maximizes genomic signal while minizing noise as in 1.4. We denote this as the oPRS.

FIGURE 1.4: oPRS plot for schizophrenia predicting major depresive disorder status. Adapted from Euesden et al. (2014)



On a high level, this score involves iterating through a list of P value thresholds T, constructing a score using all $m_{P<T}$ and selecting either the smallest P value of association between $\hat{S}$ and $\underline{\mathbf{Y}}$ or the highest $R^2_{\hat{S},Y}$ to move in the analysis.

More formally, we fix individual n and construct a vector of estimated scores $\underline{\hat{S}}$ with length n_T equal to the number of attempted P value thresholds.

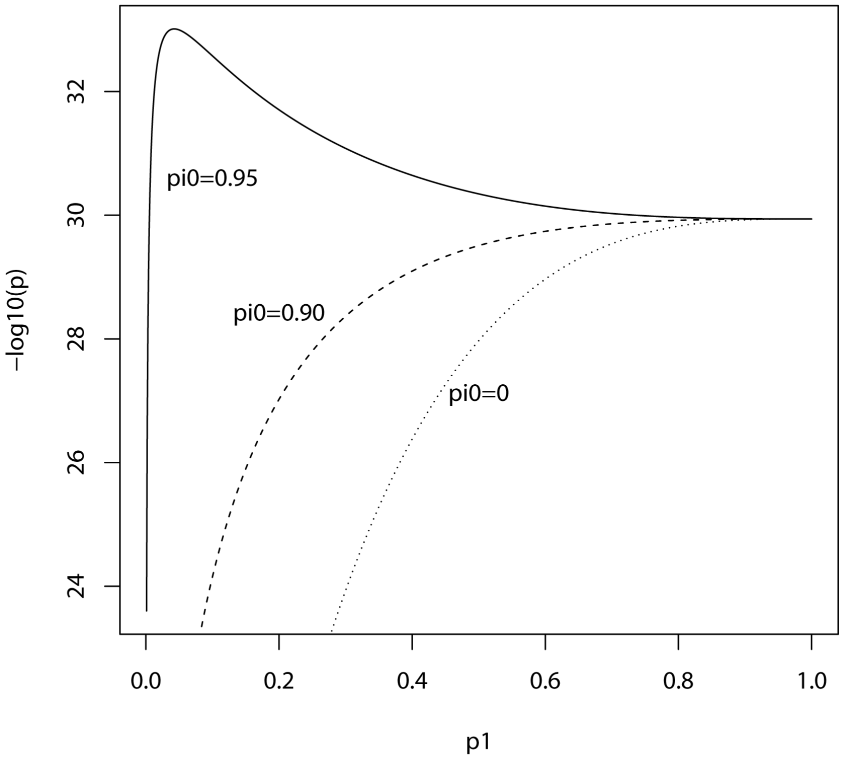
$$\underline{\hat{S}} \equiv \begin{bmatrix} \hat{S}_{T_1} \\ \vdots \\ \hat{S}_{n_T} \end{bmatrix} \qquad \hat{S}_T = \sum_{i=1}^{m_{P \leq T}} \hat{\beta}_i G_{ni}$$

Note, however, that when we build a score at each threshold for each individual, an $n \times T$ matrix is constructed, where n is the number of individuals and T is the number of thresholds. The entries are the estiamted score $\hat{S}_{nT}$ for individual n at threshold T:

$$\hat{\mathbf{S}} = \begin{bmatrix} \hat{S}_{1,1} & \dots & \hat{S}_{1,T} \\ \vdots & \ddots & \vdots \\ \hat{S}_{n,1} & \dots & \hat{S}_{n,T} \end{bmatrix} \tag{1.2}$$

It is from the matrix described in 1.2 that the rest of our model will be built.

FIGURE 1.5: Expected $-\log_{10}(P)$ value of linear regression estimate as a function of P-value threshold for selecting markers into Polygenic score. Note that π_0 refers to the proportion of *true null* markers, that is, markers which have no effect on phenotype. (Dudbridge (2013))



Remark 3 *Though it is always possible to construct an optimal score, only in certain circumstances is the P value threshold* $P_T < 1$*, depending on the internal structure of the disease under question. At different heritability levels, a disease may only have an optimal score with* $P_T = 1$*, as described in Figure 1.4.*

1.5 Summary and Study Goals

In this section, relevant background information pertaining to genetics and GWAS was presented an explored. We introduce notation and theory for PRS and oPRS as well as touching on some theoretical properties which will be exploited later in the analysis.

Make sure to add in study goals here more clearly

Methods 2

Building on the previously introduced notation, we introduce the present study.

2.1 Study Population

2.1.1 Population Descriptions

2.2 Genotyping and Imputation

2.3 Training Populations

2.3.1 GIANT Consortium

2.3.2 Global Lipids Consortium

2.4 Polygenic Prediction of CAD

2.4.1 Traditional Risk Score

2.4.2 Cardiometabolic Risk Score

2.4.3 Optimal Cardiometabolic Risk Score

Bibliography

Frank Dudbridge. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, 9(3):e1003348, mar 2013. ISSN 15537390. doi: 10.1371/journal.pgen.1003348. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3605113{&}tool=pmcentrez{&}rendertype=abstract`.

J. Euesden, C. M. Lewis, and P. F. O'Reilly. PRSice: Polygenic Risk Score software. *Bioinformatics*, 31(9):1466–1468, dec 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu848. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4410663{&}tool=pmcentrez{&}rendertype=abstract`.

Ruth McPherson and Anne Tybjaerg-Hansen. Genetics of Coronary Artery Disease. *Circulation Research*, 118(4):564–578, feb 2016. ISSN 0009-7330. doi: 10.1161/CIRCRESAHA.115.306566. URL `http://circres.ahajournals.org/content/118/4/564.abstract`.

Lei Sun, Radu V. Craiu, Andrew D. Paterson, and Shelley B. Bull. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genetic Epidemiology*, 30(6):519–530, sep 2006. ISSN 07410395. doi: 10.1002/gepi.20164. URL `http://www.ncbi.nlm.nih.gov/pubmed/16800000`.

Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of GWAS discovery. *American journal of human genetics*, 90(1):7–24, jan 2012. ISSN 1537-6605. doi: 10.1016/j.ajhg.2011.11.029. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3257326{&}tool=pmcentrez{&}rendertype=abstract`.