Exam 2009-06-11

You may answer in English or in Finnish.

1. Explain the following terms

   a) overdispersion (2 points)

   b) proportional odds model (2 points)

   c) Fisher scoring (2 points)

2. Read the pages 237–239 from the article *The impact of newly-identified loci on coronary heart disease, stroke and total mortality in the MORGAM prospective cohorts* (Appendix 1) and answer the questions.

   a) What link function was used when the disease status at baseline was analyzed? (1 point)

   b) How does this study differ from previous studies where the SNP rs1333049 was studied? (1 point)

   c) What is the point estimate of the relative hazard of incident CHD for the genotype CC (risk homozygote) of SNP rs1333049 compared to the genotype GG (non-risk homozygote)? (1 point)

   d) How missing covariates were handled? (1 point)

   e) The total number of the SNPs was 42. How this was taken into account when the results were interpreted? (1 point)

   f) How the subjects for the association analysis between SNPs and blood pressure were selected? (1 point)

3. Data on a continuous covariate x_i and a count response Y_i are collected for $i = 1, 2, \ldots, n$. The goal is to model Y_i by a generalized linear model and to estimate the regression coefficient β for the covariate x_i. The response Y_i is missing for $i \in S$ where $S \subset \{1, 2, \ldots, n\}$ but a binary variable Z_i is observed when $i \in S$. Variable Z_i has value 1 if $Y_i \geq 2$ and 0 otherwise. The missingness does not depend x_i or Y_i.

   a) Choose a suitable generalized linear model and write the log-likelihood based on observations where Y_i is available. (3 points)

   b) Write the log-likelihood based on all the observations $i = 1, 2, \ldots, n$. (3 points)

4. Derive the formula for the deviance residuals for a situation where a binomial response is modeled by a covariate x using a generalized linear model with the probit link.

5. Show that when the canonical link is used in a generalized linear model, the expected information and observed information are equal.

Genetic Epidemiology 33:237–246 (2009)

The Impact of Newly Identified Loci on Coronary Heart Disease, Stroke and Total Mortality in the MORGAM Prospective Cohorts

Juha Karvanen,[1]* Kaisa Silander,[2,3] Frank Kee,[4] Laurence Tiret,[5] Veikko Salomaa,[1] Kari Kuulasmaa[1] Per-Gunnar Wiklund,[6] Jarmo Virtamo,[1] Olli Saarela,[1] Claire Perret,[5] Markus Perola,[2,3,7] Leena Peltonen,[5] Francois Cambien,[5] Jeanette Erdmann,[10] Nilesh J. Samani,[11] Heribert Schunkert[11] and Alun Evans[4] for the MORGAM Project

[1]*Department of Health Promotion and Chronic Disease Prevention, National Public Health Institute, Helsinki, Finland*
[2]*Department of Molecular Medicine, National Public Health Institute, Helsinki, Finland*
[3]*FIMM, Institute for Molecular Medicine Finland, Helsinki, Finland*
[4]*The Queen's University of Belfast, Belfast, UK*
[5]*INSERM 525, Faculte de Medecine Pitie-Salpetriere, Paris, France*
[6]*Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden*
[7]*Department of Medical Genetics, University of Helsinki, Helsinki, Finland*
[8]*Department of Human Genetics, Wellcome Trust Sanger Institute, Cambridge, UK*
[9]*The Broad Institute of MIT and Harvard, Boston, Massachusetts, USA*
[10]*Universität zu Lübeck, Lübeck, Germany*
[11]*University of Leicester, Leicester, UK*

Recently, genome wide association studies (GWAS) have identified a number of single nucleotide polymorphisms (SNPs) as being associated with coronary heart disease (CHD). We estimated the effect of these SNPs on incident CHD, stroke and total mortality in the prospective cohorts of the MORGAM Project. We studied cohorts from Finland, Sweden, France and Northern Ireland (total $N = 33{,}282$, including 1,436 incident CHD events and 571 incident stroke events). The lead SNPs at seven loci identified thus far and additional SNPs (in total 42) were genotyped using a case-cohort design. We estimated the effect of the SNPs on disease history at baseline, disease events during follow-up and classic risk factors. Multiple testing was taken into account using false discovery rate (FDR) analysis. SNP rs1333049 on chromosome 9p21.3 was associated with both CHD and stroke (HR = 1.20, 95% CI 1.08–1.34 for incident CHD events and 1.15, 0.99–1.34 for incident stroke). SNP rs11670734 (19q12) was associated with total mortality and stroke. SNP rs2146807 (10q11.21) showed some association with the fatality of acute coronary event. SNP rs2943634 (2q36.3) was associated with high density lipoprotein (HDL) cholesterol and SNPs rs599839, rs4970834 (1p13.3) and rs17228212 (15q22.23) were associated with non-HDL cholesterol. SNPs rs2943634 (2q36.3) and rs12525353 (6q25.1) were associated with blood pressure. These findings underline the need for replication studies in prospective settings and confirm the candidacy of several SNPs that may play a role in the etiology of cardiovascular disease. *Genet. Epidemiol.* 33:237–246, 2009.

Key words: cardiovascular disease; genes; risk factors

Contract grant sponsor: European Community's Seventh Framework Programme; Contract grant number: FP7/2007-2013; Contract grant sponsor: ENGAGE project; Contract grant number: HEALTH-F4-2007-201413; Contract grant sponsor: Finnish Heart Association.
*Correspondence to: Juha Karvanen, Department of Health Promotion and Chronic Disease Prevention, National Public Health Institute, Mannerheimintie 166, 00300 Helsinki, Finland. E-mail: juha.karvanen@ktl.fi
Received 1 July 2008; Revised 19 August 2008; Accepted 23 August 2008
Published online 31 October 2008 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20374

1. INTRODUCTION

While the role of lifestyle risk factors is well established in the development of cardiovascular disease (CVD), the identification of genetic factors involved in the susceptibility to CVD has been more challenging, with only a few candidate genes reproducibly associated with disease [Arnett et al., 2007; Cambien and Tiret, 2007]. Recently, genome-wide association studies (GWAS) have opened a fresh avenue of research by affording the possibility to explore the whole genome without any a priori biological hypotheses. This approach had led to the identification of a new locus on chromosome 9p21 associated with coronary heart disease (CHD) and myocardial infarction in several GWAS [Helgadottir et al., 2007; McPherson et al., 2007; Samani et al., 2007; Wellcome Trust Case Control Consortium (WTCCC), 2007]. This association has now been replicated in several large cohorts and appears to be a robust finding, even though the underlying mechanism is not yet elucidated [Schunkert et al., 2008]. In addition, the association has been shown with two other arterial diseases [Helgadottir et al., 2008].

In the WTCCC Study [Wellcome Trust Case Control Consortium, 2007] and the German Myocardial Infarction

(MI) Study recently reported by the *Cardiogenics* Consortium [Samani et al., 2007], several loci, including chromosome 9p21.3, were identified as putative loci for CHD, namely 1p13.3, 1q41, 2q36.3, 6q25.1, 10q11.21 and 15q22.33. The natural sequence of this research is to replicate the GWAS findings in other populations and to address other interesting questions, such as the effect of the identified loci on the classic CVD risk factors or on other cardiovascular endpoints such as ischemic stroke. Moreover, these variants have been identified in case-control settings, but their effect on the risk of CHD, ischemic stroke and all-cause mortality has not been studied in a prospective setting. Of course it is widely acknowledged that the task of distinguishing meaningful signals from noise—"separating the gold from the fool's gold" [Dupuis and O'Donnell, 2007]—is not a solely problem with a genetic or technical solution, but must be addressed from epidemiological principles.

In consortium studies, a major challenge is to ensure that phenotypes have been assured with the same rigour as the genotypes, and furthermore, it can be dangerous to assume that case-control studies will reflect the impact of specific SNPs on incident disease in diverse populations with different absolute risks [McCarthy et al., 2008; Pearson and Manolio, 2008]. The vast majority of subjects in the WTCCC were British residents but, even so, there were several loci, some of which were associated with disease, which demonstrated substantial geographical variation in allele frequencies across Britain. In general, the differences in allele frequencies in Europe [Cavalli-Sforza and Piazza, 1993] and even in more isolated populations such as Finns [Pastinen et al., 2001] are known for long. Thus, replication of the findings in other European populations is essential. Although the WTCCC findings might assuage concerns about the use of common control groups for several disease phenotypes, when CHD studies from different European populations are pooled, it is important to ensure like is being compared with like. Lastly, given the fact that a substantial proportion of incident CHD cases die within 28 days and most of those who die do not even reach hospital [Tunstall-Pedoe et al., 1994], we should not assume that the findings in survivors will be identical to those in whom the outcome is fatal.

The case for replication is compelling and for this purpose we have used several extensively phenotyped prospective cohorts from Europe assembled within the MORGAM Project [Evans et al., 2005]. These cohorts have been followed up for between 5 and 10 years for CHD, stroke events and total mortality. Thus, we can assess whether the findings reported solely for CHD have relevance for the separate (though related) endpoints of stroke and total deaths, in cohorts free of disease at baseline. We can also explore the effect of these variants on classic CVD risk factors measured at baseline.

DATA AND METHODS

STUDY DESIGN

The study included two population-based cohorts from the FINRISK Study from Finland (follow-up 1992–2001 and 1997–2004), the ATBC study from Finland (follow-up 1992–1999), two cohorts from Northern Sweden (follow-up 1990–1999 and 1994–1999) and the PRIME cohort comprising three centers in France and one in Northern Ireland (5-years follow-up at the period 1991–1999). All the cohorts are part of the MORGAM Project [Evans et al., 2005; MORGAM Project, 2001-] and the cohort descriptions have been published [Kulathinal et al., 2005]. The study employed a case-cohort design in which all CVD cases from the prospective follow-up and a random subset of the cohort members were selected for genotyping [Kulathinal et al., 2007].

DEFINITION OF STUDY VARIABLES

The outcomes analyzed included disease status at baseline and disease events during the follow-up. History of MI at baseline (yes or no) and history of stroke at baseline (yes or no) were primarily based on health information sources such as a hospital discharge registers and on self-reports (questionnaire responses about doctor diagnoses). The cohorts were followed up for all fatal and non-fatal acute CHD and stroke events as well as all other deaths. The time and the type of the event were recorded. An event was considered as fatal if the subject died within 28 days of the onset of the CHD or stroke event. The main CHD outcome was defined as first fatal or non-fatal CHD event which included definite and possible acute MI or coronary death, unstable angina pectoris, revascularization and unclassifiable fatal events. The main stroke outcome was defined as first fatal or non-fatal likely cerebral infarction, which includes events validated as cerebral infarction and events that were not validated but most likely were cerebral infarctions on the basis of the clinical or death diagnoses. The details of the follow-up procedures are described in the cohort descriptions [Kulathinal et al., 2005].

Risk factors measured at baseline included total and high density lipoprotein (HDL) cholesterol, systolic and diastolic blood pressure (two measurements), daily smoking, height and weight. In the analysis, we used non-HDL cholesterol (difference of total cholesterol and HDL cholesterol), HDL-cholesterol, mean blood pressure (the mean of the two diastolic and the two systolic blood pressure measurements), current daily smoking (yes or no) and body mass index (weight in kilograms divided by square of height in meters). Other variables collected at baseline included self-reported history of diabetes, drug treatment for high cholesterol and drug treatment for high blood pressure. The baseline measurement procedures were highly standardised [Niemelä et al., 2007].

GENOTYPING

The SNP markers selected for genotyping included the seven lead SNPs (1p13.3, 1q41, 2q36.3, 6q25.1, 9p21.3, 10q11.21 and 15q22.33) identified in the WTCCC Study and German MI Study and six lead SNPs (1q32.2, 1q43, 5q21, 16q23, 19q12 and 22q12) identified in the WTCCC study but not replicated in the German MI Study. The genotyping plan also included proxies that were in almost complete linkage disequilibrium with the lead SNP and other SNPs that were selected because they brought some additional information about the haplotypic structure of the locus. The full list of the 42 SNPs is given in the web supplement http://www.ktl.fi/publications/morgam/cardiogenics/index.html.

Forty single nucleotide polymorphism (SNP) markers out of 42 were genotyped with the iPLEX chemistry on the

MassARRAY system (Sequenom, San Diego, CA), using a protocol specified by the manufacturers, and 12.5–20 ng of genomic DNA. For 233 samples with less than 7.5 μg genomic DNA, the DNA was whole genome amplified prior to genotyping [Silander et al., 2005]. Genotyping was done in 384-well plates which contained eight non-template controls and eight plate-specific duplicates in plate-specific positions and 5% blind duplicates. The case status was unknown to the laboratory staff. Assay information has been provided elsewhere [Schunkert et al., 2008].

Because of difficulties in genotyping SNPs rs599839 and rs2943634 with the iPLEX technique, they were genotyped using the 5′ nuclease assay with MGB TaqMan probes (Applied Biosystems, Courtaboeuf, France). Fluorescence was measured with an ABI PRISM 7000 sequence detection system (Applied Biosystems). Primer and probe sequences can be found at the GeneCanvas website (http://www.gencanvas.org).

STATISTICAL METHODS

We used logistic regression to analyze the case fatality and the disease status at baseline, a Cox proportional hazards model to analyze the disease events during the follow-up and linear regression to analyze the association between genotypes and risk factors. Because subjects were selected for genotyping according to the case-cohort design, cases and subcohort members had to be weighted appropriately in the analyses. In logistic regression models, subjects had weights proportional to the inverse of the selection probabilities [Zhao and Lipsitz, 1992]. In time-to-event models, subcohort members had weights proportional to the inverse of the selection probabilities and cases had a weight of one at the time of the event. In linear regression models, only the subcohort members without history of CVD were included and no weighting was used. The analysis methods for the MORGAM case-cohort design are described in detail elsewhere [Kulathinal et al., 2007]. Statistical analysis was carried out using R [R Development Core Team, 2008].

A separate model was fitted for each SNP implying that there were 42 models for each outcome. Each analysis included only subjects who had complete data on the risk factors in the model. In all models, the heterozygote was coded value 0 and the homozygotes were coded as 1 and −1. Models for disease status at baseline were adjusted for age at baseline, sex and cohort. Models for events during the follow-up were adjusted for cohort, HDL cholesterol, non-HDL cholesterol, mean of systolic and diastolic blood pressure, body mass index, current daily smoking and history of diabetes. The age of the subject was used as the time variable and separate baseline hazards were assumed for men and women. For comparison we also fitted models that were adjusted only for cohort and current daily smoking and stratified by sex. Models for the risk factors were fitted for the subcohort and were adjusted for age at baseline, sex and cohort. Subcohort members who had drug treatment for high cholesterol were excluded from the analyses of the cholesterol measurements and subcohort members who were on drug treatment for high blood pressure were excluded from the blood pressure analyses.

The study had a power of 86% to detect an effect on CHD with a hazard ratio of 1.2 and a power of 41% for detecting an effect with a hazard ratio of 1.1 in a single test with a nominal significance level of 5% [Cai and Zeng, 2004]. For stroke (cerebral infarction) the corresponding powers were 61 and 26%. The details of the power calculations are given in the Supplement.

False-discovery rate (FDR) analysis [Benjamini and Hochberg, 1995] with a conservative a priori assumption that there were no true positive findings in the results was used to address multiple testing. The FDR analysis was performed separately for each outcome. An SNP was identified as interesting if it was one of the seven lead SNPs found in the WTCCC Study and German MI Family Study (rs1333049, rs6922269, rs2943634, rs599839, rs17465637, rs501120 and rs17228212) or if it was identified in one of our main analyses (disease status at baseline, events during follow-up and baseline risk factors) using an FDR threshold of 20%.

RESULTS

The total number of individuals genotyped under the case-cohort design was 5,613 of which 2,341 were subcohort members. The characteristics of the case-cohort set and the subcohort are summarized in Table I. The tables in the web supplement http://www.ktl.fi/publications/morgam/cardiogenics/index.html report the genotyping success rates, allele frequencies and Hardy-Weinberg test statistics for each cohort. No major departures from the Hardy-Weinberg equilibrium were found and 99.89% of blind duplicate genotypes were consistent. A total of 247 samples from the PRIME cohort were genotyped using both iPLEX and TaqMan chemistries for two of the SNPs, rs1333049 and rs6922269. Among 457 successful genotype pair comparisons, five discrepancies in four samples were present, resulting in a 98.9% genotype concordance.

In addition to the seven lead SNPs, six SNPs were identified as interesting by the FDR analysis: rs4970834 (1p13.3), rs2972147 (2q36.3), rs12525353 (6q25.1), rs10738610 (9p21.3), rs2146807 (10q11.21) and rs11670734 (19q12). The genotype distributions of the interesting SNPs in different study populations are reported in Table II. The results are not reported for rs2972147 (2q36.3) and rs10738610 (9p21.3) because they are almost complete proxies of the lead SNPs. The largest differences between the populations were found in chromosome 9p21.3 (rs1333049) where the risk allele C of rs1333049 had a frequency of 52.1% (95% confidence interval 44.2–60.0) in PRIME/Belfast and a frequency of 39.5% (36.2–42.8) in FINRISK and in chromosome 19q12 (rs11670734) where allele C had a frequency of 40.2% (32.5–47.9) in PRIME/Belfast and a frequency of 25.5% (18.9–32.1) in Northern Sweden.

The results concerning disease status at baseline (Table III) showed that chromosome 9p21.3 (rs1333049) was associated with both a history of MI and a history of stroke. The odds ratio *per allele* for rs1333049 was 1.24 (1.11–1.39) for MI and 1.22 (1.06–1.41) for stroke. Also chromosome 19q12 (rs11670734) was associated with history of MI and history of stroke. Chromosome 1p13.3 (rs599839) and 1q41 (rs17465637) were associated with history of MI. Chromosome 10q11.21 was also associated with history of MI but this association was not seen on the lead SNP (rs501120) but in a proxy (rs2146807). No

Exponential family

$$f_{Y_i}(y_i;\theta_i,\phi) = \exp\left(\frac{a_i(y_i\theta_i - b(\theta_i))}{\phi} + c(y_i,\phi/a_i)\right).$$

Normal distribution

$$f(y;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right).$$

Poisson distribution

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

Binomial distribution

$$f(k;m,\pi) = \binom{m}{k}\pi^k(1-\pi)^{m-k}.$$

Gamma distribution

$$f(y;\lambda,\nu) = \frac{1}{\lambda^\nu\Gamma(\nu)}y^{\nu-1}e^{-y/\lambda}, \quad y>0,$$

where $\nu > 0$ is the shape parameter, $\lambda > 0$ is the scale parameter and Γ is the gamma function.

Raw residuals (response residuals)

$$r_i = y_i - \hat{\mu}_i.$$

Pearson residuals

$$r_{P,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)/a_i}}.$$

Deviance residuals

$$r_{D,i} = \operatorname{sign}(y_i - \hat{\mu}_i)\sqrt{d_i},$$

where

$$d_i = 2a_i\left(y_i\left(\theta_i(y_i) - \theta_i(\hat{\mu}_i)\right) - b\left(\theta_i(y_i)\right) + b\left(\theta_i(\hat{\mu}_i)\right)\right).$$