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 ÃÛ gene (GÃÛG) interactions, and gene ÃÛ environment (G $\times$ 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

List of Tables

List of Acronyms

CAD Coronary Artery Disease 1

Colophon

This document was typeset using the XeTeX typesetting system created by the Non-Roman Script Initiative and the memoir class created by Peter Wilson. The body text is set 10pt with Adobe Caslon Pro. Other fonts include Envy Code R, Optima Regular and. Most of the drawings are typeset using the TikZ/PGF packages by Till Tantau.

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.

1.1 Coronary Artery Disease

Coronary Artery Disease (CAD) is

1.2 Genome Wide Association Studies

1.3 Polygenic Prediction of Complex Disease

1.4 Polygenic Sliding Window Optimization

1.5 Summary