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

Throughout my undergraduate degree, I have made my fair share of mistakes. In truth, I have made many people's share of mistakes. Like the Rolling Stones, these failures have not always been what I wanted, but oftentimes they were exactly what I needed. This thesis represents the culmination of thousands of lessons taught to me by some truly incredible people.

I would first like to thank Professor Aris-Brosou for taking a chance on a Biomedical Science student first year who knew nothing about programming or statistics, without whom my life would doubtlessly be very different.

I would also like to thank Majid Nikpay, who has been a mentor to me for all of the years I have been researching statistical genetics.

Additionally, I would like to thank all of my friends for the support and motivation they have provided throughout my degree. In particular, I would thank Aaron R. Shifman. He has been a comrade in a lonely field, and a sounding board for my oftentimes less than genius ideas.

My most heartfelt thanks goes out to my girlfriend of four years, Kennedy Hao. She has motivated me to continue on no matter the obstacle and has proven time and again to be my closest ally, no matter the battle.

Through any and all obstacles, I have had no further to look for support than to my family. They have helped me through stress, desperation, and have been the only reason I have been able to accomplish the things that I have. Without them, nothing else would be possible.

My final acknowledgment is to my supervisor, Dr. Ruth McPherson. Working with her has been doubtlessly one of the single best decisions of my life. Three years have passed and yet Dr. McPherson continues to help me more in my career and my personal life than I could ever expect of anyone. Truly any success I may claim in the future will be a direct consequence of Dr. McPherson's kindness, generosity, and time.

To everyone, both named and not, I give my most heartfelt thanks.

Christopher B. Cole

Ottawa

May 2016

Dedication

I, Christopher B. Cole, declare that any and all work presented or implied in this thesis is my own original work and has not been previously presented or submitted for academic publication.

Abstract

Background and Rationale: *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, *Genome Wide Association Studies* (GWAS) 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), a linear combination 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.

Purpose and Specific Objectives: This study develops and validates two novel methodologies which integrate meta-data from co-morbid conditions to further elucidate the genetic underpinnings of CAD while improving phenotype prediction.

Materials and Methods: Three PRS were created using summary information from several large consortia. The first *traditional risk score* (TRS) model uses the typical procedure for constructing a PRS, simply summing over all significant variants weighted by a previously identified risk score. The second risk score incorporates information from several co-morbid traits to re-prioritize the rankings of SNPs and introduce new variants into the TRS. The third creates an optimal score which maximizes the P value of association while minimizing environmental noise. These three scores are constructed, validated in a meta analysis of a population composing 5831 cases and 3832 controls, and compared mathematically in terms of their model fit and their predictive accuracy.

Results: Both of our novel methodologies proved significantly better than the TRS at predicting CAD. Additionally, both the TRS and the *cardiometabolic* (CMB) score proved superior to 1000 scores composed of equal number of SNPs, while the *Optimal Polygenic Risk Score* (oPRS) did not. We additionally lend evidence to the assertion that genetic risk for high *Body Mass Index* (BMI) is consistent with a model where many small variants act in concert, rather than a few of larger effect size. We demonstrate the superiority of our scores and show them to be important steps forward in the development of better PRS.

Contents

List of Figures

List of Tables

List of Acronyms

AIC Akaike Information Criterion 24

CAD Coronary Artery Disease iii

PRS Polygenic Risk Score iii

oPRS Optimal Polygenic Risk Score iii

CARDIOGRAMC4D Coronary ARtery DIsease Genome wide Replication and Meta analysis plus The Coronary Artery Disease Genetics 1

GWAS Genome Wide Association Study iii

GLC Global Lipids Consortium 1

GIANT The Genetic Investigation of ANthropometric Traits 1

BMI Body Mass Index iii

MI Myocardial Infarction 2

kb kilobase 3

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 5

InDel insertion/deletion 5

RNA ribonucleic acid 5

SNP single nucleotide polymorphism iii

LD linkage disequilibrium

FWER family wise error rate 8

FDR false disovery rate 8

PRDS positive regression dependence on subsets 8

OR odds ratio 6

TG triglyceride 15

HDLc high density lipoprotein cholesterol 15

LDLc low density lipoprotein cholesterol 15

CMB cardiometabolic iii

AUC area under the curve 22

ROC receiver operator characteristic

OHGS Ottawa Heart Genomics Study 22

TRS traditional risk score iii

A note on notation

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

1. A hat ($\hat{.}$) denotes the estimator of a variable (i.e. $\hat{\beta}$ is the estimator of β).

2. Underlining a variable ($\underline{.}$) 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 (McPherson et al., 2007), obesity (Qi et al., 2008), and schizophrenia (Consortium, 2014). 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 co-morbid conditions using only summary statistics.

In brief, we present preliminary evidence for the use of PRS 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 BMI *per se*. We use this data alongside previously identified variants to construct first a simplistic PRS using only genome wide statistically significant

($P_{Bonferonni} < 0.05$ or $q_{FDR} < 0.05$) variants, then expand our search to variants which may not be as robustly linked to phenotype. (Storey, 1995; Dudbridge, 2013) We use an empirical maximization 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 disease scores in order to better predict CAD in four cohorts.

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. (McPherson and Tybjærg-Hansen, 2014) CAD is known to be a complex genetic disease with heritability estimated by twin studies between 40 and 60%. (Vinkhuyzen et al., 2013) Several important variants have been identified which have been shown to robustly increase risk to CAD by altering lipid transporting pathways, structural collagen bodies, and others factors. (Mega et al., 2015; The CARDIoGRAMplusC4D Consortium, 2015; Visscher et al., 2012; McPherson and Tybjaerg-Hansen, 2016)

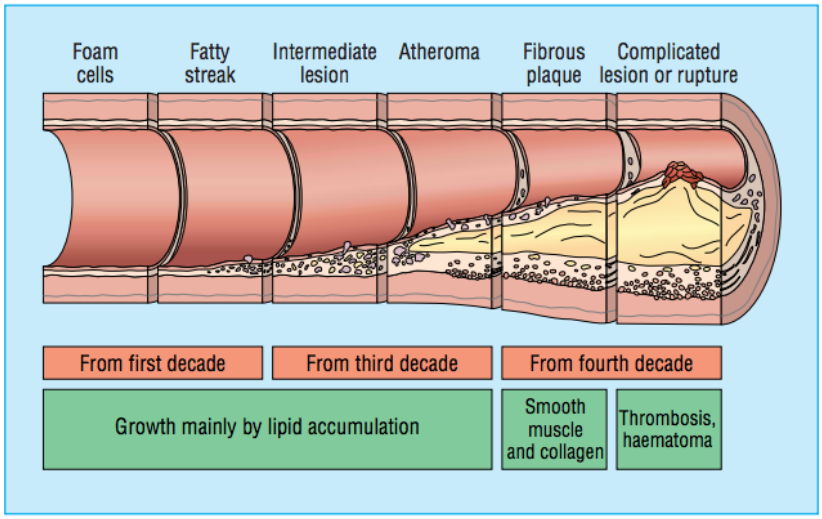


FIGURE 1.1: **Progression of the formation of plaque causing CAD.** Genetic and environmental variants may influence lipid transporting pathways, structural collagen bodies, and others factors which may hasten or influence the development of CAD. Adapted from Gretch 2003.

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 beginnings, initial attempts to understand and explain CAD through genetics were largely unsuccessful. (Visscher et al., 2012) The first variant to be successfully and robustly linked to risk for CAD was the 9p21.3 locus. Discovered by a team of researchers at the University 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 Caucasian individuals. (McPherson and Tybjaerg-Hansen (2016))

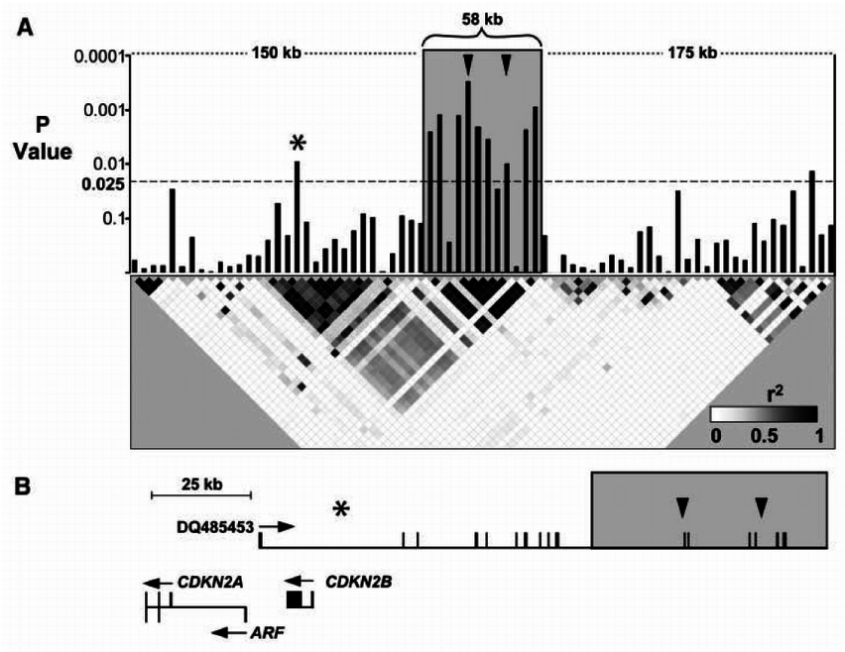


FIGURE 1.2: **Fine mapping of the genomic interval on chromosome 9 associated with Coronary Heart Disease.** (A) SNPs spaced 5 kb apart in the interval extending 175 kb upstream and downstream of rs10757274 and rs2383206 were assayed in 500 cases and 500 controls from the OHS population with GeneChip Human Mapping 500K Array Sets (Affymetrix, Santa Clara, CA). Bars represent P values (determined with χ^2 tests) for differences in allele frequency between cases and controls. Arrowheads indicate rs10757274 and rs2383206. The asterisk represents rs518394. The risk interval is indicated with a gray box. The linkage disequilibrium map indicates pairwise R^2 values. Blocks are shaded on a continuous scale, where white represents an R^2 of 0 and black represents an R^2 of 1. (B) Physical map of the region showing the location of the risk interval (gray box) relative to the noncoding RNA DQ485453 and adjacent genes CDKN2A, ARF, and CDKN2B. Arrowheads indicate rs10757274 and rs2383206, and the asterisk represents rs518394 [see (A)]. Adapted from McPherson et al. (2007)

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. As the dust settles and the low hanging fruit have been picked, common variants have been shown to explain approximately 28% of the heritability of CAD (The CARDIoGRAMplusC4D Consortium, 2015), 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. 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 and assumptions of GWAS must be explored and explained. Genome wide association studies (GWAS) seek to identify associations between individual genotypes and disease phenotypes in a hypothesis free manner. This means that the researcher has no preconceived notions of which areas of the genome may be associated with the phenotype; they are simply looking for anything that appears. 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. These sites, whether they be a change in a single base called a 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.*

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 disease state. 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 further 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 continuous 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, researchers 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 $\underline{\hat{\chi^2}}$.

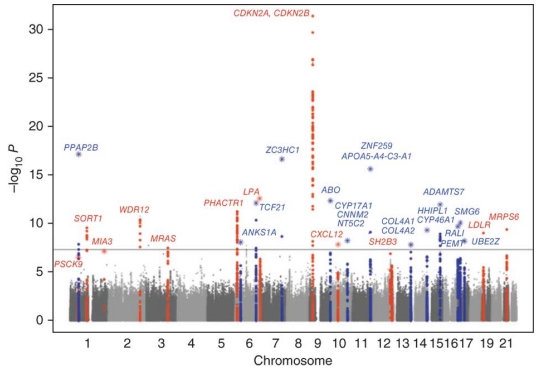


FIGURE 1.3: **Example of a Manhattan plot from a GWAS for CAD performed by Shunkert et al. 2011.** Chromosomal position of SNPs are sorted and plotted on the x axis while the $-\log_{10}$ of their P value derived through either linear regression estimation or χ^2 is plotted on the y axis. Notice that "genome wide significant" loci pass the threshold of $P < 5 \times 10^{-8}$. Those variants which have "towers" supporting them are the most canonically likely to be truly associated; these towers represent loci in strong LD to a causal variant, or one linked to one.

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 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 positives are guaranteed. It is against this backdrop that we recall in any relevant genetic context, M is large, and false positives are almost guaranteed.

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. 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 structure 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 independent 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 discoveries 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 premises.

1.3 Polygenic Prediction of Complex Disease

Referring to the definitions 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 $N(0,1)$ distribution:

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

It has been previously proposed to combine genetic variants in order to crease a score S which encompass 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 take 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 minimizing noise as in 1.4. We denote this as the oPRS.

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.

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

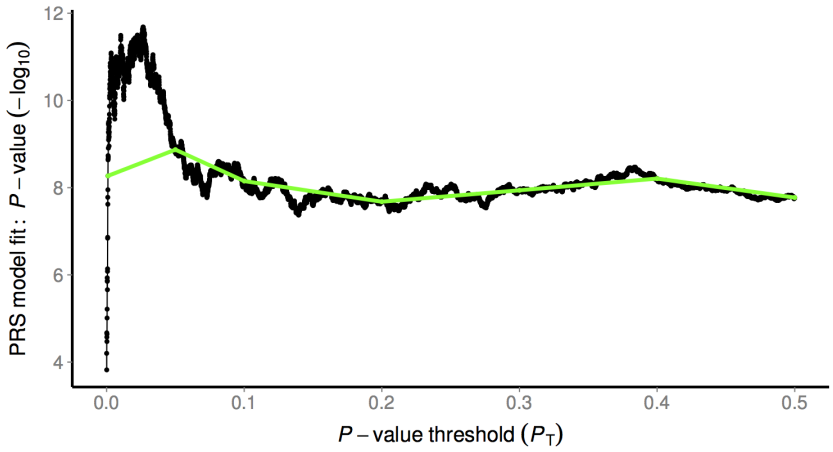


FIGURE 1.4: **High definition plot of optimal PRS association for schizophrenia associated loci predicting major depressive disorder status.** 5000 thresholds were created from $P = 0$ to $P = 0.5$ for SNP inclusion; subsequently created scores were tested for association with major depressive disorder through logistic regression adjusted for principal components. The optimal score is the threshold which maximizes the $-\log_{10} P$ value of association. Adapted from Euesden et al. (2014)

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 estimated 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.

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 Specific Aims

This study plans to address the following issues:

1. Characterize and evaluate the predictive accuracy of the traditional risk score (TRS).

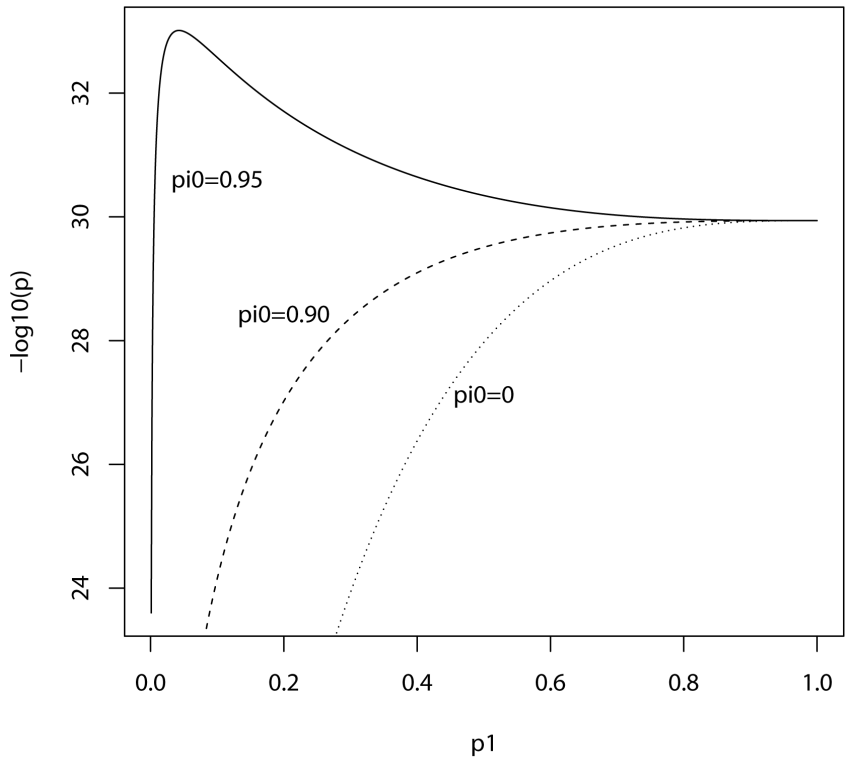


FIGURE 1.5: **Expected $-\log_{10}(P)$ value of linear regression estimate as a function of P-value threshold for selecting markers into PRS.** π_0 refers to the proportion of *true null* markers, that is, markers which have no effect on phenotype. raining sample, 3322 cases and 3587 controls; replication sample, 2687 cases and 2656 controls. Marker panel of 74062 independent SNPs. Variance explained by markers, 28.7%.(Dudbridge, 2013)

Evaluate summary statistics for regression models and perform a meta analysis across cohorts to estimate an overall effect for the model.

2. Propose two novel applications of PRS to better predict CAD. Develop these methods mathematically, validate them in test sets, and compare their performance to the traditional risk score.

3. Perform an exploratory analysis on secondary data to further evaluate the effectiveness of risk scores in predicting CAD.

4. Discuss possible theoretical ramifications of findings, applications, and further research.

Methods 2

2.1 Study Population

There are four major cohorts used as a "test" set in this study, comprising a total $n = 13371$.

Ottawa Heart Genomics Study (OHGS): Details of this cohort have been previously described (Davies et al., 2012). Both cases (1) and controls (0) were recruited from the Lipid Clinic at the University of Ottawa Heart Institute (UOHI). Cases with diabetes melliitus were entirely excluded. Cases were required to have at least one of: a stenosis in a major epicardial vessel of at least 50%; have had a percuteneous intervention (PCI); have had coronary artery bypass surgery (CABG); or have had a myocardial infarction (MI). Earlier studies using this cohort examined the effect of age, and cases were required to be ≤ 55 years old for men and ≤ 65 years old for women. The controls were either healthy elderly patients recruited from the catherization laboratory or the UOHI; they had no stenosis $\geq 50\%$ in any major epicardial vessel and were required to be at minimum 65 years old for men and 70 years old for women. The study protocol was approved by the Human Research Ethics Board of the University of Ottawa Heart Institute and all participants provided informed consent.

Cleveland Clinic (CCGB): Cases and controls from the Cleveland Clinic Cohort followed the same collection procedure as outlined for OHGS except were collected at the catherization laboratory of the Cleveland Clinic.

Duke Cathgen Study (DUKE): Both cases and controls were recruited from the

catherization laboratory at Duke University. Cases were required to have at least one epicardial coronary vessel with $\geq 50\%$ stenosis while being at most 55 years old for males and 65 years old for females. Controls were asymptomatic and required to have ≤ 30 % stenosis in all coronary vessels. Subjects with diabetes melliitus, severe pulmonary hypertension or congenital heart disease were excluded. The study protocol was approved by the ethics committee and all participants provided informed consent.

INTERHEART Cohort (ITH): INTERHEART is a standardized case-control study of acute myocardial infarction from across the world. Only Caucasian participants were analyzed in this study due to issues with differing gene frequencies among ethnicities. Cases – those showing acute MI, were age matched to within 5 years of controls who were community based individuals with no previous history or diagnosis of heart disease and exertional chest pain. The study protocol was approved by the ethics committees in all participating centers and all participants provided informed consent. A full list of ITH investigators is found at http://www.phri.ca/interheart/index2.html.

2.2 Genotyping and Imputation

SNP genotyping of the above cohorts was performed on either Affymetrix 6.0 or 500K chip arrays at the University of Ottawa Heart Institute using the recommended procedure from the manufacturer. They were processed as in Dandona et al. (2010); Schunkert et al. (2011). Imputation was performed using IMPUTE2 and the August 2009 1000 Genomes reference panel. (Howie et al., 2009). Approximately 5.5 million SNP passed quality control measures including info > 0.5, Hardy Weinburg Equilibrium $> 1 \times 10^{-6}$ and missingness $< 10\%$.

2.3 Training Populations

This study additionally comprised two "training" populations which were used to estimate the $\hat{\beta}$ effects necessary for the construction of PRS.

GIANT Consortium: GIANT consortium attempts to identify genetic loci which may modulate human body size, height, and obesity. We use for this study their data on BMI predicting SNP calculated from approximately $n = 123,865$ on close to 2M SNPs. Collection methodologies and specific information is outlined in Speliotes et al. (2010).

Global Lipids Consortium: The Global Lipids Consortium estimates genetic effects in $n = 188,577$ individuals using whole genome and custom genotyping arrays. We use their estimated additive genetic effects for SNPs predicting *triglyceride* (TG), *high density lipoprotein cholesterol* (HDLc), and *low density lipoprotein cholesterol* (LDLc). Collection methodologies and further information are outlined in (Consortium, 2013).

CARDIoGRAMplusC4D: The Coronary ARtery DIsease Genome wide Replication and Meta-analysis (CARDIoGRAM) plus The Coronary Artery Disease (C4D) Genetics is a collaborative effort combining GWAS data from multiple data sources to better estimate potential genetic effects in CAD. Their data represents 22 studies with 22,233 cases and 64,762 controls of European individuals. Further information and collection methodology is outlined in The CARDIoGRAMplusC4D Consortium (2015).

2.4 Polygenic Prediction of CAD

In the following analysis we primarily compare three different methods for constructing PRS $\hat{S}$.

2.4.1 Traditional Risk Score

The first, which we denote as the "traditional risk score", or $\hat{S}_{TRS}$. This score uses the "traditional" approach of only using the highest confidence genome wide significant loci for CAD in the construction of the score. We derive the estimated $\underline{\hat{\beta}}$ effects from (The CARDIoGRAMplusC4D Consortium, 2015), whose methodology is described above. We only use the 212 variants from this section which have been shown to be FDR significant with $q < 0.05$ across the whole genome, as is common practice. Recall from the derivation leading up to equation 1.1 that PRS S for individual n is described as

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

Therefore for this score, we define $\hat{\beta}$ as a vector of length 212 with each of the estimated additive genetic effects derived from CARDIoGRAM plus C4D, and construct estimated score $\hat{S}$ for individual n as

$$\hat{S}_{n,TRS} \equiv \sum_{i=1}^{212} \hat{\beta}_i \mathbf{G}_{n,i}$$

This forms the basis for our first model.

2.4.2 Cardiometabolic Risk Score

The second score which we estimate is a novel derivation. We aim to use genetic information from several co-morbid conditions together to better explain CAD. The motivation is that important signals may be spuriously insignificant in large meta analyses, or simply have too low effect to be accurately categorized as significant; taking information from co-morbid conditions allows researchers a wider span of information to integrate.

We use meta data from four co-morbid conditions to estimate the genetic effects of these traits and **re-prioritize** variants with the intention of creating a minimal score for CAD which better predicts the phenotype.

First, we introduce some new notation. We denote $\underline{\hat{\beta}}_{LDLc}, \underline{\hat{\beta}}_{HDLc}, \underline{\hat{\beta}}_{TG}, \underline{\hat{\beta}}_{BMI}$ as the vectors of estimated effects for LDLc, HDLc, TG, and BMI respectively derived from the training sets outlined in section 2.3.

We separately order variants by their P value and say m^* is the number of genome-significant significant ($q \leq 0.05$) hits found in each study. We take $1 \dots m^*$ from each data set and call this set of variants $\underline{G}^*$ for important genetic effects. We define the set G^* as containing all genetic elements $\mathbf{G}_i$ such that i is a part of our selected significant ordered set $1 \dots m^*$.

$$G^* \equiv \{\mathbf{G}_i | i \in 1 \dots m^*\} \qquad m^* \,|\, q < 0.05$$

We then take all genetic effects $i \in G^*$ and calculate a score based on these variants instead of the 212. We define this new CMB score for any individual n as $\hat{S}_{CMB}$:

$$\hat{S}_{CMB} \equiv \sum_{i \in G^*_{LDLc}} \hat{\beta}_i \mathbf{G}_{n,i} + \sum_{i \in G^*_{HDLc}} \hat{\beta}_i \mathbf{G}_{n,i} + \sum_{i \in G^*_{TG}} \hat{\beta}_i \mathbf{G}_{n,i} + \sum_{i \in G^*_{TG}} \hat{\beta}_i \mathbf{G}_{n,i}$$

We use this new score to predict CAD, with the hypothesis that incorporating several co-morbid conditions will better prioritize variants in order to achieve increased predictive accuracy.

2.4.3 Optimal Cardiometabolic Risk Score

We further extend this $\hat{S}_{CMB}$ using oPRS as introduced in section 1.4. Instead of selecting m^* to be all variants such that $q < 0.05$, we select all P values such that $P < T_o$ where T_o is the optimal threshold found by iterating through P value thresholds for score inclusion.

$$G^* \equiv \{\mathbf{G}_i | i \in 1 \dots m^*\} \qquad m^* \,|\, P < T_o$$

And similarly construct our optimal cardiometabolic risk score as before:

$$\hat{S}_{oCMB} \equiv \sum_{i \in G^*_{LDLc}} \hat{\beta}_i \mathbf{G}_{n,i} + \sum_{i \in G^*_{HDLc}} \hat{\beta}_i \mathbf{G}_{n,i} + \sum_{i \in G^*_{TG}} \hat{\beta}_i \mathbf{G}_{n,i} + \sum_{i \in G^*_{TG}} \hat{\beta}_i \mathbf{G}_{n,i}$$

This forms the new optimal score for testing.

2.5 Statistical Analysis

2.5.1 Construction of PRS

PRS were constructed in both R and Plink v1.90 (https://www.cog-genomics.org/plink2) with validation in Plink v 1.07 (http://pngu.mgh.harvard.edu/ purcell/plink/). Each of approximately 5.5M imputed, post quality control, GWAS were coded as 0,1, or 2, depending on the number of minor alleles present. PRS was calculated as described in previous sections and a normal distribution was validated using both empirical tests (Levene, Shapiro-Wilk) and qualitative observations on qq plots and histograms. Each PRS was confirmed to have a normal distribution, validating one of the assumptions of future models.

Optimal PRS (oPRS) were constructed again using R and Plink 1.90/1.07, though with the assistance of PRSice. (Euesden et al., 2014) To determine optimal thresholds for SNP inclusion, SNPs were ordered by P value and 2500 "slices" or thresholds were created between $P = 0.0001$ and $P = 0.25$. For each of these thresholds, a logistic regression model was created, as detailed below, and the P value of association was calculated. The model was adjusted for the first two principal components and sex. The maximal P value of association was observed and extracted. Based on permutation analysis in the original publication, since our oPRS showed P-values of association less than $\alpha = 0.001$, our association was deemed to be significant overcoming multiple testing correction. Results were graphically displayed

using both base plotting in R and ggplot2. (Wickham, 2009)

2.5.2 Predictive Model

In order to predict CAD with our PRS, covariate adjusted logistic (binomial) regression models were used. Scores were always adjusted for the first two principal components to adjust for population stratification. Nagelkerke's Pseudo R^2, gives an estimation of the scaled explained variance of a logistic model. It is given by

$$R^2 = 1 - \left\{ \frac{L(M_{intercept})}{L(M_{full})} \right\}^{\frac{2}{N}}$$

Where $L(M)$ is the conditional probability of the outcome variable (CAD) given the independent variables and L is the likelihood function operator. A full description of the operator is given in Nagelkerke (1991).

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.

2.6 Computational Resources

All analyses were performed at the Center for Advanced Computing, a large scale Red Hat Enterprise linux parallel computing cluster used to quickly analyze large sets of data. All analysis was parallelized either using inbuilt libraries or OpenMPI standards.

Analyses were performed in R version 3.2.3 (https://www.r-project.org/), Python legacy version 2.7.9 (https://www.python.org/), Plink (http://pngu.mgh.harvard.edu/ purcell/-plink/), and GCTA (http://cnsgenomics.com/software/gcta/).

All analysis was logged and stored securely and anonymously; back ups of all data were made and encrypted.

All analysis was version controlled using git + github and this thesis is entirely reproducible. Any code available upon request.

Results 3

General characteristics of the study population are displayed in Table 3.1.

3.1 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 $5.06 \times 10^{-4} \pm 1.73 \times 10^{-4}$. $\hat{S}_{TRS}$ predicted CAD status better than chance in all cohorts. Area under the receiver operator characteristic curve was calculated for each

	All Participants	Cases	Controls
n	9663	5831	3832
Age[1] (years)	62.8 ± 12.3	56.2 ± 10.1	73.0 ± 7.4
Smoke Current (%)	29.6	36	20
Male (%)	65.3	76.7	47.9
Obese [2] (%)	29	35.1	19.7
BMI (kg/m^2)	28.1 ± 5.3	28.9 ± 5.3	26.7 ± 4.9
TG [3] (mmol/L)	1.46 ± 1.47	1.66 ± 1.70	1.18 ± 0.99
HDLc[3] (mmol/L)	1.27 ± 0.44	1.13 ± 0.39	1.46 ± 0.44
LDLc[3] (mmol/L)	3.29 ± 1.08	3.18 ± 1.17	3.43 ± 0.93

Table 3.1: General Population description. All values are expressed as mean ± one standard deviation unless otherwise noted. [1] Age represents age at consent for controls and age at diagnosis for cases [2] Obesity is defined as having a BMI of greater or equal to 30 kg/m2 at time of collection [3]T G (triglyceride), LDLc (low density lipoprotein cholesterol), HDLc (high density lipoprotein cholesterol).

model and meta-analyzed. The resulting AUC ± 95% confidence intervals are displayed in Figure 3.1.

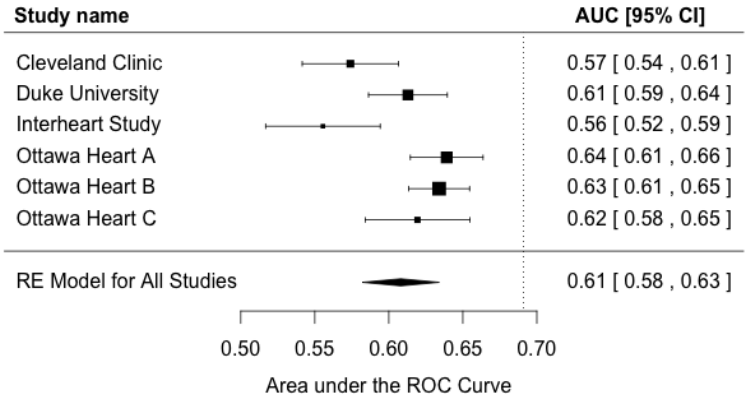


FIGURE 3.1: **Random effects meta analysis of AUC ROC from six cohorts.** Logistic regression models were constructed with $\hat{S}_{TRS}$ and the first two principal components adjusting for population stratification were used to predict CAD. Various thresholds for false positives and true positives were used to construct ROC curves. Area under these curves were estimated along with error. The associated effects were analyzed assuming cohorts were random effects and an overall effect was derived.

Logistic regression was used to predict observations following the model. Summary statistics from the models employed are displayed in Table 3.2

$$CAD = X_0 + \hat{\beta}_1 \hat{S}_{TRS} + \hat{\beta}_2 PC_1 + \hat{\beta}_3 PC_3 + \epsilon$$

Cohort	OR	SE	R^2	AIC	AUC
Cleveland Clinic	153.899	36.427	0.0162	1877.724	0.574
Duke University	255.785	33.443	0.0471	2335.567	0.613
Interheart	83.956	46.459	0.0169	1172.994	0.555
OHGS A2	310.497	34.527	0.0765	2558.277	0.639
OHGS B2	259.593	24.945	0.0742	3681.768	0.634
OHGS C2	276.967	45.920	0.0549	1318.919	0.619

Table 3.2: **Summary statistics from Logistic association model.** $\hat{S}_{TRS}$ along with the first two principal components to adjust for population stratification were used to predict CAD. OR corresponds to the odds ratio of $\hat{S}_{TRS}$ along with its standard error (SE). R^2 corresponds to NagelKerke's Pseudo-R^2, while AIC corresponds to Akaike Information Criterion, a measure of model fit. AUC corresponds to the area under the ROC curve as derived in the pROC package in R.

An *area under the curve* (AUC) ≥ 0.5 indicates that the model predicts CAD better than chance. The overall random effects meta analyzed AUC was 0.61 $\pm$0.03 for $\hat{S}_{TRS}$, with an average NagelKerke's Pseudo-R^2 of 0.047. Interheart was the worst fit model, with NagelKerke's Pseudo-$R^2 = 0.017$ and AUC $= 0.555$, barely predicting above chance. This is in contrast to the OHGS cohorts, which consistently fit better than Cleveland Clinic, Duke, or Interheart. 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 randomly selected loci in 1000 bootstraps in Figure 3.1.

The score remains significantly associated when adjusted for individual's sex, a known cardiovascular risk factor. The predictive accuracy of the model significantly increases after inclusion of sex, as would be expected. The random effects meta analysis AUC for all six cohorts becomes $0.69[0.62, 0.76]$.

In *Ottawa Heart Genomics Study* (OHGS) B2, sex significantly ($P = 0.00226$) interacts with the effect of the risk score. This shows that sex modulates the effect of genetics in this cohort. The remainder of the cohorts do not interact.

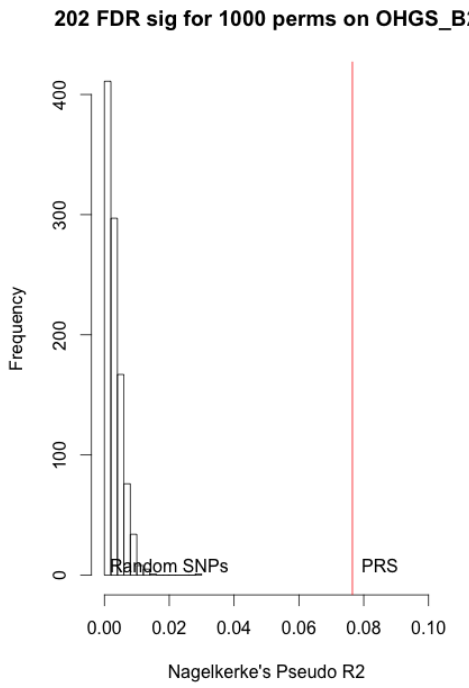


FIGURE 3.2: $\hat{S}_{TRS}$ **predicts CAD significantly better than 1000 PRS constructed with an equal number of SNPs.** 202 SNPs were randomly selected from post quality control imputed SNPs and a PRS was constructed using summary information from The CARDIoGRAMplusC4D Consortium (2015) and Nagelkerke's Pseudo R^2 was plotted against frequency. The red line denote the Nagelkerke's R^2 of the true TRS which predicts significantly $P \approx 0$ better than the random model.

These results are in line with literature values. Those in the top quintile of the PRS were 70% more likely to have CAD than not (2045 cases vs 1198 controls). Similarly, those in the bottom quintile were 5% less likely to be diagnosed with CAD than not (1576 cases vs 1659 controls).

3.2 Cardiometabolic Risk Score

We add in each co-morbid score in a stepwise manner. $\hat{S}_{CMB;1}$ uses just the information available for CAD and is equivalent to the above section. $\hat{S}_{CMB;2}$ uses information from CAD and BMI, while $\hat{S}_{CMB;3}$ uses CAD, BMI, and LDLc. $\hat{S}_{CMB;4}$ uses information from CAD, BMI, LDLc, and TG, while $\hat{S}_{CMB;5}$ uses CAD, BMI, LDLc, TG, and HDLc. Note that in this section we restrict our analysis to the OHGS cohorts along with Cleveland Clinic due to the lack of quality in lipid data present in the other cohorts.

The summary statistics of this analysis are presented in Table 3.2. Higher genetic risk scores were uniformly and significantly ($P < 2.2 \times 10^{-16}$) associated with increased risk of CAD.

Score	Cohort	OR	SE	R2	AIC	AUC
$\hat{S}_{CMB;1}$	OHGS A2	310.498	34.527	0.077	2558.277	0.639
	OHGS B2	259.593	24.945	0.074	3681.768	0.634
	OHGS C2	276.967	45.920	0.055	1318.919	0.619
	Cleveland Clinic	153.899	36.428	0.017	1877.724	0.574
$\hat{S}_{CMB;2}$	OHGS_A2	252.964	25.292	0.091	2547.617	0.656
	OHGS_B2	267.348	22.487	0.092	3654.756	0.650
	OHGS_C2	280.924	38.981	0.077	1300.585	0.646
	Cleveland	337.371	35.182	0.084	1791.596	0.666
$\hat{S}_{CMB;3}$	OHGS_A2	236.0145	26.173	0.077	2570.063	0.643
	OHGS_B2	255.740	24.099	0.077	3688.513	0.637
	OHGS_C2	284.979	41.161	0.071	1305.337	0.640
	Cleveland	344.918	37.879	0.076	1802.024	0.655
$\hat{S}_{CMB;4}$	OHGS_A2	266.505	29.249	0.078	2568.062	0.644
	OHGS_B2	270.889	26.455	0.073	3697.51	0.633
	OHGS_C2	298.983	44.838	0.066	1309.38	0.634
	Cleveland	372.189	41.802	0.072	1806.542	0.651
$\hat{S}_{CMB;5}$	OHGS_A2	297.017	34.105	0.072	2576.54	0.640
	OHGS_B2	299.070	30.767	0.067	3709.488	0.628
	OHGS_C2	323.917	50.903	0.061	1313.958	0.628
	Cleveland	410.996	48.574	0.065	1815.742	0.644

Table 3.3: **Summary statistics from Logistic association model.** $\hat{S}_{CMD;x}$ along with the first two principal components to adjust for population stratification were used to predict CAD. OR corresponds to the odds ratio of $\hat{S}_{TRS}$ along with its standard error (SE). R^2 corresponds to NagelKerke's Pseudo-R^2, while AIC corresponds to Akaike Information Criterion, a measure of model fit. AUC corresponds to the area under the ROC curve as derived in the pROC package in R.

In permutation analyses, each of the scores performed significantly better than an equivalent number of randomly selected SNPs in 1000 bootstraps. The random effects meta analysis AUC values were $0.65[0.64, 0.67], 0.64[0.63, 0.66], 0.64[0.63, 0.65], 0.63[0.62, 0.65]$ for scores 1 through 5 respectively. Interestingly, the more SNPs added in, the worse the model was at predicting the phenotype. The *Akaike Information Criterion* (AIC) is also uniformly smaller in score 2 than in subsequent scores, meaning that this model fit the data the best. 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.

The score maintained its significance even after inclusion of biologically relevant covariates such as gender and smoking status. Additionally, the predictive accuracy of the score increased substantially after the inclusion of these covariates. After inclusion of individual's sex, a known cardiovascular risk factor, predictive accuracy increased substantially. The random effects meta analysis AUC values were increased to $0.69[0.62, 0.76], 0.74[0.67, 0.81], 0.73[0.66, 0.81], 0.73[0.66, 0.81], 0.7$ for $\hat{S}_{CMB;1}$ through $\hat{S}_{CMB;5}$ respectively. Again it appears as though $\hat{S}_{CMB;2}$ is the best predictive model.

We additionally investigated whether sex significantly modulates the effect of $\hat{S}_{CMB;\cdot}$ and found no significant interactions.

3.2.1 Optimal Cardiometabolic Risk Score

When optimal scores were derived, the thresholds described in Table 3.4 were observed.

	LDLc	**HDLc**	**TG**	**BMI**
OHGS_A2	0.0001	0.0055	0.1743	1
OHGS_B2	0.2484	0.0999	0.0002	1
OHGS_C2	0.1299	0.0085	0.1528	1
CCGB_2	0.1807	0.2039	0.004	1

Table 3.4: **Optimal PRS P value thresholds (T_o) derived for each trait in each cohort.** 2500 thresholds were created between $P = 0.0001$ and $P = 0.25$ and used as inclusion threshold T for PRS. These SNPs were used to construct a PRS, which was used alongside the first two principal components and sex as covariates to predict CAD. Their respective P-values of association were recorded and the maximal $-\log_{10} P$-value of association was used as the optimal threshold T_0.

The optimal P value cutoff threshold T_o for BMI was found to be 1 for all cases as demonstrated in 3.2.1. We discuss this result further below, however, for now we exclude BMI from the analysis of optimal risk scores, and opt to use just scores for the lipid traits.

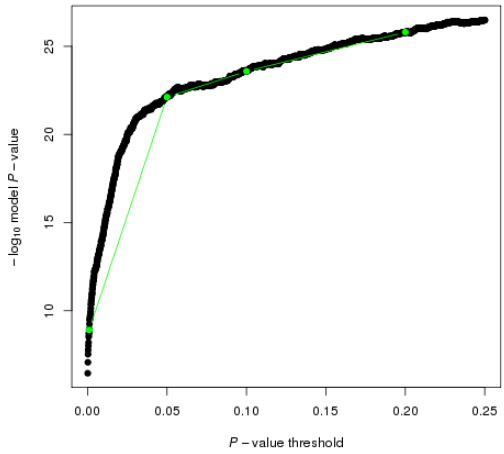


FIGURE 3.3: **High definition assocaition plot for BMI SNPs with CAD.**2500 thresholds were created between $P = 0.0001$ and $P = 0.25$ and used as inclusion threshold T for PRS. These SNPs were used to construct a PRS, which was used alongside the first two principal components and sex as covariates to predict CAD. Their respective P-values of association were recorded and the maximal $-\log_{10} P$-value of association was used as the optimal threshold T_0. There was no maximal $-\log_{10} P$ value threshold less than 0.25, consistent with the model shown in figure 1.4 with low π_0, or proportion of truly null SNPs.

Constructing a logistic regression model, we find that our cumulative score (combining optimal scores from all three traits) is significantly ($P < 2.2 \times 10^{-16}$.) associated with CAD status, and those who have a higher PRS tend to have higher risk for CAD. We detail summary statistics for this association in table 3.2.1.

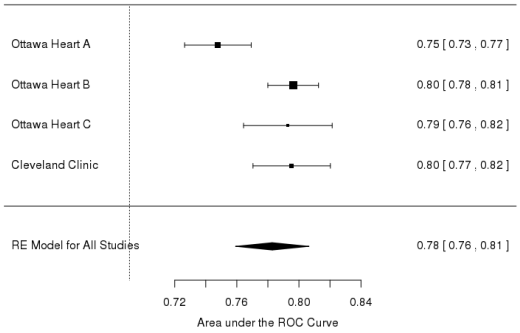
Study	n_{SNPs}	OR	SE	R^2	AIC	ROC
OHGS_A2	439971	2.2×10^4	1.3×10^3	0.245	2313	0.748
OHGS_B2	790128	3.42×10^4	1.5×10^3	0.338	3061.9	0.796
OHGS_C2	622050	3.34×10^4	2.5×10^3	0.301	1095.1	0.793
CCGB_2	847335	3.78×10^4	2.3×10^3	0.284	1517.21	0.795

Table 3.5: **Summary statistics from Logistic association model.** $\hat{S}_{oCMD}$ along with the first two principal components to adjust for population stratification were used to predict CAD. OR corresponds to the odds ratio of $\hat{S}_{TRS}$ along with its standard error (SE). R^2 corresponds to NagelKerke's Pseudo-R^2, while AIC corresponds to Akaike Information Criterion, a measure of model fit. AUC corresponds to the area under the ROC curve as derived in the pROC package in R.

The oPRS, comprising optimal scores for LDLc, HDLc, and TG predicts CAD with a very high accuracy, explaining between 24.5 and 33.8 percent of variance in CAD. The random effects meta analysis of AUC of the ROC curve reveals an overall AUC of $0.78[0.75, 0.81]$,

an extremely high value (Figure 3.2.1).

FIGURE 3.4: Random effects meta analysis of oPRS predicting risk for CAD



Though the AUC varies significantly between the cohorts, it is significantly higher than any observed in either the TRS or the CMB scores. However, as noted in table 3.2.1, each oPRS comprises several hundred thousand SNPs, and it is difficult to assess whether the increase in predictive accuracy is simply a consequence of increasing the number of SNPs used to predict the phenotype, or whether it is a true biological occurrence. When an equal number of SNPs were selected at random in 1000 bootstraps, it was found that the R^2 of our oPRS was not significantly $P > 0.05$ different than that obtained through permutation.

3.2.2 Model Comparisons

When the cardiometabolic $\hat{S}_{CMB;2}$, determined to be the best predictive candidate score for CAD was compared against the TRS score in 1000 random bootstraps in all cohorts combined, it was found to have a significantly larger AUC ($P < 2.2 \times 10^{-16}$). A similar result was found when the oPRS was compared to both the TRS and the CMB scores.

Discussion 4

In this thesis we have introduced a novel method for using summary information from co-morbid conditions identified through GWAS for the prediction of complex disease. We validate this technique in a meta analysis of four cohorts.

Previously, polygenic risk scores (PRS) have been used to predict complex diseases and to measure to genetic overlap between conditions. However, often researchers restrict themselves to using the loci which have the highest confidence – those which reach genome-wide significance. Theoretical work disputes this notation, indicating that many low or modest effect variants may be hidden in the region of P-values classically called non-significant. To help better identify candidate variants which may be useful in predicting a condition, we use variants shown to be linked to co-morbid conditions in order to construct a cardiometabolic risk score for coronary artery disease (CAD).

We use meta data from the recent large scale meta-analysis conducted by the CARDIOGRAMC4D consortium to construct the "traditional" PRS for CAD in our four cohorts. The score performs as is expected, significantly predicting CAD with a relatively low explained variance, with NagelKereke's Pseudo-R^2 averaging 4.7% between the cohorts. This model performs decently at predicting CAD as well, with a meta analyzed overall AUC of the ROC curve of 0.61 with 95% confidence interval between 0.58 and 0.63. Adding biologically relevant covariates increases the predictive accuracy whilst the PRS maintains its significance. Our new model, adding together the PRS from co-morbid conditions as in Section 2.4.2, showed

an improvement over the traditional model in all cases. However, the largest (and only statistically significant) difference occurs between the traditional risk score and the score incorporating BMI loci. Interestingly, though the subsequent scores contain all SNPs present in the second BMI score, the increased noise makes the improvement over the TRS less pronounced. This is directly conflicting with the notion that increasing the number of SNPs used to construct the PRS will usually increase the predictive accuracy, even if it does not increase the significance of the model.

We additionally showed that though our method involves additional SNPs to the PRS, the increase in predictive accuracy occurs that which would be expected by chance, as shown in figure 3.1. This was not the case with the optimal model, as will be described below.

We can only speculate on the true reasons for the observed pattern in PRS association. LDLc, HDLc, and TG are well known risk factors for CAD, and perhaps these results shed some light on the underlying relationship between these phenotypes and CAD. It has been well documented that CAD and BMI share a substantial genetic relationship: in a recent study, the r_G, or genetic correction, between these two phenotypes was estimated to be $r_G = 0.66$ with standard error 0.2, $P = 5 \times 10^{-4}$. (Cole et al., 2015b) Full results for our cohorts are displayed in table 4.1.

Study	**Sample Size**	r_G	**SE**
OHGS_A2	3578	0.566	0.244
OHGS_B2	5718	1.00	0.787
OHGS_C2	2816	0.778	0.713
CCGB_2	4365	0.868	0.496

Table 4.1: **Genetic Correlations between Obesity and CAD.** Bivariate generalized restricted maximum likelihood (GREML) was used to estimated the correlation between the genetic relationship matrices (GRM) for CAD and BMI. Adapted from Cole et al. (2015b)

It is evident that BMI and CAD share substantial genetic pleiotropy. Similar analyses for LDLc, HDLc, and TG have not been conducted. It has recently been demonstrated that adiposity, that is, BMI *per se*, significantly interacted with a PRS for dyslpidemia. (Cole et al., 2014) This adds to a body of evidence suggesting multiple gene by environment interactions may play crucial roles in dyslipidemia risk. Cole et al. (2015a) Thus, it remains an open question whether the genetic predisposition to BMI is interacting with the genetic predisposition to lipids such that the results observed are unexpected.

The combined optimal model was shown to be highly effective at predicting CAD,

explaining between 24.5 and 33.8 percent of variance in the phenotype. Recall that estimated heritability h^2 is approximately 40%, meaning that our score explains between 60% and 80% of the variance in CAD attributable to genetics. However, it was found that predictive accuracy explained by the score may be attributable to the number of SNPs which were used in its construction, as the NagelKerke's Pseudo-R^2 was insignificantly different than 1000 bootstrap permutations using the same number of SNPs. Though it performed well, it is difficult to compare because of this reason.

Additionally, we observed that the optimal solution for BMI was trivial; that is, no P-value threshold $T_o < 1$ was optimal. Referring to Figure 1.4, this may lend evidence to the assertion to that there is a relatively large number of small effect size SNPs influence BMI rather than a smaller set of large effect size SNPs. This would correspond to a lower π_0, or proportion of true null SNPs and no maxima < 1 for $P \in [0, 1]$.

Because of this, it is difficult to compare the oPRS to either the TRS or the CMB scores.

4.1 Future Directions

Our study has provided a solid foundation upon which to further elucidate the roll which polygenic risk scores may play in both the prediction and understanding of CAD. Future research may examine the exact algorithm by which the cardiometabolic score is calculated. Specifically, the roll and weighting of duplicated variants must be investigated to see if predictive variants which occur in more than one PRS should be upweighted or prioritized to better facilitate phenotype prediction. Additionally, the roll of variant independence must be examined in depth; currently the optimal score does not include any reservations about variant dependency conditions. Whether or not this has impacted the predictive accuracy remains to be seen.

Additionally, the trends observed must be validated mathematically; we are unsure of the exact mechanics and expected outcomes when scores are combined such as we have done. More research must be done to better understand the benefits and drawbacks of combining information in this way.

Conclusions 5

In this thesis submitted to partially fulfill the requirements of an honours undergraduate degree in biomedical science, we introduce and validate a novel methodology for combining summary information from GWAS in order to better predict complex disease from co-morbid conditions and phenotypes. We introduce GWAS and their statistical properties, leading into a discussion of PRS. We build upon these definitions to create our new model, which combines scores across several conditions. We find that our new CMB score predicted CAD significantly better than did the TRS. We additionally constructed an optimal PRS (oPRS), which predicted between 60 and 80 % of the variance in CAD attributable to genetics, but was not significantly different from a random bootstrap. We have also shown that the characteristics of oPRS P value threshold selection in BMI are consistent with those observed when a relatively high number of low effect SNPs affect the phenotype, lending evidence to this notion.

Our novel scoring technique represents a significant improvement over the traditional polygeneic prediction of CAD. However, much theoretical research is needed to validate and explore the trends observed. We believe that by pushing further into this area of knowledge, eventually PRS, a widely used but often naive methodology, will be improved to the point of clinical utility. Using further iterations of methods like ours, eventually clinicians may be able to predict patient's clinical phenotypes with a high degree of accuracy from a small number of SNPs.

Bibliography

C. B. Cole, M. Nikpay, P. Lau, a. F. R. Stewart, R. W. Davies, G. a. Wells, R. Dent, and R. McPherson. Adiposity significantly modifies genetic risk for dyslipidemia. *The Journal of Lipid Research*, 55(11):2416–2422, sep 2014. ISSN 0022-2275. doi: 10.1194/jlr.P052522. URL http://www.jlr.org/cgi/doi/10.1194/jlr.P052522.

Christopher B Cole, Majid Nikpay, and Ruth McPherson. GeneâĂŞenvironment interaction in dyslipidemia, 2015a. URL http://journals.lww.com/co-lipidology/Abstract/2015/04000/Gene{_}environment{_}interaction{_}in{_}dyslipidemia.10.aspx.

Christopher B Cole, Majid Nikpay, Alexandre Fr Stewart, and Ruth McPherson. Increased genetic risk for obesity in premature coronary artery disease. *European journal of human genetics : EJHG*, jul 2015b. ISSN 1476-5438. doi: 10.1038/ejhg.2015.162. URL http://dx.doi.org/10.1038/ejhg.2015.162.

Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nat Genet*, 45(11):1274–1283, nov 2013. ISSN 1061-4036. URL http://dx.doi.org/10.1038/ng.2797http://10.1038/ng.2797http://www.nature.com/ng/journal/v45/n11/abs/ng.2797.html{#}supplementary-information.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, jul 2014. ISSN 0028-0836. URL http://dx.doi.org/10.1038/nature13595http://10.1038/nature13595http://www.nature.com/nature/journal/v511/n7510/abs/nature13595.html{#}supplementary-information.

Sonny Dandona, Alexandre F R Stewart, Li Chen, Kathryn Williams, Derek So, Ed O'Brien, Christopher Glover, Michel LeMay, Olivia Assogba, Lan Vo, Yan Qing Wang, Marino

Labinaz, George A Wells, Ruth McPherson, and Robert Roberts. Gene Dosage of the Common Variant 9p21 Predicts Severity of Coronary Artery Disease. *Journal of the American College of Cardiology*, 56(6):479–486, aug 2010. ISSN 0735-1097. doi: http://dx.doi.org/10.1016/j.jacc.2009.10.092. URL http://www.sciencedirect.com/science/article/pii/S0735109710019583.

Robert W Davies, George A Wells, Alexandre F R Stewart, Jeanette Erdmann, Svati H Shah, Jane F Ferguson, Alistair S Hall, Sonia S Anand, Mary S Burnett, Stephen E Epstein, Sonny Dandona, Li Chen, Janja Nahrstaedt, Christina Loley, Inke R König, William E Kraus, Christopher B Granger, James C Engert, Christian Hengstenberg, H-Erich Wichmann, Stefan Schreiber, W H Wilson Tang, Stephen G Ellis, Daniel J Rader, Stanley L Hazen, Muredach P Reilly, Nilesh J Samani, Heribert Schunkert, Robert Roberts, and Ruth McPherson. A Genome Wide Association Study for Coronary Artery Disease Identifies a Novel Susceptibility Locus in the Major Histocompatibility Complex. *Circulation. Cardiovascular genetics*, 5(2):217–225, apr 2012. ISSN 1942-325X. doi: 10.1161/CIRCGENETICS.111.961243. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3335297/.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44(3):837–845, 1988. ISSN 0006341X, 15410420. doi: 10.2307/2531595. URL http://www.jstor.org/stable/2531595.

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.

Bryan N Howie, Peter Donnelly, and Jonathan Marchini. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS*

Genet, 5(6):e1000529, 2009. doi: 10.1371/journal.pgen.1000529. URL http://dx.doi.org/10.1371{%}2Fjournal.pgen.1000529.

Ruth McPherson and Anne Tybjærg-Hansen. Genetics of coronary artery disease. *Circulation Research*, 114(12):1890–1903, 2014. ISSN 15244571. doi: 10.1161/CIRCRESAHA.114.302692. URL http://circres.ahajournals.org/content/114/12/1890.abstract?etoc.

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.

Ruth McPherson, Alexander Pertsemlidis, Nihan Kavaslar, Alexandre Stewart, Robert Roberts, David R Cox, David A Hinds, Len A Pennacchio, Anne Tybjaerg-Hansen, Aaron R Folsom, Eric Boerwinkle, Helen H Hobbs, and Jonathan C Cohen. A Common Allele on Chromosome 9 Associated with Coronary Heart Disease. *Science (New York, N.Y.)*, 316(5830):1488–1491, jun 2007. ISSN 0036-8075. doi: 10.1126/science.1142447. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2711874/.

Jessica L Mega, Nathan O Stitziel, J Gustav Smith, Daniel I Chasman, Mark J Caulfield, James J Devlin, Francesco Nordio, Craig L Hyde, Christopher P Cannon, Frank M Sacks, Neil R Poulter, Peter S Sever, Paul M Ridker, Eugene Braunwald, Olle Melander, Sekar Kathiresan, and Marc S Sabatine. Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *The Lancet*, 385:2264–2271, 2015. ISSN 01406736. doi: 10.1016/S0140-6736(14)61730-X. URL http://linkinghub.elsevier.com/retrieve/pii/S014067361461730X.

N. J D Nagelkerke. A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692, 1991. ISSN 00063444. doi: 10.1093/biomet/78.3.691.

Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy a Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, aug 2006. ISSN 1061-4036. doi: 10.1038/ng1847. URL http://www.ncbi.nlm.nih.gov/pubmed/16862161.

Lu Qi, Kihwa Kang, Cuilin Zhang, Rob M van Dam, Peter Kraft, David Hunter, Chih-Hao Lee, and Frank B Hu. Fat MassâĂŞand Obesity-Associated (FTO) Gene Variant Is Associated With Obesity: Longitudinal Analyses in Two Cohort Studies and Functional Test . *Diabetes*, 57(11):3145–3151, nov 2008. doi: 10.2337/db08-0006. URL `http://diabetes.diabetesjournals.org/content/57/11/3145.abstract`.

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics*, 12(1):77, jan 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-77. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3068975{&}tool=pmcentrez{&}rendertype=abstract`.

Heribert Schunkert, Inke R König, Sekar Kathiresan, Muredach P Reilly, Themistocles L Assimes, Hilma Holm, Michael Preuss, Alexandre F R Stewart, Maja Barbalic, Christian Gieger, Devin Absher, Zouhair Aherrahrou, Hooman Allayee, David Altshuler, Sonia S Anand, Karl Andersen, Jeffrey L Anderson, Diego Ardissino, Stephen G Ball, Anthony J Balmforth, Timothy A Barnes, Diane M Becker, Lewis C Becker, Klaus Berger, Joshua C Bis, S Matthijs Boekholdt, Eric Boerwinkle, Peter S Braund, Morris J Brown, Mary Susan Burnett, Ian Buysschaert, Cardiogenics Carlquist John F, Li Chen, Sven Cichon, Veryan Codd, Robert W Davies, George Dedoussis, Abbas Dehghan, Serkalem Demissie, Joseph M Devaney, Ron Do, Angela Doering, Sandra Eifert, Nour Eddine El Mokhtari, Stephen G Ellis, Roberto Elosua, James C Engert, Stephen E Epstein, Ulf de Faire, Marcus Fischer, Aaron R Folsom, Jennifer Freyer, Bruna Gigante, Domenico Girelli, Solveig Gretarsdottir, Vilmundur Gudnason, Jeffrey R Gulcher, Eran Halperin, Naomi Hammond, Stanley L Hazen, Albert Hofman, Benjamin D Horne, Thomas Illig, Carlos Iribarren, Gregory T Jones, JWouter Jukema, Michael A Kaiser, Lee M Kaplan, John J P Kastelein, Kay-Tee Khaw, Joshua W Knowles, Genovefa Kolovou, Augustine Kong, Reijo Laaksonen, Diether Lambrechts, Karin Leander, Guillaume Lettre, Mingyao Li, Wolfgang Lieb, Patrick Linsel-Nitschke, Christina Loley, Andrew J Lotery, Pier M Mannucci, Seraya Maouche, Nicola Martinelli, Pascal P McKeown, Christa Meisinger, Thomas Meitinger, Olle Melander, Pier Angelica Merlini, Vincent Mooser, Thomas Morgan, Thomas W Mühleisen, Joseph B Muhlestein, Thomas Münzel, Kiran Musunuru, Janja Nahrstaedt, Christopher P Nelson, Markus M Nöthen, Oliviero Olivieri, Riyaz S Patel, Chris C

Patterson, Annette Peters, Flora Peyvandi, Liming Qu, Arshed A Quyyumi, Daniel J Rader, Loukianos S Rallidis, Catherine Rice, Frits R Rosendaal, Diana Rubin, Veikko Salomaa, M Lourdes Sampietro, Manj S Sandhu, Eric Schadt, Arne Schäfer, Arne Schillert, Stefan Schreiber, Jürgen Schrezenmeir, Stephen M Schwartz, David S Siscovick, Mohan Sivananthan, Suthesh Sivapalaratnam, Albert Smith, Tamara B Smith, Jaapjan D Snoep, Nicole Soranzo, John A Spertus, Klaus Stark, Kathy Stirrups, Monika Stoll, W H Wilson Tang, Stephanie Tennstedt, Gudmundur Thorgeirsson, Gudmar Thorleifsson, Maciej Tomaszewski, Andre G Uitterlinden, Andre M van Rij, Benjamin F Voight, Nick J Wareham, George A Wells, H-Erich Wichmann, Philipp S Wild, Christina Willenborg, Jaqueline C M Witteman, Benjamin J Wright, Shu Ye, Tanja Zeller, Andreas Ziegler, Francois Cambien, Alison H Goodall, L Adrienne Cupples, Thomas Quertermous, Winfried März, Christian Hengstenberg, Stefan Blankenberg, Willem H Ouwehand, Alistair S Hall, Panos Deloukas, John R Thompson, Kari Stefansson, Robert Roberts, Unnur Thorsteinsdottir, Christopher J O'Donnell, Ruth McPherson, Jeanette Erdmann, Nilesh J Samani, and for the CARDIoGRAM Consortium. Large-scale association analyses identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics*, 43(4):333–338, mar 2011. ISSN 1061-4036. doi: 10.1038/ng.784. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3119261/`.

Elizabeth K Speliotes, Cristen J Willer, Sonja I Berndt, Keri L Monda, Gudmar Thorleifsson, Anne U Jackson, Hana Lango Allen, Cecilia M Lindgren, Jian'an Luan, Reedik Mägi, Joshua C Randall, Sailaja Vedantam, Thomas W Winkler, Lu Qi, Tsegaselassie Workalemahu, Iris M Heid, Valgerdur Steinthorsdottir, Heather M Stringham, Michael N Weedon, Eleanor Wheeler, Andrew R Wood, Teresa Ferreira, Robert J Weyant, Ayellet V Segré, Karol Estrada, Liming Liang, James Nemesh, Ju-Hyun Park, Stefan Gustafsson, Tuomas O Kilpeläinen, Jian Yang, Nabila Bouatia-Naji, Tõnu Esko, Mary F Feitosa, Zoltán Kutalik, Massimo Mangino, Soumya Raychaudhuri, Andre Scherag, Albert Vernon Smith, Ryan Welch, Jing Hua Zhao, Katja K Aben, Devin M Absher, Najaf Amin, Anna L Dixon, Eva Fisher, Nicole L Glazer, Michael E Goddard, Nancy L Heard-Costa, Volker Hoesel, Jouke-Jan Hottenga, Åsa Johansson, Toby Johnson, Shamika Ketkar, Claudia Lamina, Shengxu Li, Miriam F Moffatt, Richard H Myers, Narisu Narisu, John R B Perry, Marjolein J Peters, Michael Preuss, Samuli Ripatti, Fernando Rivadeneira, Camilla

Sandholt, Laura J Scott, Nicholas J Timpson, Jonathan P Tyrer, Sophie van Wingerden, Richard M Watanabe, Charles C White, Fredrik Wiklund, Christina Barlassina, Daniel I Chasman, Matthew N Cooper, John-Olov Jansson, Robert W Lawrence, Niina Pellikka, Inga Prokopenko, Jianxin Shi, Elisabeth Thiering, Helene Alavere, Maria T S Alibrandi, Peter Almgren, Alice M Arnold, Thor Aspelund, Larry D Atwood, Beverley Balkau, Anthony J Balmforth, Amanda J Bennett, Yoav Ben-Shlomo, Richard N Bergman, Sven Bergmann, Heike Biebermann, Alexandra I F Blakemore, Tanja Boes, Lori L Bonnycastle, Stefan R Bornstein, Morris J Brown, Thomas A Buchanan, Fabio Busonero, Harry Campbell, Francesco P Cappuccio, Christine Cavalcanti-Proença, Yii-Der Ida Chen, Chih-Mei Chen, Peter S Chines, Robert Clarke, Lachlan Coin, John Connell, Ian N M Day, Martin den Heijer, Jubao Duan, Shah Ebrahim, Paul Elliott, Roberto Elosua, Gudny Eiriksdottir, Michael R Erdos, Johan G Eriksson, Maurizio F Facheris, Stephan B Felix, Pamela Fischer-Posovszky, Aaron R Folsom, Nele Friedrich, Nelson B Freimer, Mao Fu, Stefan Gaget, Pablo V Gejman, Eco J C Geus, Christian Gieger, Anette P Gjesing, Anuj Goel, Philippe Goyette, Harald Grallert, Jürgen Gräßler, Danielle M Greenawalt, Christopher J Groves, Vilmundur Gudnason, Candace Guiducci, Anna-Liisa Hartikainen, Neelam Hassanali, Alistair S Hall, Aki S Havulinna, Caroline Hayward, Andrew C Heath, Christian Hengstenberg, Andrew A Hicks, Anke Hinney, Albert Hofman, Georg Homuth, Jennie Hui, Wilmar Igl, Carlos Iribarren, Bo Isomaa, Kevin B Jacobs, Ivonne Jarick, Elizabeth Jewell, Ulrich John, Torben Jørgensen, Pekka Jousilahti, Antti Jula, Marika Kaakinen, Eero Kajantie, Lee M Kaplan, Sekar Kathiresan, Johannes Kettunen, Leena Kinnunen, Joshua W Knowles, Ivana Kolcic, Inke R König, Seppo Koskinen, Peter Kovacs, Johanna Kuusisto, Peter Kraft, Kirsti Kvaløy, Jaana Laitinen, Olivier Lantieri, Chiara Lanzani, Lenore J Launer, Cecile Lecoeur, Terho Lehtimäki, Guillaume Lettre, Jianjun Liu, Marja-Liisa Lokki, Mattias Lorentzon, Robert N Luben, Barbara Ludwig, MAGIC, Paolo Manunta, Diana Marek, Michel Marre, Nicholas G Martin, Wendy L McArdle, Anne McCarthy, Barbara McKnight, Thomas Meitinger, Olle Melander, David Meyre, Kristian Midthjell, Grant W Montgomery, Mario A Morken, Andrew P Morris, Rosanda Mulic, Julius S Ngwa, Mari Nelis, Matt J Neville, Dale R Nyholt, Christopher J O'Donnell, Stephen O'Rahilly, Ken K Ong, Ben Oostra, Guillaume Paré, Alex N Parker, Markus Perola, Irene Pichler, Kirsi H Pietiläinen, Carl G P Platou, Ozren Polasek, Anneli Pouta, Suzanne Rafelt, Olli Raitakari, Nigel W Rayner, Martin Ridderstråle, Winfried Rief, Aimo Ruokonen, Neil R

Robertson, Peter Rzehak, Veikko Salomaa, Alan R Sanders, Manjinder S Sandhu, Serena Sanna, Jouko Saramies, Markku J Savolainen, Susann Scherag, Sabine Schipf, Stefan Schreiber, Heribert Schunkert, Kaisa Silander, Juha Sinisalo, David S Siscovick, Jan H Smit, Nicole Soranzo, Ulla Sovio, Jonathan Stephens, Ida Surakka, Amy J Swift, Mari-Liis Tammesoo, Jean-Claude Tardif, Maris Teder-Laving, Tanya M Teslovich, John R Thompson, Brian Thomson, Anke Tönjes, Tiinamaija Tuomi, Joyce B J van Meurs, Gert-Jan van Ommen, Vincent Vatin, Jorma Viikari, Sophie Visvikis-Siest, Veronique Vitart, Carla I G Vogel, Benjamin F Voight, Lindsay L Waite, Henri Wallaschofski, G Bragi Walters, Elisabeth Widen, Susanna Wiegand, Sarah H Wild, Gonneke Willemsen, Daniel R Witte, Jacqueline C Witteman, Jianfeng Xu, Qunyuan Zhang, Lina Zgaga, Andreas Ziegler, Paavo Zitting, John P Beilby, I Sadaf Farooqi, Johannes Hebebrand, Heikki V Huikuri, Alan L James, Mika Kähönen, Douglas F Levinson, Fabio Macciardi, Markku S Nieminen, Claes Ohlsson, Lyle J Palmer, Paul M Ridker, Michael Stumvoll, Jacques S Beckmann, Heiner Boeing, Eric Boerwinkle, Dorret I Boomsma, Mark J Caulfield, Stephen J Chanock, Francis S Collins, L Adrienne Cupples, George Davey Smith, Jeanette Erdmann, Philippe Froguel, Henrik Grönberg, Ulf Gyllensten, Per Hall, Torben Hansen, Tamara B Harris, Andrew T Hattersley, Richard B Hayes, Joachim Heinrich, Frank B Hu, Kristian Hveem, Thomas Illig, Marjo-Riitta Jarvelin, Jaakko Kaprio, Fredrik Karpe, Kay-Tee Khaw, Lambertus A Kiemeney, Heiko Krude, Markku Laakso, Debbie A Lawlor, Andres Metspalu, Patricia B Munroe, Willem H Ouwehand, Oluf Pedersen, Brenda W Penninx, Annette Peters, Peter P Pramstaller, Thomas Quertermous, Thomas Reinehr, Aila Rissanen, Igor Rudan, Nilesh J Samani, Peter E H Schwarz, Alan R Shuldiner, Timothy D Spector, Jaakko Tuomilehto, Manuela Uda, André Uitterlinden, Timo T Valle, Martin Wabitsch, Gérard Waeber, Nicholas J Wareham, Hugh Watkins, James F Wilson, Alan F Wright, M Carola Zillikens, Nilanjan Chatterjee, Steven A McCarroll, Shaun Purcell, Eric E Schadt, Peter M Visscher, Themistocles L Assimes, Ingrid B Borecki, Panos Deloukas, Caroline S Fox, Leif C Groop, Talin Haritunians, David J Hunter, Robert C Kaplan, Karen L Mohlke, Jeffrey R O'Connell, Leena Peltonen, David Schlessinger, David P Strachan, Cornelia M van Duijn, H-Erich Wichmann, Timothy M Frayling, Unnur Thorsteinsdottir, Gonçalo R Abecasis, Inês Barroso, Michael Boehnke, Kari Stefansson, Kari E North, Mark I McCarthy, Joel N Hirschhorn, Erik Ingelsson, and Ruth J F Loos. Association analyses of 249,796 individuals reveal eighteen new loci associated with body mass index.

Nature genetics, 42(11):937–948, nov 2010. ISSN 1061-4036. doi: 10.1038/ng.686. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3014648/.

John D Storey. A direct approachto false discovery rates. *J. R. Statist.Soc. B*, 1995.

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.

The CARDIoGRAMplusC4D Consortium. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nature Genetics*, 47(10):1121–1130, sep 2015. ISSN 1546-1718. doi: 10.1038/ng.3396. URL http://dx.doi.org/10.1038/ng.3396.

Wolfgang Viechtbauer. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48, 2010. URL http://www.jstatsoft.org/v36/i03/.

Anna A E Vinkhuyzen, Naomi R Wray, Jian Yang, Michael E Goddard, and Peter M Visscher. Estimation and Partitioning of Heritability in Human Populations using Whole Genome Analysis Methods. *Annual review of genetics*, 47:75–95, aug 2013. ISSN 0066-4197. doi: 10.1146/annurev-genet-111212-133258. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4037293/.

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.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL http://had.co.nz/ggplot2/book.

Yiwei Zhang, Weihua Guan, and Wei Pan. Adjustment for Population Stratification via Principal Components in Association Analysis of Rare Variants. *Genetic epidemiology*, 37(1):99–109, jan 2013. ISSN 0741-0395. doi: 10.1002/gepi.21691. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4066816/.