The handle http://hdl.handle.net/1887/25760 holds various files of this Leiden University dissertation.

**Author:** Vaarhorst, Anika  
**Title:** Genetic and metabolomic approaches for coronary heart disease risk prediction  
**Issue Date:** 2014-05-13
Summary and general discussion
The prediction of incident coronary heart disease (CHD) in individuals free from cardiovascular disease (CVD) is currently based on traditional risk factors (TRFs) such as age, gender, self-reported diabetes, total cholesterol, HDL cholesterol, systolic blood pressure, current smoking, body mass index and a parental history of myocardial infarction (MI).\textsuperscript{1,2} In this thesis we investigated whether CHD risk prediction can potentially be improved by applying a metabolomics approach and by including information on common genetic variation previously reported to be associated with CVD or risk factors associated with CVD. We tested if metabolic profiles generated by \textsuperscript{1}H-NMR spectroscopy could mark individuals at high risk for CHD using the Cardiovascular Registry Maastricht (CAREMA) cohort and the Erasmus Rucphen Family (ERF) study. In addition we compared two different approaches for analysing \textsuperscript{1}H-NMR spectroscopy data. Next, using the CAREMA cohort, it was investigated if single nucleotide polymorphisms (SNPs) associated with CHD and/or CHD risk factors identified by genome-wide-association studies (GWAS) could improve CHD risk prediction. In the Leiden Longevity Study (LLS), we examined comprehensive lipid profiles for their ability to mark metabolic health in middle-aged individuals. Finally we focused on combining genetic information of the FADS1 locus with a comprehensive fatty acid profile for understanding and predicting CHD using the CAREMA cohort.

\textbf{Main findings}

In \textit{chapter 2} we explored the potential use of metabolomics technology for estimating CHD risk. In a case-cohort study design within the CAREMA cohort, it was investigated whether a single-point blood measurement of the metabolome, using \textsuperscript{1}H-NMR spectroscopy, can be instrumental to predict CHD risk in individuals free from CVD. By applying a targeted approach we identified and quantified 100 signals in the obtained \textsuperscript{1}H-NMR spectra representing 36 different low-molecular-weight metabolites. To select the \textsuperscript{1}H-NMR signals associated with incident CHD, Least Absolute Shrinkage and Selection Operator (LASSO) regression was applied.\textsuperscript{3,4} This method performs variable selection and shrinkage at the same time, which is useful when aiming at finding prediction rules in high-dimensional data.\textsuperscript{3,4} Using this method we found that a combination of 16 signals out of the 100 was the best subset for CHD risk prediction. These signals represent valine, various lipid fractions, glycoproteins, glutamate, citrate, ornithine, trimethylamine N-oxide (TMAO), 1,5-anhydroseritol, glucose, serine and creatine. A metabolite score based on these 16 signals was associated with incident CHD with a hazard ratio (HR) of 1.91 per standard deviation (SD). Indicating that someone with 1 SD increase in the metabolite score had a 1.91 increased chance of developing CHD. After adjusting for the traditional CHD risk factors the HR decreased from 1.91 to 1.50 but remained significant, indicating that the metabolite score also captured information that
Summary and general discussion

Chapter 7

is not reflected by TRFs. Next, it was tested if this metabolite score improved CHD risk prediction beyond TRFs using Harrell's C-index. This index assesses the ability of a marker to discriminate between events and non-events, by ranking participants who develop an event during follow-up higher than the ones who stay free from an event.\textsuperscript{5} For scores based on traditional CHD risk factors, like the Framingham risk score, the C-index ranges between 0.71 to 0.84.\textsuperscript{6} The metabolite score, when combined with age and sex, had a C-index of 0.81, which is comparable to a score based on TRFs (C-index = 0.82). Using the ERF study, an independent study population, it was found that the same \textsuperscript{1}H-NMR signals were also relevant for distinguishing between prevalent CHD cases and CHD-free controls.

In chapter 3 we compared our own targeted approach for data reduction (see chapter 2) to a more commonly applied method which involves partitioning the \textsuperscript{1}H-NMR spectra into discrete sections, and uses the sum of the spectral intensity in each section.\textsuperscript{7} This process is also known as equidistant binning.\textsuperscript{7} The reason for this comparison is that different methods for data reduction may have a different impact on subsequent data analysis and biomarker discovery.\textsuperscript{8} The \textsuperscript{1}H-NMR spectra were partitioned in sections of 0.02 parts per million (ppm), resulting in 401 bins. As mentioned in chapter 2, the targeted approach resulted in 100 \textsuperscript{1}H-NMR signals representing 36 metabolites. In addition we also evaluated the impact of two different scaling methods (autoscaling versus pareto scaling) on data analysis. LASSO regression in combination with autoscaling was performed to select the most important bins (when using the untargeted approach) or signals (when using the targeted approach) for CHD risk prediction. The selected bins or signals and their corresponding coefficients were used to construct a metabolite score. When using autoscaling, the metabolite scores based on both approaches were associated with incident CHD, independent of TRFs. The metabolite score based on the untargeted showed better discrimination as assessed by the C-index than the metabolite score based on the targeted approach. However the untargeted approach might be more prone to overfitting than the targeted approach. To test this presumption, the metabolite scores need to be validated in independent prospective cohorts. To avoid the selection on bins located in the noise region of the \textsuperscript{1}H-NMR spectrum, pareto scaling can be used as a data pre-treatment step. When using pareto scaling as data pretreatment, both the targeted and the untargeted metabolite score had comparable discriminatory capabilities. However, these metabolites scores had a lower discriminatory capability when compared to the metabolite score based on the targeted approach using autoscaling as a data pre-treatment step.

In chapter 4, using the LLS, we investigated whether a comprehensive lipid profile obtained by \textsuperscript{1}H-NMR spectroscopy,\textsuperscript{9} can classify middle-aged individuals according to their metabolic health. Information on classical lipids was also available. The LLS consists of long-lived
siblings, their offspring and the partners of the offspring (see figure 1 for a visualization of the study design). As a group, the middle-aged offspring of long-lived individuals show a survival advantage, and a lower prevalence of CHD, hypertension and type 2 diabetes than their spouses, who act as controls. The offspring therefore represent individuals with a better metabolic health as compared with controls of comparable age, sex and BMI representing the general population. We investigated whether lipoprotein particle sizes obtained by \textsuperscript{1}H-NMR spectroscopy represented improved markers of familial longevity compared to traditional lipids. We found that indeed for males, metabolic health was best characterized by larger mean low-density lipoprotein (LDL) particle sizes, whereas in females metabolic health was characterised by lower triglyceride levels.

In chapter 5, using the CAREMA case-cohort study, it was investigated if genetic risk scores (GRSs) based on SNPs associated with CHD or CHD risk factors according to previously published GWAS could improve CHD risk prediction beyond TRFs. Several approaches to construct a GRS were explored. The first approach was based on counting the number of risk alleles in every individual. These counted GRSs were not associated with CHD, nor did these GRS improve risk reclassification or discrimination. Second, to account for different effect sizes, every risk allele was multiplied by its corresponding weight. These weights were based on two published meta-GWAS. For practical reasons only the 29 SNPs associated with CHD were used to construct a weighted GRS. This weighted CHD GRS was associated with CHD before and after adjusting for TRFs and improved risk reclassification with 2.8%, but not risk discrimination. From these results it can be concluded that it is important to account for the different effect sizes of the SNPs. A GRS composed of 14 SNPs selected by LASSO regression improved risk discrimination and reclassification. This LASSO regression based GRS needs to be validated in independent populations.
In chapter 6 we investigated if the product C20:4n-6 to precursor C20:3n-6 and the product C18:2n-6 to precursor C18:3n-6 ratios, as markers of δ-5 and δ-6 desaturase activity, influence CHD risk (see for more information on the synthesis of these polyunsaturated fatty acids figure 1 in chapter 6 and for more information on the nomenclature of fatty acids see box 1). We found that an increased C20:4n-6 to C20:3n-6 ratio marking an increased δ-5 desaturase activity,13,14 was associated with a reduced CHD risk. This suggests that a high δ-5 desaturase activity is protective against the development of CHD, whereas no effect was found for δ-6 desaturase. δ-5 desaturase is encoded by FADS1. The minor G allele of a common genetic variant located in the FADS1 gene (rs174547) has been associated with a lower C20:4n-6 to C20:3n-6 and C18:3n-6 to C18:2n-6 ratio,15 increased plasma lipid levels,16-19 and resting heart rate,20 but never with CHD. In the study described in this thesis it was also observed that carriers of the minor G allele of that SNP had lower C20:4n-6 to C20:3n-6 and C18:3n-6 to C18:2n-6 ratios, but not an increased CHD risk. This seems counterintuitive, however we did find that the protective effect of a high δ-5 desaturase activity was mainly observed in AA carriers of the FADS1 genotype, which could be indicative for a gene-environment interaction.

**Box 1 | The nomenclature of the fatty acids**

A fatty acid consists of a carboxylic acid (the -COOH group) with a long aliphatic tail and can be saturated or unsaturated.21 Saturated fatty acids have no double-bounds.21 Mono unsaturated fatty acids (MUFAs) have one double bound and polyunsaturated fatty acids (PUFAs) are characterized by multiple double bounds.21 In this thesis the fatty acids are named by their common or trivial names, but also the lipid numbers and the ω-x nomenclature are used to identify the fatty acids in a more systematic way. The lipid numbers take the form C:D, were C is the number of carbon atoms and D in the number of double bounds.21 For example as shown in the figure the fatty acid arachidonic acid consists of 20 carbon atoms and has 4 double bounds, thus its lipid number is C20:4. The ω-x nomenclature indicates were the first double bound is located, counted from the CH₃ end of the fatty acid.21 The first double bound of Arachidonic acid starts at the 6th carbon atom (ω-6).

![Figure](image_url) The chemical structure of arachidonic acid (C20:4ω -6). This fatty acids consists of a carboxylic acid with a long aliphatic tail. There are 20 carbon atoms present and 4 double bounds (C20:4). The first double bound is located at the sixth carbon from the CH₃ end of the fatty acid (ω-6).

**Metabolomics and CHD risk prediction**

Based on chapter 2 of this thesis, it can be concluded that a subset of 16 signals from a total of 100 ¹H-NMR signals, when combined into a weighted score was predictive for incident CHD. When this metabolite score was combined with age and sex, its predictability was comparable to a score based on traditional CHD risk factors. This metabolite score however
Chapter 7

could not improve risk prediction beyond traditional CHD risk factors. However the 100
$^1$H-NMR signals represent only 36 metabolites and the human serum metabolome consists
of at least 4229 confirmed metabolites.22
Several other studies have investigated if a metabolomics approach using either $^1$H-NMR
based or mass spectrometry (MS) based platforms could be useful to predict incident CHD.23-
26 Mora et al. found that information on lipoproteins obtained by $^1$H-NMR spectroscopy was
comparable, but not superior to standard lipids when predicting incident CAD in women,23
confirming the results described in this thesis. According to Würtz et al., when standard
lipids were replaced by $^1$H-NMR determined LDL-cholesterol and medium HDL-cholesterol,
docosahexaenoic acid, and tyrosine, the prediction of incident high intima media thickness
improved in comparison to traditional lipids.24 Shah et al. showed that a MS based
metabolite profile consisting of short-chain dicarboxylacylcarnitines was associated with
subsequent incident cardiovascular events in CAD patients after adjusting for traditional
CHD risk factors.25 Wang et al. showed that the metabolites choline, TMAO and betaine
were associated with CVD risk independent of traditional cardiovascular risk factors and
medication use.26 Thus it seems that there is potential for metabolomics approaches to
improve CHD risk prediction. Thus for future research, metabolites from various platforms
should be tested in independent prospective cohort studies by meta-analysis to evaluate
their usefulness for CHD risk prediction in comparison to the traditional CHD risk factors.

When using the untargeted approach with autoscaling as a data pre-treatment step, LASSO
regression resulted in the selection of bins located in noise region of the $^1$H-NMR spectra.
Since other studies indicate that more reliable results are obtained with a targeted approach
than binning approaches,27,28 this could indicate that the untargeted approach could result
in overfitting. With the targeted approach, this part of $^1$H-NMR spectrum is eliminated. The
metabolite score based on this approach was associated with incident CHD independent
of TRFs. Moreover, when combined with age and sex this score had a comparable C-index
to a score based on TRFs. Thus it seems that with a relative low number of metabolites
measured by $^1$H-NMR spectroscopy it is possible to predict who will get an incident CHD
event and that this score gives comparable results to a risk score based on TRFs. However,
these results need to be confirmed in an independent population.

Novel risk makers tested in prospective studies indicate that it possible to improve CHD
risk prediction.29 For example, a coronary artery calcium score, a family history of MI,
ankle brachial index or high-sensitivity C-reactive protein improved risk discrimination and
classification of incident CHD beyond the Framingham score.29 But the decision of using
a novel risk marker should not only be based on improvement in risk discrimination and
reclassification, but also on improvement in clinical outcome and cost-effectiveness.30 It took
information from 52 prospective studies that included 246,669 participants, to estimate that after initial screening with conventional CHD risk factors, the additional assessment of high sensitive C-reactive protein or fibrinogen in people at intermediate risk (i.e. between 10 and 20% chance of developing a CVD event) could help prevent one additional CHD event over a period of 10 years for every 400 to 500 people. This illustrates how difficult it is to improve CHD risk prediction.

**Genetics and CHD risk prediction**

As shown in chapter 5, adding a weighted GRS based on SNPs associated with CHD to a CHD risk score based on TRFs resulted in a slight improvement of CHD risk reclassification. Other studies also investigated if comparable GRSs can improve CHD risk prediction. In general, when using a GRS based on SNPs associated with CHD risk factors, no improvement in CHD risk prediction was observed, which is confirmed in the study described in this thesis. For GRSs based on SNPs associated with CHD the results are mixed. For example, Paynter et al. showed that a non-weighted CHD GRS failed to improve CHD risk prediction, whereas three other studies showed that a GRS based on CHD associated SNPs improved both risk discrimination and/or reclassification. Hughes et al. showed that adding a weighted CHD GRS did not significantly improve risk discrimination beyond the Framingham score, but risk reclassification was significantly improved, which is in line with the results described in this thesis. Concluding so far, GRSs based on SNPs show a minor improvement in CHD risk prediction beyond conventional risk markers but this improvement is not clinically relevant.

On the other hand, Ganna et al. showed that 318 people at intermediate risk (a 10-years risk of CHD between 10% and 20%) needed to be screened with a CHD based GRS to prevent 1 additional CHD event. This is a better outcome than when C-reactive protein would be added to a score based on conventional risk markers, which would prevent 1 additional CHD event per 400-500 individuals at intermediate risk.

One may argue that a family history of CHD can also be a marker of genetic risk, and with a family history you will also capture the lifestyle factors not included in scores based on TRFs like the Framingham score. As described in chapter 5 of this thesis, when adjusting for a parental history of MI, the weighted CHD GRS remained associated with incident CHD in our study. This indicates that a GRS captures some genetic information not available in a parental history of MI. The question if a family history of CHD is a marker of genetic risk was investigated in a study performed by Do et al. In this study it was found that a family history of CHD (only first degree relatives are included) resulted in a better risk discrimination than a GRS based on SNPs associated with CHD. This finding is also supported by a study from So et al., were it was found that 12 CHD associated SNPs combined could explain 25.15% of
the variance in heritability of coronary artery disease.\textsuperscript{40} This means that SNPs discovered in GWAS so far only explain a small proportion the heritability found in CHD.\textsuperscript{41}

Several hypotheses have been proposed to explain this missing heritability problem. According to the infinitesimal model many variants of small effect are responsible for the missing heritability.\textsuperscript{42} The rare allele model assumes that many rare alleles of large effect can explain the missing heritability and the broad sense heritability model assumes that non-additive gene-gene and gene-environment interactions and epigenetic effects can explain the missing heritability.\textsuperscript{43} Recently studies have been published that favour the infinitesimal model.\textsuperscript{44-49} For example, common variants explain 61.8\% of the variance in HDL-C and rare variants only explain 7.8\% of the variance in HDL-C.\textsuperscript{44} Other studies also found that common SNPs explain a significant proportion of the heritability of traits like height,\textsuperscript{45,46} BMI,\textsuperscript{46} rheumatoid arthritis,\textsuperscript{47} and other traits.\textsuperscript{48} These results might favour a polygene approach for disease risk prediction, which already has been proved to work for schizophrenia.\textsuperscript{49} To test this approach for CHD, Ganna and co-workers used a relative small data set (1972 cases and 2891 controls) to select common SNPs with a very small effect size (i.e. SNPs with P-value lower than 0.2).\textsuperscript{37} These SNPs were combined into a weighted polygenic GRS.\textsuperscript{37} When tested in an independent population, this polygenic GRS performed worse than a GRS based on 46 SNPs associated with CHD according to GWAS.\textsuperscript{37} Since Ganna and co-workers used a relative small study for SNP discovery, especially when compared to the large meta-analysis performed to date, it would be interesting to select SNPs with very small effect sizes using large meta-analysis studies. Next, these SNPs need to be tested in independent prospective cohorts to evaluate their usefulness in CHD risk prediction.

**Metabolic health**

In this thesis it is described that lower triglyceride levels characterize familial longevity in females and that familial longevity is characterized by larger LDL particle sizes in males. This indicates that females and males show a different metabolic pattern indicating metabolic health and familial longevity. This is confirmed in a study conducted by Gonzalez et al.\textsuperscript{50} In this study a comprehensive lipid profile, consisting of 128 different lipid species, was obtained in the participants of the LLS.\textsuperscript{50} In males no differences between offspring from long-lived siblings and controls were found.\textsuperscript{50} However, female offspring were characterized by higher levels of ether phosphocholines and lower levels of phosphoethanolamine and long-chain triglyceride species, independent of classical triglyceride levels.\textsuperscript{50} This confirms our observation that a lipid profile indicating metabolic health and familial longevity is gender specific.
In the LLS it was further reported that healthy ageing and metabolic health are characterized by lower glucose levels, higher insulin sensitivity, lower free T3 levels, lower vitamin D levels, and a lower RPTOR gene expression in blood. Since offspring of long-lived siblings are characterized by a lower prevalence of MI and hypertension, it would be interesting to test if markers for metabolic health and familial longevity can be included in scores for CHD risk prediction, especially the lipid profile. These markers for CHD risk prediction should be evaluated in males and females separately.

Combining genetic and metabolomics data
In thesis it is shown that δ-5 desaturase activity is negatively associated with incident CHD, whereas no association between δ-6 desaturase activity and incident CHD was found. δ-5 desaturase is encoded by FADS1. The minor G allele of a SNP (rs174547) located in this gene has been associated with a lower desaturase activity before, but never with an increased CHD risk. Both observations have been confirmed in this thesis. Since the protective effect of a high δ-5 desaturase activity was mostly confined to the carriers of AA genotype, this could indicate that some metabolic markers may only be useful for CHD risk prediction in people with a specific genotype. This indicates that it is relevant to combine genetics and metabolomics data for improving CHD risk prediction. Another reason for combining genetics and metabolomics was shown in a study by Yu and co-workers. In this study it was found that novel SNPs for heart failure can be discovered when applying a GWAS on metabolites previously associated with heart failure. Individually these SNPs were not significantly associated with heart failure, but when combined into a risk score these SNPs predict heart failure. It will be interesting to test if the same approach leads to the discovery of SNPs that can improve risk prediction of CHD, since this approach would also lead to the inclusion of SNPs that are associated with CHD, but not at a genome-wide significance level. This may also be a better strategy than the polygenic approach used by Purcell and co-workers, were only a minority of the included SNPs are truly associated with the disease or phenotype of interest.

In a study performed by Ganna et al. it was found that a CHD specific GRS based on 46 SNPs associated with CHD according to a large meta-GWAS was not associated with the Framingham risk score. This could indicate that these SNPs represent other pathways than the conventional TRFs for increasing CHD risk. It would be interesting to test with which metabolites these SNPs are associated and if these metabolites combined predict CHD risk. Moreover, this approach can also be used to gain knowledge about the CHD pathogenesis.
Conclusions and future research

The risk of CHD is influenced by genetic, environmental and lifestyle factors. The current scores for estimating CHD risk in individuals free from CVD do not incorporate metabolomics nor genetic information. Therefore we evaluated the usefulness of these types of data for CHD risk prediction. However, based on the results described in this thesis it remains inconclusive if a metabolomics profile based on \(^1\)H-NMR spectroscopy or a GRS based on SNPs associated with CHD can improve 10-year CHD risk prediction beyond the currently used risk scores based on TRFs only. Based on the results described in this thesis it can be argued that markers for healthy ageing such as a LDL particle size should also be evaluated for their usefulness in CHD risk prediction and that these markers should be evaluated in males and females separately.

Most scores developed to predict CHD or CVD risk are developed to estimate CHD or CVD risk in individuals at middle age free from CVD. These scores cannot be used to estimate CHD risk in elderly individuals. For example a high blood pressure is associated with less mortality and strokes in the healthy elderly.\(^{57,58}\) This means that the field of CVD risk prediction in the 85+ part of the population is relatively unexplained. The metabolomics platforms described in this thesis could be used to discover if metabolites could indicate metabolic health, co-morbidity and mortality in the oldest old. In this thesis it is described that metabolic health is characterized by different markers in males and females, therefore such analysis should done in females and males separately.

In a study conducted by Pencina \textit{et al.} a cohort with individuals between 20 and 60 years old was followed for 30 years.\(^{59}\) It was found that an adverse risk profile leads to a high 30-year CVD risk in an individual of 25 years.\(^{59}\) When using a follow-up time of 10 years, this individual would have been characterized as having a very low CVD risk.\(^{59}\) Especially within a young population health gains can be achieved by using intervention strategies. That is why we need large cohorts with a younger population and longer follow-up times than currently used for finding risk markers for CHD. Since genetic markers remain stable throughout live, it is possible to use genetic markers for CHD risk prediction much earlier in life than conventional CHD risk factors.\(^{37}\) This hypothesis can be investigated in the CAREMA cohort using a case-cohort study design.\(^{60}\) Cardiologic follow-up for this study was performed in December 2003 and only for the participants between 30 and 59 years of age at moment of inclusion. At that point the median follow-up time was 12.1 years. If cardiologic follow-up would be performed again in December 2013, including the participants between 20 and 30 years of age at moment of inclusion, this would result in a study with a longer follow-up time (approximately 22.1 years). Moreover the study population would be younger. In this younger population with a longer follow-up time the hypothesis that genetic markers have
more value at a younger age for CHD risk prediction can be tested for relative low costs, because only additional samples need to be analysed.

It can be argued that to prevent CHD, instead of estimating CHD risk and treat high risk individuals, everybody of 55 years and older should be prescribed a polypill.\textsuperscript{61} This pill contains a statin to lower cholesterol levels, three blood pressure lowering drugs, folic acid and aspirin.\textsuperscript{61} It was estimated that such a pill would prevent 88\% of heart attacks and 80\% of strokes after two years of treatment.\textsuperscript{61} Several randomized clinical trials have investigated the efficacy of this polypill and were summarized in a meta-analysis.\textsuperscript{62} According to this meta-analysis, the use of the polypill reduced systolic blood pressure total cholesterol and LDL-cholesterol levels, but this reduction was less than previously estimated by Wald \textit{et al.}\textsuperscript{62} This can be explained by the fact that the actual reduction in risk factors is dependent on the baseline levels of these risk factors.\textsuperscript{62} Thus it could be that the polypill is not as effective as expected in persons with a low cardiovascular risk. On the other hand, according to van Gils \textit{et al.} the polypill strategy for primary prevention in the Netherlands is a cost effective strategy when given to participants of 40 years and older with a cardiovascular risk of more the 5\%.\textsuperscript{63} Thus the combination of screening for cardiovascular risk with prescribing a polypill if this risk exceeds 5\% might be an effective strategy.\textsuperscript{63}

To get a better understanding of CHD pathology, it makes sense to integrate different omics datasets together with lifestyle and environmental factors. Moreover this type of data should be gathered in longitudinal cohorts in whom follow-up information on CHD or other metabolic disease is available or can be collected. Blood should not be only collected at baseline but also during follow-up (repeated measures design). This will make it possible to do a time-dependent analysis were it is possible to account for changes in risk factor levels (for example stopping with smoking or changes in BMI).\textsuperscript{59}
References


23. Mora S, Otvos JD, Rifai N, Rosenson RS, Buring JE, Ridker PM. Lipoprotein particle profiles by nuclear magnetic resonance compared with standard lipids and apolipoproteins in predicting incident


45. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of the herita-


