Author's response to reviews

Title: Genotype-Informed Calculation of Risk of Coronary Heart Disease Based on Genome-Wide Association Data Linked to the Electronic Medical Record

Authors:

Keyue Ding (ding.keyue@mayo.edu)
Kent R Bailey (baileyk@mayo.edu)
Iftikhar J Kullo (kullo.iftikhar@mayo.edu)

Version: 2 Date: 5 August 2011

Author's response to reviews: see over
Reviewer #1

I applaud the authors for tackling a critical problem how genotype alters traditional risk stratification using a creative EMR-based approach.

My major concern with the manuscript is that the major findings are essentially calculations with no validation from a patient cohort, EMR-based or conventional.

We have acknowledged this limitation in our manuscript. Our main intent was to demonstrate that a genetic risk score can be created from high-density genotyping data linked to the EMR.

The authors state (in the second-to-last sentence of the manuscript) that the structure of their EMR study does not allow them to ascertain CHD phenotypes, which is what a reader might be expecting from an EMR approach.

Coronary heart disease (CHD) was ascertained by the International Classification of Disease-9-Clinical Modification (ICD-9-CM) diagnosis codes for ischemic heart disease or a history of percutaneous coronary intervention or coronary artery bypass surgery (See below).

Major Revisions:

1. p. 10: The authors state that the genetic risk scores did not correlate with traditional risk factors or the Framingham risk score. However, 3 of the 11 SNP loci are related to lipid phenotypes (PCSK9, LDLR, SORT1). Does the genotype at these loci (either one locus at a time, or in aggregate) correlate with the lipid values (esp LDL) extracted from the EMR? This in and of itself would be an interesting contribution to the EMR-based genetics literature, and would potentially help validate their overall approach in this manuscript and their phenotype ascertainment.

We thank the reviewer for the comments. We tested the association between three SNPs known to be related to lipid traits – rs11206510 in PCSK9, rs599839 in SORT1, and rs1122608 in LDLR – with total cholesterol, LDL cholesterol, and HDL cholesterol. We found that rs11206510 was associated with total cholesterol ($P = 0.026$) and LDL cholesterol ($P = 0.027$), and rs599839 was marginally associated with HDL cholesterol ($P = 0.050$) and LDL cholesterol ($P = 0.078$). However, the overall weighted GRS was not associated with Framingham risk score ($P=0.78$). We have added these statements in the revised manuscript (Page 6, paragraph 3).

2. The lack of EMR validation of the revised risk scores should be made clearer.

-For instance, the title should be modified to something along the lines of: Genotype-informed calculations of risk of ...’ or ‘Genotype-informed calculated risk of ...’
Based on the reviewer’s comment, we have changed the title to ‘Genotype-informed calculation of risk of coronary heart disease based on genome-wide association data linked to the electronic medical record.’

-The need for validation in traditional or EMR cohorts should be added to Table 5 (Limitations)

We have mentioned this limitation in the revised Table 5.

-The absence of empiric validation should be stated more prominently in the Discussion, e.g. at the end of the first paragraph.

In the first paragraph of the Discussion section, we have added the following statement of ‘Such an approach to risk reclassification needs to be validated in prospectively followed cohorts.’ (Page 8, paragraph 3)

Minor Revisions:

1. In the Methods section, it would be helpful to briefly summarize what methods were used in the prior CVD publication to determine that the current study subjects did not have CHD, given the centrality of this to the current study.

CHD was defined as the presence of ICD-9-CM diagnosis codes for ischemic heart disease including 410.33-414.33, or a history of percutaneous coronary intervention or coronary artery bypass surgery (ICD-9-CM procedure codes 36.10-36.14). We have incorporated this statement in the revised manuscript (Page 3, paragraph 2).

2. Some discussion of why the authors chose to use an additive model would be informative (especially since they note in the Discussion that others use multiplicative models).

An additive model was chosen because it performs well even when the true genetic model may not be known or may be incorrectly specified [1]. We have added this statement in the revised manuscript (Page 4, paragraph 2)
Reviewer #3

Major Compulsory Revisions

The clinical indices used in the calculation of Framingham risk scores are standard measures and it is unclear why the authors have presented this as an advantage of using EMRs. It is likely that the unique advantage of EMRs in genetic studies may be access to longitudinal data rather than cross-sectional data presented here.

We thank the review for the comment. Our intent was to demonstrate that a genetic risk score could be calculated from high-density genotyping data linked to the EMR with significant implications for disease risk reclassification. As a first step in this direction, we needed to calculate the Framingham risk score from cross-sectional data in the EMR.

There is no mention of family history and perhaps this is not captured in EMRs. In any test of genetic prediction, family history is essential.

We agree with the reviewer. Capture of family history from the EMR is challenging. The Centers for Disease Control and Prevention has created the software ‘Family Healthware’, a family history-screening tool for common chronic diseases that can be incorporated into the EMR. We have mentioned this limitation in the revised manuscript (Page 9, paragraph 1).

The lack of prospective outcomes does not allow any interpretation of the risk scores and the validity of the reclassification. There need to be evidence presented that the reclassification are indeed valid and this requires prospective data.

We acknowledged this limitation in our study. Prospective studies are needed to empirically validate the clinical utility (i.e., risk reclassification) of genotype information (Page 8, paragraph 3). Our main intent was to demonstrate that a genetic risk score can be created from high-density genotyping data linked to the EMR.
Reviewer #2

1. **The biggest limitation is the lack of outcomes which the authors acknowledge.**
   We have highlighted this limitation in the revised manuscript.

2. **The authors use the Framingham risk score as published by Wilson et al. However, the clinically used ATP-III based risk prediction score is a modified version of this. Diabetes, for example, will be considered a `high risk equivalent' but is used as a variable in the risk prediction tool per their methods. Why did the authors choose to use the FRS (original) rather than that which is clinically used (ATP III)? Were the results similar if the ATP III version was used.**
   We chose the Wilson paper [2] as the authors provided Beta-coefficient for each risk factor and survival function allowing us to calculate the hazard ratio for CHD (ie, based on conventional risk factors) and 10-y CHD risk. These parameters were not available in the APT-III based risk score although the risk score itself can be calculated from the Tables B1 and B2 in [3].

3. **Better characterization of the study population is needed: The authors state that these were 1262 controls. How were they chosen? Were they consecutive? Was absence of cardiovascular disease based on history alone or history + ECG or.???
   CHD was defined as the presence of ICD-9-CM diagnosis codes for ischemic heart disease including 410.33-414.33, or a history of percutaneous coronary intervention or coronary artery bypass surgery (ICD-9-CM procedure codes 36.10-36.14). We used the following criteria to define the presence of peripheral arterial disease (PAD): an ankle-brachial index (ABI) of 0.9 or less at rest or one minute after exercise; or the presence of poorly compressible arteries; or normal ABI but a history of revascularization for PAD. Patients without CHD and PAD were selected as controls in the present study. A significant proportion of controls underwent ECG testing and those with a study positive for ischemia were excluded. We have incorporated these statements in the revised manuscript (Page 3, paragraph 2).

4. **Most of the SNPs presented were likely associated with CHD in Whites alone. No information is provided regarding the racial make up, i.e. were all subjects included Whites? If different ethnicities were enrolled how were ORs/ RRs estimated for them?**
   We excluded non-whites based on the quality control steps developed by the eMERGE genomics work group [4]. The identity-by-state (IBS) matrix was constructed for all samples genotyped in the eMERGE network (n=17,358). The first two dimensions were derived from multidimensional
decomposition analysis (using `cmdscale' function in R) of the 1-IBS matrix. Samples > 6 standard error (s.e.) from the mean of self-reported white ethnicity on dimensions 1 and 2 were excluded (Page 4, paragraph 1). After using these QC steps, we kept 1243 individuals of European ancestry without CHD in the revised manuscript. We have made corresponding changes in the revised manuscript.

5. The authors need to explain why different strategies were used in estimating the Odds Ratio and 95% CI (i.e. one strategy for 4 and another for the remaining 11). Uniformity in general would be preferred.

Four out of 12 SNPs (rs599838, rs501120, rs11206510, and rs1746048) were replicated in different studies. Therefore, we used a fixed effect model to calculate the summary odds ratio and its 95% confidence interval based on different studies. The remaining SNPs were discovered and replicated in a single study, therefore we used the odds ratios from the combined analysis (discovery + replication cohorts). The references for each SNP are summarized in an article that we have provided a reference to [5].

6. The quality control for the genotypes in eMERGE is referenced; however a couple of key findings could be included as this would be important to the paper as not all readers will be able to access the reference.

The quality control procedures for the genotype data were developed by the eMERGE genomic work group [4]. The following QC criteria were used: SNP call rate > 98%, samples call rate > 98%, minor allele frequency > 0.05, Hardy-Weinberg equilibrium \( P > 0.001 \), and 99.99% concordance rate in duplications. This information was incorporated into the revised manuscript (Page 3, paragraph 3; Page 4, paragraph 1).

7. The construction of the odds ratio/ genotype effect from multiple SNPs is a little difficult to follow. Would suggest formal statistical review of the same as this is a very important part of the paper. The overall risk score ranged from 6-to 18 suggesting that all individuals had at least 6/22 of the risk alleles. The authors also suggest that having 12 risk alleles will be associated with a lesser relative risk than the population which is a bit confusing. What this suggests is that most (all) of the population have some risk alleles which in turn then throws up the question related to the value of these risk alleles when, even in their presence, your risk could be less than that of the population.

Given a population of average relative risk of 1, the average genotype effect (ie, combined odds
ratios and relative risks) in patients with number of risk alleles <= 12 is less than 1 (Fig. 2b). Both genetic risk scores were marginally associated with CHD based on logistic regression ($P=0.08$ for both genetic risk scores) (Page 7, paragraph 1).

8. The authors state in the results and again in discussion that they only chose SNPs associated with CHD but not with CHD risk factors? Is this correct? Looking at the SNPs on Table 2, surely LDLR and PCSK9 are associated with LDL-cholesterol. So am not sure if what the authors state is correct. If the argument is that LDL-c is not part of FRS, by considering both Total Cholesterol and HDL-c in the FRS, the LDL-c is accounted for. Further, LDL-c is clearly a CHD risk factor.

We regret the misunderstanding. In the present study, we selected SNPs associated with CHD but not risk factors alone. Some selected SNPs are associated with both CHD and risk factors. In the calculation of Framingham risk score, we used total cholesterol instead of LDL cholesterol.
Reviewer #4

Minor Essential Revisions

This is a statistical review. I list specific questions, which need to be answered in the manuscript.

1. In the first paragraph on Genetic marker selection and imputation, the last two sentences describe methods used to calculate odds ratios and 95% confidence intervals for the 12 SNPs. How did the methods differ for the two groups of SNPs? Is there a difference between `calculated' and `derived'? between published studies and original studies? Between summary or combined or pooled odds ratios? Were fixed effects used for all of them? Why not use the same method for all 12?

Four out of 12 SNPs (rs599838, rs501120, rs11206510, and rs1746048) were replicated in different studies. Therefore, we used a fixed effect model to calculate the summary odds ratio and its 95% confidence interval based on different studies. The remaining SNPs were discovered and replicated in a single study. We used the odds ratios from the combined analysis (discovery + replication cohorts). The references for each SNP were summarized in our review article [5].

2. I understand that the SNPs on the Human660W chip, together with the HapMap II CEU were used for imputation by MACH. Why could the genotypes for rs3008621 not be inferred?

The rs3008621 was not genotyped in HapMap II CEU (release 21).

3. In equation (1), is each \( w_i \) the log(odds ratio for gene \( i \))? If yes, why call it `derived from the risk coefficient for gene \( i \) based on odds ratios,...'?

Yes. The \( w_i \) is the log(odds ratio for gene \( i \)). We have replaced the phrase as ‘multiplying the logarithm of odds ratio (\( w_i \)) by 0, 1, or 2 according to the number of risk alleles carried by each person’ (Page 4, paragraph 3).

4. Provide a clearer reference to the derivation of the method used for equation (4). I could not find it on www.decodeme.com.

The reference for Eq. (4) is provided in the revised manuscript (http://www.decodeme.com/health-watch-information/risk-calculation, reference 29 in the revised manuscript).
References


