ELSEVIER

Archives of Medical Research ■ (2017) ■

Archives of Medical Research

ORIGINAL ARTICLE

Conditional Inference Tree for Multiple Gene-Environment Interactions on Myocardial Infarction Among Chinese Men

Zhijun Wu,[a,1] Xiuxiu Su,[a,1] Haihui Sheng,[b] Yanjia Chen,[a] Xiang Gao,[c] Le Bao,[d] and Wei Jin[a]

[a]*Department of Cardiology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China*
[b]*National Engineering Center for Biochip at Shanghai, Shanghai, China*
[c]*Department of Nutritional Sciences,* [d]*Department of Statistics, Pennsylvania State University, State college, Pennsylvania*

Received for publication July 4, 2017; accepted December 6, 2017 (ARCMED-D-17-00401).

Background and Aims. Identifying gene-environment interaction in the context of multiple environmental factors has been a challenging task. We aimed to use conditional inference tree (CTREE) to strata myocardial infarction (MI) risk synthesizing information from both genetic and environmental factors.

Methods. We conducted a case-control study including 1440 Chinese men (730 MI patients and 710 controls). We first calculated a weighted genetic risk score (GRS) by combining 25 single nucleotide polymorphisms (SNPs) that had been identified to be associated with coronary artery diseases in previous genome wide association studies. We then developed a CTREE model to interpret the gene-environment interaction network in predicting MI.

Results. We detected high-order interactions between dyslipidemia, GRS, smoking status, age and diabetes. Of all the variables examined, high density lipoprotein cholesterol (HDL-C) of 1.25 mmlo/L was identified as the key discriminator. The subsequent splits of MI were low density lipoprotein cholesterol (LDL-C) of 4.01 mmol/L and GRS of 20.9. We found that individuals with HDL-C $\leq$1.25 mmol/L, GRS $>$20.9 and lipoprotein (a) $>$ 0.09 g/L had a higher risk of MI than those who at the lowest risk group (OR: 5.89, 95% CI: 3.99−8.69). This magnitude of MI risk was similar to the combination of HDL-C $\leq$1.25 mmol/L, GRS $\leq$20.9, smoking and lipoprotein (a) $>$ 0.15 g/L (OR: 5.49, 95% CI: 3.51−8.58).

Conclusions. The multiple interactions between genetic and environmental factors can be visually present via the CTREE approach. The tree diagram also simplifies the decision making procedure by answering a sequence of questions along the branches.

Key Words: Data mining, Conditional inference tree, Myocardial infarction, Genome wide association study.

Introduction

Myocardial infarction (MI) is a terminal manifestation of multiple pathophysiology processes involving in genetic risk factors, intermediate conditions (e.g., hypertension, diabetes and dyslipidemia) and lifestyle factors (e.g., smoking and drinking habits) (1,2). Recently, genome-wide association studies (GWASs) have made great efforts on characterizing susceptible genes/single nucleotide polymorphisms (SNPs) to MI. However, the reproducibility of those findings is limited because the effect of single SNPs might depend on the gene-gene interactions or gene-environment interactions (3). There remains a great challenge in understanding the genetic architecture of common diseases with the existence of complex and non-linear associations between genes, environmental factors and diseases (4).

[1]These authors contributed equally to this work.

Address reprint requests to: Wei Jin, MD, PhD, Department of Cardiology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, No 197, Ruijin Er Road, Shanghai, 200025, PR China; Phone: +86-21-64370045; FAX: +86-21-64457177; E-mail: jinwei10724@126.com

0188-4409/$ - see front matter.
https://doi.org/10.1016/j.arcmed.2017.12.001

New modelling tools are needed to precisely evaluate complex interplay between disease-susceptible genes/SNPs and multiple environmental exposures (5).

The linear and logistic regression models seem helpful in identifying whether a variable or a combination of multiple variables is a significant risk factor by adjusting for other covariates. However, these methods are underpowered to depict the multitude and complexity of interaction network due to over-parameterization. For instance, the logistic regression model with 8 predictors would require comparison among $2^{8+28+56} = 4.95 \times 10^{27}$ candidate models that allow up to three way interactions.

There has been a growing interest in the use of recursive partitioning algorithms to assist in risk stratification and clinical decision making (6,7). These approaches, such as classification and regression tree (CART) and conditional inference tree (CTREE), are non-parametric and nonlinear techniques that separate the feature space into smaller and smaller sub-regions according to impurity criterion and finally generate a tree-structured visualization (7,8). CART has been applied in previous health care studies (6,9,10) and genetic studies (11,12). However, this approach has several weaknesses such as overfitting and a selection bias of covariates with a lot of possible splits. As an advanced splitting algorithm, CTREE is more effective to construct hierarchical models after an unbiased selection among arbitrary scaled variables that might be hard to be directly handled by traditional statistical approaches (13). However, CTREE has not been used in risk stratification synthesizing information from both genetic and environmental factors.

In the current study, we elucidated the gene-environment interactions using a CTREE model. Because a single common SNP generally has a small-effect on diseases and cannot represent the whole genetic burden, we created a weighted genetic risk score (GRS) by combining 25 SNPs which were identified in previous GWASs of coronary artery disease (CAD). The weighted GRS and other potential predictors for MI (e.g., age, lifestyle behavior, medical comorbidity and lipid profiles) were used to create a risk stratification algorithm for MI risk in a case-control study including 1440 Chinese men.

Methods

Participants

We recruited 1440 Chinese adult males who lived in Shanghai and visited the Ruijin Hospital from January 2008 to December 2012, including 730 MI patients and 710 controls. All individuals were of Southern Han Chinese ethnicity.

Both MI cases and controls underwent coronary angiography (CAG) using standard Judkins techniques (14). The standard CAG operating procedures recommended by the ACC/AHA guidelines were followed (15,16). Individuals who had normal coronary arteries, as diagnosed by CAG, were regarded as the control group. MI cases were diagnosed on the basis of positive cardiac enzymes, clinical symptoms and typical electrocardiographic changes. To be specific, MI patients were identified with ≥50% obstruction or atherothrombosis in at least one major epicardial coronary vessel (17). The physicians performing CAG were masked to the study design.

Finally, people with any of the following were excluded from the analysis: prior atherosclerotic cardiovascular diseases (e.g., CAD, MI or peripheral vascular disease) with or without revascularization; cardiomyopathy; congenital heart disease; pulmonary heart disease; valvular heart disease; infection; chronic kidney disease; tumor; or autoimmune disease.

This study was approved by the ethics committee of the Ruijin Hospital, Shanghai Jiao Tong University School of Medicine. The study protocol followed the norms of the World's Association Declaration of Helsinki. All participants provided their written informed consent at the time of enrollment.

DNA Extraction and Genotyping

Whole blood specimens were drawn after fasting overnight within 24 h of admission. Genomic DNA was extracted from peripheral blood leukocytes according to the standard phenol-chloroform methods, and was normalized to a concentration of 50 ng/μL. Laboratory personnel who performed genotyping were blinded to the clinical characteristics of DNA samples.

We genotyped 35 SNPs that have been identified to be relevant to atherosclerotic cardiovascular diseases by prior GWAS (18−29). The distribution of these genotypes was examined for the conformance of Hardy Weinberg Equilibrium (HWE) (30). We then tested the linkage disequilibrium within these SNPs, using the SHEsis platform (http://analysis.bio-x.cn/myAnalysis.php) (31). Inclusion criteria were a) minor allele frequency ≥1%; b) genotypes were consistent with HWE ($p > 0.05$); c) genotype call>95%; and d) SNPs were independent and in weak/no linkage disequilibrium (32). Finally, we included 25 SNPs for each individuals and constructed a weighted GRS by calculating the weighted number of risk alleles of the specific locus and dividing the sum by the total number of SNPs. Each allele was weighted by the effect size reported by previous GWASs (Supplementary Table 1) (33).

The flanking sequence of these SNPs was obtained from the NCBI database (http://www.ncbi.nlm.nih.gov/snp). The amplification and extension primers were designed using MassARRAY Assay Design 2.0 software. Genotyping was carried out using the matrix-assisted laser desorption/ionization time-of-flight mass spectrometry based on MassARRAY system (Sequenom, San Diego, CA, USA). More than 98% samples were successfully genotyped.