

Cover Page



Universiteit Leiden



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

Genetic and metabolomic approaches for coronary heart disease risk prediction.

Anika Antoinette Maria Vaarhorst

Genetic and metabolomic approaches for coronary heart disease risk prediction.

Drs. Anika A.M. Vaarhorst

PhD thesis with summary in Dutch

ISBN: 978-90-8891-869-8

Cover design: Proefschriftmaken.nl || Uitgeverij BOXpress

Printed & Lay Out by: Proefschriftmaken.nl || Uitgeverij BOXpress

Published by: Uitgeverij BOXpress, 's-Hertogenbosch

Genetic and metabolomic approaches for coronary heart disease risk prediction.

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College van Promoties
te verdedigen op dinsdag 13 mei 2014
Klokke 16.15 uur

door

Anika Antoinette Maria Vaarhorst
Geboren te Enschede
in 1981

Promotiecommissie

Promotor: Prof. Dr. P. Slagboom

Co-promotor: Dr. B.T. Heijmans

Overige leden: Prof. Dr. H. Putter

Prof. Dr. E.J.M. Feskens
Wageningen University

Prof. Dr. Y.T. van der Schouw
University Medical Center Utrecht

The research described in this thesis was supported by a grant of the Dutch Heart Foundation (DHF-2006B195) and by BBMRI-NL, a Research Infrastructure financed by the Dutch government (NWO 184.021.007).

Financial support by the Dutch Heart Foundation for the publication of this thesis is gratefully acknowledged. Financial support for the printing of this thesis was provided by the Netherlands Consortium for Healthy Ageing (GI 050060810).

Contents

Chapter 1	General Introduction	7
Chapter 2	A metabolomics profile is associated with the risk of incident coronary heart disease	21
Chapter 3	CHD risk prediction using regularized Cox regression: Comparing a targeted and an untargeted approach for extracting information from ¹ H-NMR spectra	37
Chapter 4	Lipid metabolism in long-lived families: The Leiden Longevity Study	49
Chapter 5	Literature-based genetic risk scores for coronary heart disease; the CAREMA prospective-cohort study	63
Chapter 6	Markers of endogenous desaturase activity and risk of coronary heart disease in the CAREMA cohort Study	81
Chapter 7	Summary and general discussion	97
Chapter 8	Nederlandse samenvatting	111
	Supplement I	119
	Supplement II	131
	Supplement III	133
	Supplement IV	155
	Dankwoord	161
	Curriculum Vitae	163
	List of publications	165

General Introduction

What is coronary heart disease?

Coronary heart disease (CHD) arises from atherosclerosis in the coronary arteries.^{1,2} Atherosclerosis is a slowly progressing disease resulting in the formation of atherosclerotic plaques and narrowing of the arteries.^{1,2} Rupture of an atherosclerotic plaque leads to thrombus formation and temporal or permanent obstruction of the coronary artery.^{1,2} As a consequence, the underlying tissue is no longer supplied with oxygen and nutrients, resulting in ischemia and possible necrosis of myocardial tissue.^{1,2} The patient experiences this as chest pain that may radiate to the neck, back, or arms, which is usually accompanied with nausea, excessive sweating, shortness of breath and fear of death.^{3,4} Although 20% of the patients (mostly women, diabetics, postoperative patients and elderly people) with a temporal or permanent obstruction of one of the coronary arteries lack chest pain and experience atypical symptoms.^{3,4}

Acute CHD can appear as an acute myocardial infarction (MI) or unstable angina pectoris (UAP).^{3,4} If a patient experiences chest pain and/or other complaints concerning an obstruction of the coronary arteries an electrocardiogram (ECG) should be performed.^{3,4} As shown in figure 1, an abnormal ECG is indicative for CHD, but is not sufficient by itself to distinguish between acute MI and UAP.⁵ The most important difference between acute MI and UAP is the presence of necrosis of myocardial tissue.³ Therefore, as shown in figure 1, the final diagnosis of acute MI or UAP depends on the detection of proteins that are released in the blood due to myocardial tissue necrosis.^{4,5}

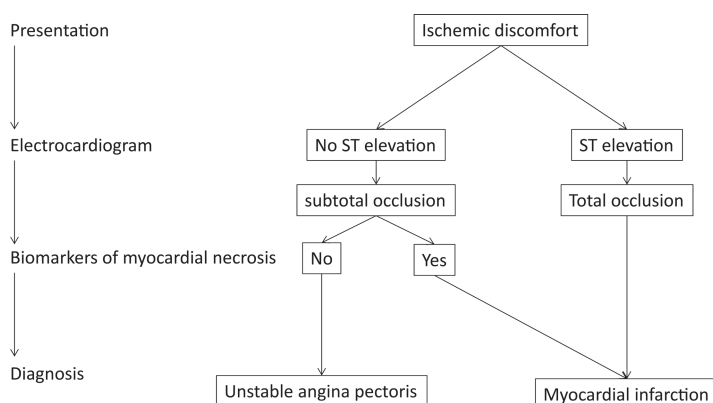


Figure 1 | A schematic representation illustrating the steps needed to diagnose acute myocardial infarction or unstable angina pectoris in patients experiencing ischemic discomfort (adapted from figure 1 in Wright et al. 2011 JACC).⁵

Mortality of coronary heart disease

CHD is the main cause of death worldwide.⁶ In the Netherlands, 10,400 individuals died because of CHD (6000 males and 4400 females), accounting for 7.6% of all deaths in 2010, making CHD the second most important cause of death in the Netherlands (see table 1).⁷

Table 1 | Top ten of diseases with the highest mortality in the Netherlands in 2011 (www.nationaalkompas.nl)⁷

Disease/illness	Males	Females	Combined
Lung cancer	6.586	3.958	10.544
Coronary heart disease	5.724	4.152	9.876
Dementia (including Alzheimer)	2.544	6.617	9.161
Stroke	3.315	5.165	8.480
Heart failure	2.499	3.999	6.498
COPD	3.466	2.887	6.353
Infections of the lower respiratory tract	2.613	3.068	5.681
Colon cancer	2.619	2.508	5.127
Breast cancer	22	3.261	3.283
Private accidents	1.246	1.575	2.821
All causes of death combined	65.259	70.482	135.741

COPD, Chronic Obstructive Pulmonary Disease

Since 1970 CHD mortality is declining in western countries,^{8,9} including the Netherlands.¹⁰ This can be explained by technological advances such as bypass surgery and percutaneous coronary interventions, which prevent death during an acute manifestation of CHD.¹¹ Second, there are interventions that prolong survival once CHD is manifest (secondary prevention).¹² This includes stimulating a healthy lifestyle and medical treatment to reduce high cholesterol levels and high blood pressure in individuals diagnosed with CHD.¹² And finally by primary prevention aimed at preventing CHD in individuals that never had a CHD event.¹² This can be done by stimulating a healthy lifestyle in the general population but also by treating individuals with a high risk for CHD with the same interventions that are used for secondary prevention.^{12,13} For the latter approach to be successful it is important to select only the individuals at high risk for CHD. To do so risk scores have been developed to select these high risk individuals.

Risk scores for coronary heart disease

In 1948 the Framingham Heart Study was initiated, the goal of this study was to identify risk factors for CHD.¹⁴ At that point it was already assumed that CHD was a multifactor disease

that developed over a long time.¹⁴ The investigators included 5209 individuals, free from CHD aged between 30 and 59 years living in Framingham, Massachusetts and hoped that they could follow the included individuals for as long as 20 years.¹⁴ It was expected that by comparing individuals that developed CHD with individuals that remained healthy during the course of time, risk factors for CHD could be found.¹⁴

In 1957, the first report on the Framingham study was published.¹⁵ It was found that older age, male sex, high blood pressure, obesity and high cholesterol levels were associated with atherosclerotic heart disease.¹⁵ Twenty years later a general cardiovascular risk profile was proposed.¹⁶ This risk profile included the joint effects of gender, age, systolic blood pressure, serum cholesterol levels, cigarette smoking, electrocardiographic evidence of left ventricular hypertrophy and glucose intolerance.¹⁶ The Framingham score currently used is simplified,¹⁷ and based on a larger and more recent population (*i.e.* the original Framingham cohort plus their offspring).¹⁸ This simplified version includes the traditional risk factors (TRFs) sex, age, total cholesterol or LDL cholesterol, systolic blood pressure, smoking and self-reported diabetes.¹⁷ Left ventricular hypertrophy was no longer included, since it was highly correlated with hypertension.¹⁷ To date several other risk scores have been developed to predict the risk for cardiovascular disease¹⁹⁻²² and CHD²³⁻²⁷ in individuals free from cardiovascular disease. Of these, the SCORE function is used to assess CHD risk in individuals free from cardiovascular disease in the Netherlands.^{28,29}

The performance of these risk scores can be estimated by the *C* statistic,³⁰ which is also known as the area under the receiver operator characteristic (ROC) curve³¹ or the *C*-index in case of survival data.³² The *C*-index estimates the probability that of two randomly chosen individuals, the one with a higher score will develop CHD earlier than the one with a lower score.³² The *C*-index can vary between 0.5 (50% of the cases are correctly ranked, meaning that the test is not better than flipping a coin) and 1 (all the cases are correctly ranked).³² For the Framingham risk score and comparable scores the *C*-index is around 0.75,³³ meaning that in 75% of the cases, two randomly chosen individuals are correctly ranked.³⁰ This indicates that there is room for improvement of the currently used CHD risks scores, which can be done by adding novel risk markers.

Metabolomics and coronary heart disease

Since atherosclerosis is a disease of the arteries it is reasonable to assume that most CHD risk factors can be found in blood. However of the risk factors to be found in blood the current risk scores only use total cholesterol, HDL-cholesterol and triglyceride levels as risk markers. There are a many other compounds to be found in blood, for example free fatty acids, glucose, amino acids, lipid particles that can all play a role in the development of

atherosclerosis and thus CHD. By applying metabolomics technology it is possible to acquire information about the blood metabolome by identifying and quantifying metabolites in a non-biased way.^{34,35} To date more than 4000 different metabolites are already identified and quantified in human blood, and this number is still increasing.³⁶ These metabolites reveal information about the metabolic state of an individual and this information might be useful for understanding, diagnosing or predicting CHD.

There are different techniques available for identifying and quantifying these metabolites.³⁶ Proton nuclear magnetic resonance (¹H-NMR) spectroscopy and chromatography coupled to mass spectrometry (MS) based platforms are mostly used.^{37,38} See box 1 for a description of both techniques and their advantages and disadvantages.^{37,38} ¹H-NMR spectroscopy is a

Box 1 | Techniques used for measuring metabolites at a large scale

The two most commonly used methods to obtain information about a person's metabolome in one single measurement are chromatography coupled to mass spectrometry (MS) and proton nuclear magnetic resonance (¹H-NMR) spectroscopy.^{37,38}

With chromatography coupled to MS, the compounds within the samples are first separated by separation techniques based on liquid or gas chromatography.⁴¹ Next, to identify these compounds the samples are placed in the mass spectrometer and ionized, the resulting charged molecules (ions) are then separated by their mass to charge ratio.⁴¹ Next, these ions fly into the detector, which records the current induced when this happens.⁴¹ This information is visualized in a mass spectrum.⁴¹ Depending on the location and the heights of the peaks in the mass spectrum, information about the identity and the abundance of the molecules in the biological sample can be deduced.⁴¹

¹H-NMR spectroscopy is used to find clues about the structure of unknown molecules.^{37,42} ¹H stands for hydrogen-1, which is the most common hydrogen isotope.⁴² This hydrogen isotope contains 1 proton and 1 electron and absorbs electromagnetic radiation at a specific frequency or resonance, depending on the force of the magnetic field used.⁴² In a ¹H-NMR spectrometer, the electromagnetic radiation is fixed at a specific frequency but the magnetic field can be varied from high to low.⁴² When hydrogen atoms absorb electromagnetic radiation, this is visualized by peaks in the ¹H-NMR spectrum.⁴² The location of the peaks depends on the chemical environment (*i.e.* the other atoms located nearby the hydrogen atom) of the hydrogen atoms.⁴² As a result a high or a low magnetic field is needed for the hydrogen atoms to absorb electromagnetic radiation.⁴² The heights and areas of the peaks reveal information about the number of hydrogen atoms in these specific chemical environments.⁴² Based on this information it is possible to identify molecules in a biological sample.⁴²

Table | The advantages and disadvantages of ¹H-NMR and chromatography coupled to MS based techniques.⁴³

Method	Advantages/Disadvantages
¹ H-NMR spectroscopy	+highly reproducible +fast -relatively few metabolites can be identified and quantified -low sensitivity, only high abundant metabolites can be detected
Chromatography coupled to MS	+many different metabolites can be identified and quantified +high sensitivity, also low abundant metabolites can be detected -slow -sample preparation is difficult

robust and reliable technique that is highly reproducible,³⁹ which is a prerequisite for the identification of novel biomarkers that may be used for CHD risk prediction improvement.⁴⁰ Moreover, ¹H-NMR spectroscopy is fast, especially when compared to chromatography coupled to MS based platforms. The downside of ¹H-NMR spectroscopy is its low sensitivity compared to MS based platforms. For example with ¹H-NMR spectroscopy it is possible to identify and quantify up to 44 different metabolites in blood, whereas with chromatography-MS based methods up to hundreds of different metabolites can be identified and quantified.³⁶ Thus with ¹H-NMR spectroscopy it is possible to generate a robust and reliable profile of the metabolome in large study populations in relatively little time and it is hoped that this profile can be used to improve CHD risk prediction beyond TRFs.

Genetics and coronary heart disease

Twin, family and adoption studies showed that CHD has a heritable component.⁴⁴⁻⁴⁷ In a Swedish twin-registry study it was found that in males 57% of the variance in death from CHD can be explained by heritable factors and for females this was 38%.⁴⁶ The proportion in CHD death explained by a non-shared environment was 43% for males and 62% for females, the shared environment did not explain any of the variance found.⁴⁶ In a Danish twin study comparable results were found.⁴⁷ Thus CHD is caused by a combination of genetic and environmental risk factors. Not only CHD, but also CHD risk factors have a large heritable component, including smoking,⁴⁸ blood pressure,⁴⁹ and blood lipid levels.⁵⁰

In 1996, Risch and Merikangas proposed that finding genetic variants associated with common complex disease in the common population required large scale testing by association analysis.⁵¹ With the completion of a physical map of the human genome in 2001,⁵² and the identification of more than 1.4 million single nucleotide polymorphisms (SNPs)⁵³ it became possible to associate diseases and/or traits with common genetic variants on a genome-wide scale. These studies are commonly referred to as genome-wide association studies (GWAS).

In 2009 a catalog with published GWAS came online.⁵⁴ On 13/12/12 this catalog (<http://www.genome.gov/gwastudies/>) contained 1459 publications and 8084 SNPs. Of these studies, 153 publications are related to CHD or CHD related risk factors (*i.e.* blood pressure, blood lipid levels, diabetes type and anthropomorphic related traits). Together these publications contain 914 SNPs associated with CHD or CHD risk factors.

The GWAS approach has delivered many SNPs for complex traits and with the progress made through meta-analysis leading to stable risk estimates for each SNP, the usefulness of these SNPs for CHD risk prediction should be tested. A limitation of GWAS is that every identified SNP has a very small effect size.⁵⁵ Because of the small effect sizes of SNPs

associated with CHD or its risk factors, SNPs were combined into a genetic risk score (GRS).⁵⁶⁻⁵⁸ It was expected that such a GRS, based on SNPs associated with CHD or CHD risk factors is associated with incident CHD and perhaps can improve CHD risk prediction beyond the TRFs on which current CHD risk scores are based.⁵⁶⁻⁵⁸

Metabolic health

To find risk markers that may improve CHD risk prediction, most studies focus on finding risk factors for CHD. One may question if a low CHD risk is simply caused by the absence of risk factors or that there are protective factors involved, protecting individuals against CHD even in the presence of established CHD risk factors. To answer this question it is relevant to study which factors mark metabolic health.

This can be done in the offspring of long-lived individuals that are characterized by a lower prevalence of age-related diseases, including diabetes, hypertension CHD and stroke.⁵⁹⁻⁶² This indicates that these long-lived individuals have a genetic or familial factor that somehow protects them against the development of metabolic diseases, which make them interesting for studying which factors are involved in healthy ageing and longevity and thus possible markers that are protective against CHD.

Box 2 | Studies used in this thesis

Cardiovascular Registry Maastricht cohort study - The Cardiovascular Registry Maastricht (CAREMA) cohort study was set up to study coronary heart disease (CHD) and associated risk factors.⁶³ The cohort consists of 21,148 participants born between 1927 and 1977 who were randomly sampled from Maastricht and surrounding communities between 1987 and 1997.⁶³ This cohort was followed until December 2003. During this follow-up period, 742 incident CHD cases (*i.e.* acute myocardial infarction, unstable angina pectoris or death due to CHD) occurred.^{63,64} From the CAREMA cohort a sub-cohort was drawn, consisting of 2221 participants, including 116 cases. Only the cases experiencing a CHD event and the members of the sub-cohort are included in the data-analysis.⁶⁵ This type of study design is called a case-cohort study.⁶⁵

Erasmus Rucphen Family study - The Erasmus Rucphen Family (ERF) study is a large, family-based study with participants from a genetically isolated community located in the Southwest of the Netherlands.⁶⁶ The ERF study includes individuals that are the living descendants of 22 couples that had at least six children baptized in the community church between 1850 and 1900. Participants were thus not selected based on disease of interest.⁶⁶ Participants were classified as a CHD case when they indicated that they had experienced a myocardial infarction, underwent a coronary revascularization procedure or if they reported angina symptoms during an interview with the study physician. Participants were also classified as a CHD case if they showed signs of myocardial infarction on the electrocardiogram. In this study of 2415 participants, 170 participants were classified as a CHD case and 2157 participants were categorized as controls.

Leiden Longevity Study - This study was set up to investigate if there was evidence of genetic enrichment for exceptional survival using a family approach.⁶⁷ For this study large series of sib pairs aged 90 years and over were collected.⁶⁷ In addition the offspring of the long-lived siblings and the partners of the offspring were also collected for further study.⁶⁷ The recruitment for this study started in July 2002.⁶⁷ In total 421 families were included, representing 982 long-lived siblings, their offspring (n=1671) and the partners of the offspring (n=745).

Aims of this thesis

The main goal of this thesis was to assess if CHD risk prediction based on the TRFs could be improved by adding novel risk markers. First it was tested if metabolomics data could improve CHD risk prediction beyond TRFs. Second it was investigated if genetic markers could improve CHD risk prediction. A secondary goal was to find out if a comprehensive lipid profile could mark metabolic health as opposite to age-related diseases like CHD in middle-aged individuals. Finally, as a first step in understanding more of CHD pathology we combined metabolomics data (a comprehensive fatty acid profile) with genetic information on the FADS1 gene. See box 2 for the studies used to investigate these questions.

Outline of this thesis

In **chapter 2** we investigated if a metabolomics approach could be useful for improving and understanding CHD risk prediction beyond TRFs. Therefore we obtained a metabolic profile using ^1H -NMR spectroscopy in the individuals who entered the Cardiovascular Registry Maastricht (CAREMA) cohort study from 1993 until 1997 and were selected for the sub-cohort or experienced a CHD event during follow-up. Using least absolute shrinkage and selection operator (LASSO) regression,^{68,69} a subset of the most informative metabolites for predicting incident CHD was obtained. Subsequently, it was tested if a score based on these metabolites was associated with incident CHD within the same case-cohort study. Using the Erasmus Rucphen Family (ERF) study it was tested if the selected metabolites were relevant for differentiating between prevalent CHD cases and CHD-free controls.

In **chapter 3** we compared two methods for data reduction, our own targeted approach explained in chapter 2 and a more commonly used method which involves partitioning the ^1H -NMR spectra into discrete sections and calculate the sum of the spectral intensity in each section.^{70,71} Moreover, we also evaluated the impact of two scaling methods (*i.e.* autoscaling and pareto scaling) which are commonly used as data pre-treatment methods in metabolomics data-analysis.⁷²

In **chapter 4**, as apposite to the development of CHD, factors that promote longevity and metabolic health were studied. Therefore offspring of long-lived siblings were compared to their same-aged spouses representing the general population. Siblings are considered long-lived when they are 90 years or older and have a brother or sister who is also 90 years or older.⁶⁷ These long-lived siblings could not be easily compared a control group since most subjects in their birth cohort were already deceased. However at middle-age their offspring also showed less overall mortality, and metabolic disease compared to the general Dutch population.⁶⁷ Therefore the offspring of long-lived siblings were compared to their same

aged partners. With this approach it is possible to study the differences between people that have a tendency to become long-lived and people with an average life expectancy. The focus for this chapter was on lipid metabolism. Thus a comprehensive lipid profile was obtained using ^1H -NMR spectroscopy.⁷³

In **chapter 5** it was investigated if GWAS identified SNPs known to be associated with CHD or CHD risk factors (*i.e.* blood pressure, diabetes type 2, blood lipid levels and anthropomorphic related traits) could improve CHD risk prediction beyond traditional CHD risk factors. Based on these SNPs GRSs were constructed. First we focused on counted GRSs. The first counted GRS was based on SNPs associated with CHD and CHD risk factors, the second one was based on SNPs associated with CHD risk factors and the third one was based on SNPs associated with CHD only. The fourth GRS was also based on SNPs associated with CHD, but in contrast to the counted GRS, we adjusted for the different effect sizes of the CHD associated SNPs. The effect sizes for these SNPs were based on two large meta-analysis.^{74,75} As an exploratory approach, we selected a subset of the most informative SNPs using LASSO regression.

In **chapter 6**, the fatty acid metabolism was explored for gaining more understanding in the development of CHD. Lipid metabolism, including fatty acid metabolism plays an important role in CHD etiology. For example, a high intake of polyunsaturated fatty acids reduces CHD risk.⁷⁶⁻⁷⁸ The fatty acid intake is often estimated by assessing the fatty acid contents in biological tissues.⁷⁹ This profile, however, does not only reflect dietary intake, but also fatty acid metabolism in the body itself,⁷⁹ in which δ -5 and δ -6 desaturase play an important role.^{80,81} In this chapter we explored the role of δ -5 and δ -6 desaturase in CHD, in addition we investigated the influence of a genetic variation in the fatty acid desaturase 1 gene.

References

1. Lusis AJ. Atherosclerosis. *Nature*. Sep 14 2000;407(6801):233-241.
2. Weber C, Noels H. Atherosclerosis: current pathogenesis and therapeutic options. *Nature medicine*. 2011;17(11):1410-1422.
3. Welch TD, Yang EH, Reeder GS, Gersh BJ. Modern management of acute myocardial infarction. *Current problems in cardiology*. Jul 2012;37(7):237-310.
4. Myocardial infarction redefined--a consensus document of The Joint European Society of Cardiology/American College of Cardiology Committee for the redefinition of myocardial infarction. *European heart journal*. Sep 2000;21(18):1502-1513.
5. Wright RS, Anderson JL, Adams CD, et al. 2011 ACCF/AHA focused update incorporated into the ACC/AHA 2007 Guidelines for the Management of Patients with Unstable Angina/Non-ST-Elevation Myocardial Infarction: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines developed in collaboration with the American Academy of Family Physicians, Society for Cardiovascular Angiography and Interventions, and the Society of Thoracic Surgeons. *Journal of the American College of Cardiology*. May 10 2011;57(19):e215-367.
6. World Health Organization. The top 10 causes of death. 2013; <http://www.who.int/mediacentre/factsheets/fs310/en/>. Accessed 26-02-2014, 2014.
7. Poos M. Sterfte naar doodsoorzaak: Waaraan overlijden mensen in Nederland? 2013; <http://www.nationaalkompas.nl/gezondheid-en-ziekte/sterfte-levensverwachting-en-daly-s/sterfte-naar-doodsoorzaak/waaraan-overlijden-mensen-in-nederland/>. Accessed 26-02-2014, 2014.
8. Thom TJ. International Mortality from Heart-Disease - Rates and Trends. *International Journal of Epidemiology*. 1989;18:S20-S28.
9. Nichols M, Townsend N, Scarborough P, Rayner M. Trends in age-specific coronary heart disease mortality in the European Union over three decades: 1980-2009. *European heart journal*. Oct 2013;34(39):3017-3027.
10. Reitsma JB, Dalstra JA, Bonsel GJ, et al. Cardiovascular disease in the Netherlands, 1975 to 1995: decline in mortality, but increasing numbers of patients with chronic conditions. *Heart*. Jul 1999;82(1):52-56.
11. Gaziano TA, Bitton A, Anand S, Abrahams-Gessel S, Murphy A. Growing epidemic of coronary heart disease in low- and middle-income countries. *Current problems in cardiology*. Feb 2010;35(2):72-115.
12. Ford ES, Capewell S. Proportion of the decline in cardiovascular mortality disease due to prevention versus treatment: public health versus clinical care. *Annual review of public health*. 2011;32:5-22.
13. Smolina K, Wright FL, Rayner M, Goldacre MJ. Determinants of the decline in mortality from acute myocardial infarction in England between 2002 and 2010: linked national database study. *Bmj*. 2012;344:d8059.
14. Dawber TR, Meadors GF, Moore FE, Jr. Epidemiological approaches to heart disease: the Framingham Study. *American journal of public health and the nation's health*. Mar 1951;41(3):279-281.
15. Dawber TR, Moore FE, Mann GV. Coronary heart disease in the Framingham study. *American journal of public health and the nation's health*. Apr 1957;47(4 Pt 2):4-24.
16. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile: the Framingham Study. *The American journal of cardiology*. Jul 1976;38(1):46-51.
17. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation*. May 12 1998;97(18):1837-1847.
18. Anderson KM, Wilson PW, Odell PM, Kannel WB. An updated coronary risk profile. A statement for health professionals. *Circulation*. Jan 1991;83(1):356-362.
19. Conroy R. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *European heart journal*. 2003;24(11):987-1003.
20. Ridker PM, Paynter NP, Rifai N, Gaziano JM, Cook NR. C-reactive protein and parental history improve global cardiovascular risk prediction: the Reynolds Risk Score for men. *Circulation*. Nov 25 2008;118(22):2243-2251, 2244p following 2251.
21. Woodward M, Brindle P, Tunstall-Pedoe H, estimation Sgor. Adding social deprivation and family history to cardiovascular risk assessment: the AS-SIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart*. Feb 2007;93(2):172-176.
22. Balkau B, Hu G, Qiao Q, et al. Prediction of the risk of cardiovascular mortality using a score that in-

- cludes glucose as a risk factor. The DECODE Study. *Diabetologia*. Dec 2004;47(12):2118-2128.
23. Tunstall-Pedoe H. The Dundee coronary risk-disk for management of change in risk factors. *Bmj*. Sep 28 1991;303(6805):744-747.
 24. Menotti A, Lanti M, Puddu PE, Kromhout D. Coronary heart disease incidence in northern and southern European populations: a reanalysis of the seven countries study for a European coronary risk chart. *Heart*. Sep 2000;84(3):238-244.
 25. Stevens RJ, Kothari V, Adler AI, Stratton IM, Holman RR, Grp U. The UKPDS risk engine: a model for the risk of coronary heart disease in Type II diabetes (UKPDS 56). *Clin Sci*. Dec 2001;101(6):671-679.
 26. Chambless LE, Folsom AR, Sharrett AR, et al. Coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC) study. *Journal of Clinical Epidemiology*. 2003;56(9):880-890.
 27. Assmann G, Schulte H, Cullen P, Seedorf U. Assessing risk of myocardial infarction and stroke: new data from the Prospective Cardiovascular Munster (PROCAM) study. *European journal of clinical investigation*. Dec 2007;37(12):925-932.
 28. Smulders YM, Burgers JS, Scheltens T, et al. Clinical practice guideline for cardiovascular risk management in the Netherlands. *The Netherlands journal of medicine*. Apr 2008;66(4):169-174.
 29. Graham I, Atar D, Borch-Johnsen K, et al. European guidelines on cardiovascular disease prevention in clinical practice: executive summary: Fourth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (Constituted by representatives of nine societies and by invited experts). *European heart journal*. Oct 2007;28(19):2375-2414.
 30. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*. Feb 20 2007;115(7):928-935.
 31. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. Apr 1982;143(1):29-36.
 32. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA : the journal of the American Medical Association*. May 14 1982;247(18):2543-2546.
 33. Cook NR, Paynter NP, Eaton CB, et al. Comparison of the Framingham and Reynolds Risk scores for global cardiovascular risk prediction in the multi-ethnic Women's Health Initiative. *Circulation*. Apr 10 2012;125(14):1748-1756, S1741-1711.
 34. Ellis DI, Dunn WB, Griffin JL, Allwood JW, Goodacre R. Metabolic fingerprinting as a diagnostic tool. *Pharmacogenomics*. Sep 2007;8(9):1243-1266.
 35. Nicholson JK, Lindon JC. Systems biology: Metabonomics. *Nature*. Oct 23 2008;455(7216):1054-1056.
 36. Psychogios N, Hau DD, Peng J, et al. The human serum metabolome. *PLoS one*. 2011;6(2):e16957.
 37. Beckonert O, Keun HC, Ebbels TM, et al. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature protocols*. 2007;2(11):2692-2703.
 38. Dunn WB, Broadhurst D, Begley P, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature protocols*. Jul 2011;6(7):1060-1083.
 39. Dumas ME, Maibaum EC, Teague C, et al. Assessment of analytical reproducibility of ¹H NMR spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP Study. *Analytical chemistry*. Apr 1 2006;78(7):2199-2208.
 40. Dent TH. Predicting the risk of coronary heart disease. II: the role of novel molecular biomarkers and genetics in estimating risk, and the future of risk prediction. *Atherosclerosis*. Dec 2010;213(2):352-362.
 41. Mishur RJ, Rea SL. Applications of mass spectrometry to metabolomics and metabonomics: detection of biomarkers of aging and of age-related diseases. *Mass spectrometry reviews*. Jan-Feb 2012;31(1):70-95.
 42. Keeler J. *Understanding NMR Spectroscopy*. Wiley; 2013.
 43. Mayr M. Metabolomics: ready for the prime time? *Circulation. Cardiovascular genetics*. Oct 2008;1(1):58-65.
 44. Marenberg ME, Risch N, Berkman LF, Floderus B, de Faire U. Genetic susceptibility to death from coronary heart disease in a study of twins. *The New England journal of medicine*. Apr 14 1994;330(15):1041-1046.
 45. Zdravkovic S, Wienie A, Pedersen NL, Marenberg ME, Yashin AI, De Faire U. Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. *Journal of internal medicine*. Sep 2002;252(3):247-254.
 46. Sundquist K, Winkleby M, Li X, Ji J, Hemminki K, Sundquist J. Familial transmission of coro-

- nary heart disease: a cohort study of 80,214 Swedish adoptees linked to their biological and adoptive parents. *American heart journal*. Aug 2011;162(2):317-323.
47. Wienke A, Holm NV, Skytthe A, Yashin AI. The heritability of mortality due to heart diseases: a correlated frailty model applied to Danish twins. *Twin research : the official journal of the International Society for Twin Studies*. Aug 2001;4(4):266-274.
48. Boomsma DI, Koopmans JR, Van Doornen LJ, Orlebeke JF. Genetic and social influences on starting to smoke: a study of Dutch adolescent twins and their parents. *Addiction*. Feb 1994;89(2):219-226.
49. Evans A, Van Baal GC, McCarron P, et al. The genetics of coronary heart disease: the contribution of twin studies. *Twin research : the official journal of the International Society for Twin Studies*. Oct 2003;6(5):432-441.
50. Beekman M, Heijmans BT, Martin NG, et al. Heritabilities of apolipoprotein and lipid levels in three countries. *Twin research : the official journal of the International Society for Twin Studies*. Apr 2002;5(2):87-97.
51. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science*. Sep 13 1996;273(5281):1516-1517.
52. McPherson JD, Marra M, Hillier L, et al. A physical map of the human genome. *Nature*. Feb 15 2001;409(6822):934-941.
53. Sachidanandam R, Weissman D, Schmidt SC, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*. Feb 15 2001;409(6822):928-933.
54. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*. Jun 9 2009;106(23):9362-9367.
55. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature genetics*. Mar 2003;33 Suppl:228-237.
56. Aulchenko YS, Ripatti S, Lindqvist I, et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nature genetics*. Jan 2009;41(1):47-55.
57. Paynter NP, Chasman DI, Pare G, et al. Association between a literature-based genetic risk score and cardiovascular events in women. *JAMA : the journal of the American Medical Association*. Feb 17 2010;303(7):631-637.
58. Morrison AC, Bare LA, Chambless LE, et al. Prediction of coronary heart disease risk using a genetic risk score: the Atherosclerosis Risk in Communities Study. *American journal of epidemiology*. Jul 1 2007;166(1):28-35.
59. Adams ER, Nolan VG, Andersen SL, Perls TT, Terry DF. Centenarian offspring: start healthier and stay healthier. *Journal of the American Geriatrics Society*. Nov 2008;56(11):2089-2092.
60. Westendorp RG, van Heemst D, Rozing MP, et al. Nonagenarian siblings and their offspring display lower risk of mortality and morbidity than sporadic nonagenarians: The Leiden Longevity Study. *Journal of the American Geriatrics Society*. Sep 2009;57(9):1634-1637.
61. Florez H, Ma Y, Crandall JP, et al. Parental longevity and diabetes risk in the Diabetes Prevention Program. *The journals of gerontology. Series A, Biological sciences and medical sciences*. Nov 2011;66(11):1211-1217.
62. Terry DF, Wilcox MA, McCormick MA, Perls TT. Cardiovascular disease delay in centenarian offspring. *The journals of gerontology. Series A, Biological sciences and medical sciences*. Apr 2004;59(4):385-389.
63. Merry AH, Boer JM, Schouten LJ, et al. Validity of coronary heart diseases and heart failure based on hospital discharge and mortality data in the Netherlands using the cardiovascular registry Maastricht cohort study. *European journal of epidemiology*. 2009;24(5):237-247.
64. Merry AH, Boer JM, Schouten LJ, et al. Risk prediction of incident coronary heart disease in The Netherlands: re-estimation and improvement of the SCORE risk function. *European journal of preventive cardiology*. Aug 2012;19(4):840-848.
65. Prentice RL. A Case-Cohort Design for Epidemiologic Cohort Studies and Disease Prevention Trials. *Biometrika*. Apr 1986;73(1):1-11.
66. Sayed-Tabatabaei FA, van Rijn MJ, Schut AF, et al. Heritability of the function and structure of the arterial wall: findings of the Erasmus Rucphen Family (ERF) study. *Stroke; a journal of cerebral circulation*. Nov 2005;36(11):2351-2356.
67. Schoenmaker M, de Craen AJ, de Meijer PH, et al. Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. *European journal of human genetics : EJHG*. Jan 2006;14(1):79-84.

68. Tibshirani R. The lasso method for variable selection in the Cox model. *Statistics in medicine*. Feb 28 1997;16(4):385-395.
69. Tibshirani R. Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B Met*. 1996;58(1):267-288.
70. De Meyer T, Sinnaeve D, Van Gasse B, et al. Evaluation of standard and advanced preprocessing methods for the univariate analysis of blood serum 1H-NMR spectra. *Analytical and bioanalytical chemistry*. Oct 2010;398(4):1781-1790.
71. Smolinska A, Blanchet L, Buydens LM, Wijmenga SS. NMR and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review. *Analytica chimica acta*. Oct 31 2012;750:82-97.
72. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*. 2006;7:142.
73. Heijmans BT, Beekman M, Houwing-Duistermaat JJ, et al. Lipoprotein particle profiles mark familial and sporadic human longevity. *PLoS medicine*. Dec 2006;3(12):e495.
74. Schunkert H, König IR, Kathiresan S, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics*. Apr 2011;43(4):333-338.
75. International Consortium for Blood Pressure Genome-Wide Association S, Ehret GB, Munroe PB, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*. Oct 6 2011;478(7367):103-109.
76. Mozaffarian D. Fish and n-3 fatty acids for the prevention of fatal coronary heart disease and sudden cardiac death. *The American journal of clinical nutrition*. Jun 2008;87(6):1991S-1996S.
77. Harris WS, Mozaffarian D, Rimm E, et al. Omega-6 fatty acids and risk for cardiovascular disease: a science advisory from the American Heart Association Nutrition Subcommittee of the Council on Nutrition, Physical Activity, and Metabolism; Council on Cardiovascular Nursing; and Council on Epidemiology and Prevention. *Circulation*. Feb 17 2009;119(6):902-907.
78. De Caterina R. n-3 fatty acids in cardiovascular disease. *The New England journal of medicine*. Jun 23 2011;364(25):2439-2450.
79. Raatz SK, Bibus D, Thomas W, Kris-Etherton P. Total fat intake modifies plasma fatty acid composition in humans. *The Journal of nutrition*. Feb 2001;131(2):231-234.
80. Cho HP, Nakamura M, Clarke SD. Cloning, expression, and fatty acid regulation of the human delta-5 desaturase. *The Journal of biological chemistry*. Dec 24 1999;274(52):37335-37339.
81. Cho HP, Nakamura MT, Clarke SD. Cloning, expression, and nutritional regulation of the mammalian Delta-6 desaturase. *The Journal of biological chemistry*. Jan 1 1999;274(1):471-477.

A metabolomics profile is associated with the risk of incident coronary heart disease

Anika A.M. Vaarhorst^{1*}, Aswin Verhoeven^{2*}, Claudia M. Weller^{3*}, Stefan Böhringer⁴, Sibel Göröler², Axel Meissner², André M. Deelder², Peter Henneman³, Anton P.M. Gorgels⁵, Piet A. van den Brandt^{6,7}, Leo J. Schouten⁷, Marleen M. van Greevenbroek⁸, Audrey H.H. Merry⁶, W.M. Monique Verschuren⁹, Arn M.J.M. van den Maagdenberg^{3,10}, Ko Willems van Dijk^{3,11}, Aaron Isaacs¹², Dorret Boomsma¹³, Ben A. Oostra¹², Cornelia M. van Duijn¹², J. Wouter Jukema^{14,15,16}, Jolanda M.A. Boer⁹, Edith Feskens¹⁷, Bastiaan T. Heijmans¹, P. Eline Slagboom^{1,18}

*These authors contributed equally to this work

From the departments of **1)** Molecular Epidemiology, **2)** Center for Proteomics and Metabolomics, **3)** Human Genetics, **4)** Medical Statistics, **10)** Neurology, **11)** Endocrinology, **14)** Cardiology, Leiden University Medical Center, Leiden, the Netherlands From the departments of **5)** Cardiology, **8)** Internal Medicine (CARIM School for Cardiovascular diseases), Maastricht University Medical Centre, Maastricht, the Netherlands From the departments of **6)** Epidemiology (CAPHRI School for Public Health and Primary Care), **7)** Epidemiology (GROW School of Oncology and Developmental Biology), Maastricht University, Maastricht, the Netherlands; **9)** National Institute for Public Health and the Environment, Bilthoven, the Netherlands; **12)** Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands; **13)** Biological Psychology, VU University, Amsterdam, the Netherlands; **15)** The Durrer Center for Cardiogenetic Research, Amsterdam, the Netherlands; **16)** Interuniversity Cardiology Institute of the Netherlands (ICIN), Utrecht, the Netherlands; **17)** Division of Human Nutrition, Wageningen University and Research Center, Wageningen, the Netherlands; **18)** Netherlands Consortium for Healthy Ageing

Accepted for publication in American Heart Journal

Abstract

Background Metabolomics, defined as the comprehensive identification and quantification of low-molecular-weight metabolites to be found in a biological sample, has been put forward as a potential tool for classifying individuals according to their risk of coronary heart disease (CHD). Here, we investigated whether a single-point blood measurement of the metabolome is associated with and predictive for the risk of CHD.

Methods & Results We obtained proton nuclear magnetic resonance ($^1\text{H-NMR}$) spectra in 79 cases who developed CHD during follow-up (median 8.1 years) and in 565 randomly selected individuals. In these spectra 100 signals representing 36 metabolites were identified. Applying LASSO regression, we defined a weighted metabolite score consisting of 13 $^1\text{H-NMR}$ signals that optimally predicted CHD. This metabolite score, including signals representing a lipid fraction, glucose, valine, ornithine, glutamate, creatinine, glycoproteins, citrate and 1.5-anhydrosorbitol, was associated with the incidence of CHD independent of traditional risk factors (TRFs) (HR=1.50; 95%CI=1.12-2.01). Predictive performance of this metabolite score on its own was moderate (C-index=0.75; 95%CI=0.70-0.80) but after adding age and sex the C-index was only modestly lower than that of TRFs (C-index=0.81; 95%CI=0.77-0.85 and C-index=0.82; 95%CI=0.78-0.87, respectively). The metabolite score was also associated with prevalent CHD independent of TRFs (OR=1.59; 95%CI=1.19-2.13).

Conclusion A metabolite score derived from a single-point metabolome measurement is associated with CHD and metabolomics may be a promising tool for refining and improving the prediction of CHD.

Introduction

Over the last 50 years, risk factors that are robustly and independently associated with coronary heart disease (CHD), including lipid levels, blood pressure, lifestyle factors, family history, sex and age, were identified.^{1,2} Based on these traditional risk factors (TRFs), scores have been developed to predict CHD risk for an individual.^{1,2} The discriminatory capabilities for these scores, as assessed by the C-index, ranges from 0.71 to 0.84.¹

Metabolomics refers to the identification and quantification of low-molecular-weight metabolites in a biological sample.³ Recent technological developments made it possible to generate metabolomics profiles of blood samples consisting of 10s to 100s metabolites in a single measurement.⁴ These profiles are considered to be promising tools to efficiently capture the predictive information of TRFs and may potentially contribute to further improvement of primary CHD risk prediction.³

Several studies have attempted to use a metabolomics approach to diagnose prevalent CHD^{5,6} or to predict incident CHD events in individuals free from cardiovascular disease⁷⁻⁹ or diagnosed with CHD¹⁰⁻¹³ or diabetes.¹⁴ According to some studies, a metabolomics approach might improve CHD risk prediction.⁸⁻¹³ For example, in one study, ¹H-NMR spectroscopy improves the prediction of subclinical atherosclerosis in comparison to conventional lipid testing.⁸ Other studies found that a baseline metabolomics profile based on mass-spectroscopy was associated with incident cardiovascular events in patients diagnosed with CHD¹¹ or in patients with suspected CHD.^{10,12,13} However, none of these studies investigated if a metabolite profile based on low-molecular-weight molecules identified by proton nuclear magnetic resonance (¹H-NMR) spectroscopy could predict incident CHD, defined as an acute myocardial infarction (MI), unstable angina pectoris (UAP) or dead because of CHD, in individuals free from cardiovascular disease. Therefore, we studied the association of an ¹H-NMR based metabolite profile with incident CHD in a prospective case-cohort study. Subsequently, a second study was performed in an independent population to test if the selected metabolites were also relevant to classify prevalent CHD.

Material and Methods

Study populations

Primary study: We conducted a prospective case-cohort study within the Monitoring project on chronic disease risk factors (MORGEN-Project) 1993-1997,¹⁵ one of the two monitoring studies that were included in the Cardiovascular Registry Maastricht study.¹⁶ In total 6459 men and women, between 20-59 years old at the moment of inclusion, had given informed consent to retrieve information from the municipal registries and from the

general practitioner and specialist. The study complied with the Declaration of Helsinki and was approved by the Medical Ethics Committee of TNO (Dutch Organization for Applied Scientific Research). We excluded participants who were younger than 30 years at baseline (n=1301), who had had an acute MI, UAP, a coronary artery bypass graft or a percutaneous transluminal coronary angioplasty before baseline (n=69), or were lost to follow-up (n=15), resulting in an eligible cohort of 5074 participants.

Sub-cohort selection: From the eligible cohort, a sub-cohort of 738 participants was randomly drawn. This took place before cardiologic follow-up. EDTA plasma was unavailable for 92 participants and ¹H-NMR analysis failed in 19 participants. For 62 participants information on TRFs (*i.e.* total cholesterol [TC], high-density lipoprotein cholesterol [HDL-C], systolic blood pressure [SBP], current smoking, body mass index [BMI], current diabetes status and a parental history of MI) was incomplete, resulting in a sub-cohort of 565 participants.

Cardiologic follow-up: The cardiologic follow-up has been described in detail earlier¹⁶ and ended on 31 December 2003 with a median follow-up of 8.1 years (range=0.2-10.9 years). During follow-up 125 participants developed CHD (acute MI n=55, UAP n=51, dead due to CHD n=19, [ICD-9 410-414 and ICD-10 I20-I25]). For 31 patients EDTA plasma for ¹H-NMR analysis was unavailable and in one patient ¹H-NMR analysis failed. For 14 patients information on TRFs was incomplete, resulting in 79 patients.

Determination of TRFs: At baseline, participants filled in a questionnaire on medical history (including self-reported diabetes), parental history of MI (defined as no parents with MI, one parent with MI or both parents with MI), and lifestyle factors (including smoking). During a medical examination, information on SBP and BMI was collected and non-fasting EDTA blood samples were taken.^{17,18} The blood was centrifuged for 10 minutes at 1000 rpm at 4°C and EDTA plasma aliquots were stored at -80°C in tubes of 0.5 mL for future analysis, or at -20°C for cholesterol determinations. HDL-C and TC levels were determined in the plasma samples stored at -20°C using a CHOD-PAP method.¹⁹

Secondary study: The Erasmus Rucphen Family study is a population-based study in a genetically isolated community in the Southwest of the Netherlands and includes 3465 individuals, who are living descendants of 22 couples that had at least six children baptized in the community church between 1850 and 1900. Details are described elsewhere.²⁰ The study was approved by the Medical Ethical Committee of the Erasmus Medical Center in Rotterdam and all participants gave written informed consent.

Determination of TRFs: All participants filled in questionnaires about lifestyle, personal and family medical history. During personal interviews performed by study physicians, information on lifestyle factors, medication use, personal and family medical history was collected. Physical examinations were performed, including measurements of SBP and BMI.²¹

In addition, fasted blood samples were taken. An electrocardiogram (ECG) was performed and scored by an experienced cardiologist. Plasma concentrations of HDL-C and TC were determined according to standard procedures.²² Diabetes was defined as the use of blood glucose-lowering medication and/or fasting glucose levels of ≥ 7 mmol/L.

Diagnosis of CHD in the secondary study: Participants were classified as CHD cases if they indicated during the interview or in the questionnaire that they had experienced a MI or underwent a coronary revascularization procedure, they reported angina symptoms in the interview, and/or showed signs of MI on ECG.

From 2919 participants fasting serum samples were available. Good quality ^1H -NMR spectra from 2415 participants were obtained. For 2327 of these participants data on CHD diagnosis was available and 170 were classified as having CHD.

^1H -NMR metabolite profiling

The stored EDTA plasma and serum samples were thawed at 4°C and were mixed by inverting the tubes 10 times. Next, samples (300 μL) were mixed with 300 μL TSP buffer (see Supplement I, supplementary Material & methods) and transferred into 5mm NMR tubes and kept at 6°C while queued for acquisition. Two-dimensional J-resolved and Carr-Purcell-Meiboom-Gill spectra were acquired on a 600 MHz Bruker Avance II spectrometer (Bruker BioSpin, Karlsruhe, Germany), operating at a sample temperature of 310 K. For details on acquisition, processing, quality control, scaling and calibration of the ^1H -NMR spectra see Supplement I, supplementary material & methods.

Using the procedure described in the Supplement I, supplementary Material & methods, 100 signals were detected and quantified in the ^1H -NMR spectra of every individual in the primary study. For 76 signals metabolites were assigned (see Supplement I, table S1 for all ^1H -NMR signals and their assigned metabolites). These signals represented 36 different compounds (*i.e.* after subtracting the signals representing free EDTA). Signals representing calcium-EDTA and magnesium-EDTA complexes may give an indication of the levels of calcium and magnesium ions, respectively. Using the same procedure, 68 out of 100 signals detected in the primary study were detected in the secondary study. For 54 out of these 68 signals metabolites could be assigned, representing 28 different compounds.

Statistical analysis

In the primary study, Cox regression, adjusted for delayed entry, and according to the method of Prentice to adjust for the case-cohort design was performed to see whether

baseline characteristics were associated with incident CHD.²³ Age in years was used as the time-scale variable.

Prior to analysis, the ¹H-NMR signals were transformed to Z-scores. We selected a subset of the most informative signals for CHD prediction, using least absolute shrinkage and selection operator (LASSO) regression,²⁴ and performed 10-fold cross-validation to determine the tuning parameter.^{24,25} This set was further reduced to signals that could be detected in both studies. The linear predictor of the Cox-model was used as a weighted metabolite score (sum of regression coefficients multiplied by the corresponding covariate values).

Cox regression, according to the method of Prentice to adjust for the case-cohort design, was used to calculate whether this metabolite score was associated with incident CHD before and after adjusting for TRFs.²³

To investigate whether the metabolite score improved risk discrimination, Harrell's concordance index (C-index),²⁶ the net reclassification index (NRI) and the integrative discrimination index (IDI) were calculated.²⁷ For the NRI the following risk categories were applied: 0-<5%, 5-<10%, 10-<20%, ≥20%. Next, using ANOVA, we investigated to what extent TRFs can explain the variance in the metabolite score in the sub-cohort. Stata/SE version 11.2 was used to calculate the C-index. R-package PredictABEL version 1.2-1 was used to calculate the IDI. For all other analyses, R version 2.14.1 was used.

In the secondary study, raw ¹H-NMR signal data were adjusted for kinship by linear regression in GenABEL.²⁸ The residuals for all signals were transformed into Z-scores. Logistic regression was performed to assess the association of TRFs with prevalent CHD. To assess whether the metabolite signals selected by LASSO regression in the primary study were also relevant for the identification of prevalent CHD cases, we put these signals into one logistic regression model to determine their individual regression coefficients. Next, we used logistic regression analysis to test the association of this metabolite score, thus with weights based on the secondary study, with prevalent CHD. Finally, we tested to what extent TRFs can explain the variance in the metabolite score using ANOVA. For these analyses, PASW statistics version 18 (SPSS-IBM, New York, US) was used.

Results

Primary study: The TRFs were associated with incident CHD. When all TRFs were entered into one Cox proportional hazards model with age in years as the time-scale variable, HDL-C, SBP, sex and current smoking remained independently associated with incident CHD (table 1). See supplement I, table S2 for the baseline characteristics before excluding individuals

with missing data. For the association of the individual ^1H -NMR signals with incident CHD see supplement I, table S1.

Table 1 | TRFs and their association with incident CHD in the primary study.

	Cases (n=79)	Sub-cohort (n=565)*	HR (95%CI) [†]	HR (95%CI) [‡]
Age (years)	51.9±6.1	44.8±8.3	-	-
Total cholesterol (mmol/L)	5.9±1.0	5.3±1.0	1.30 (1.03-1.64)	1.21 (0.93-1.58)
HDL cholesterol (mmol/L)	1.1±0.3	1.3±0.4	0.12 (0.05-0.30)	0.34 (0.12-0.94)
SBP (mmHg)	134.5±16.8	121.5±15.1	1.04 (1.02-1.05)	1.03 (1.01-1.05)
BMI (kg/m ²)	27.6±4.6	25.4±3.8	1.10 (1.03-1.17)	1.03 (0.95-1.11)
Men	79.6% (63)	44.6% (252)	4.95 (2.72-9.02)	3.30 (1.65-6.57)
Current smoking	51.9% (41)	38.8% (219)	2.16 (1.29-3.62)	1.84 (1.02-3.32)
Diabetes	3.8% (3)	0.5% (3)	3.99 (0.70-22.84)	3.34 (0.41-27.43)
Parental history of MI	50.6% (40)	40.2% (227)	1.32 (0.88-1.98)	1.24 (0.78-1.98)

Data are expressed as mean±SD or %(n). *HR*, hazard ratio; *HDL*, high-density lipoprotein; *MI*, myocardial infarction; *SBP*, systolic blood pressure; *BMI*, body mass index

*Including 10 cases

[†]Univariate HR were calculated per unit increase for age, total cholesterol, HDL cholesterol, SBP, BMI, and for the categorical traits. Age in years was used as the time-scale variable.

[‡]All variables were added into one multivariable Cox proportional hazards model, except for age, which was used as the time-scale variable.

We determined 16 ^1H -NMR signals as the best prediction subset using LASSO regression, of which, 13 were available in the secondary study (see figure 1). These 16 ^1H -NMR signals represent creatinine, serine, glucose, 1,5-anhydrosorbitol, trimethylamine N-oxide (TMAO), ornithine, citrate, glutamate, glycoproteins, an unsaturated lipid structure, valine and five non-annotated signals located at 3.924, 3.145, 2.412, 1.391 and 0.988 ppm. From the 13 signals present in both the primary and secondary study a weighted metabolite score was constructed using the corresponding coefficients (figure 1). This metabolite score was normally distributed in cases and sub-cohort members (supplement I, figure S1), and associated with incident CHD (hazard ratio per standard deviation [HR/SD]=1.91, 95% confidence interval [95%CI]=1.50-2.44). After adjusting for TRFs, this metabolite score remained associated with incident CHD (HR/SD=1.50, 95%CI=1.12-2.01). For the results of the metabolite score based on the 16 signals see supplement I, table S3.

The metabolite score had a C-index of 0.75, (95%CI=0.70-0.80). Adding age and sex to the metabolite score, resulted in a C-index of 0.81 (95%CI=0.77-0.85), which is similar to a C-index when only TRFs are included (C-index=0.82, 95%CI=0.78-0.87, $p=0.327$). When the metabolite score was added to a model containing all TRFs, the C-index increased

from 0.82 to 0.84, which was non-significant ($p=0.107$). Both the improvement in the NRI ($\text{NRI}_{\text{total}}=0.038$; $p=0.209$) and the IDI (0.012; $p=0.091$) were non-significant (supplement I, table S4 and S5). Inspecting C-indices for individual TRFs and evaluating improvement of adding the metabolite score indicated that the metabolite score improved the C-indices of all individual TRFs (supplement I, table S6a/b).

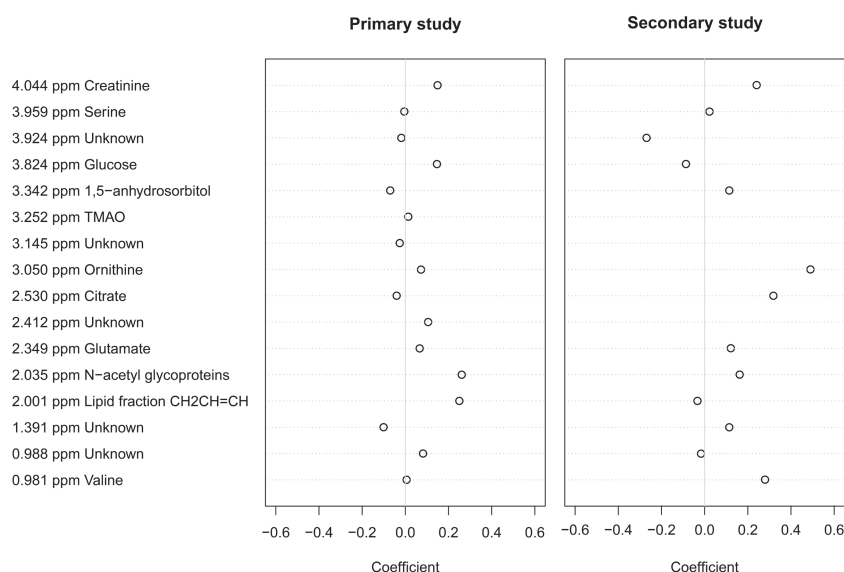


Figure 1 | The subset of signals selected using 10-fold cross-validated LASSO regression and their coefficients in the primary study (left panel) and the secondary study (right panel).

We tested to what extent TRFs explain the variance in the metabolite score in the sub-cohort of the primary study. HDL-C, sex, BMI, TC, SBP, age and diabetes explained respective 16.4%, 11.7%, 10.0%, 7.9%, 5.6%, 4.1% and 2.3% of the variance in the metabolite score. Current smoking and parental history of MI all explained less than 1% of the variance in the metabolite score (figure 2). When all TRFs were combined, 32.6% of the variance in the metabolite score was explained.

Secondary study: To test if the 13 metabolite signals selected in the primary study were relevant for the identification of prevalent CHD cases, we investigated 170 CHD cases and 2157 controls for which equivalent metabolomics profiles were obtained. In this non-prospective cohort, combining all TRFs in one logistic regression model resulted in only age, sex, and parental history of MI to be independently associated with prevalent CHD (table 2). See supplement I, table S7 for the baseline characteristics before excluding individuals with missing data.

Table 2 | TRFs and their association with prevalent CHD in the secondary study.

	Cases (n=170)	Controls (n=2157)	OR (95%CI)*	OR (95%CI)†
Age (years)	60.9±11.7	47.7±14.0	1.08 (1.07-1.10)	1.08 (1.05-1.10)
Total cholesterol (mmol/L)	5.0±1.1	5.6±1.1	0.57 (0.49-0.67)	0.47 (0.37-0.59)
HDL cholesterol (mmol/L)	1.2±0.3	1.3±0.4	0.30 (0.18-0.49)	1.25 (0.60-2.61)
SBP (mmHg)	148.2±22.9	139.1±19.7	1.02 (1.02-1.03)	1.00 (0.99-1.01)
BMI (kg/m ²)	28.1±4.4	26.8±4.6	1.06 (1.03-1.09)	1.01 (0.96-1.07)
Men	61.2% (104)	42.5% (917)	2.40 (1.79-3.22)	1.93 (1.18-3.14)
Current smoking	36.5% (62)	39.1% (844)	0.90 (0.66-1.24)	1.18 (0.75-1.88)
Diabetes	12.9% (22)	4.3% (93)	3.44 (2.15-5.49)	1.25 (0.56-2.79)
Parental history of MI	31.2% (53)	19.9% (430)	2.03 (1.53-2.71)	1.60 (1.16-2.21)

Data are expressed as mean±SD or %(n). OR, odds ratio; HDL, high-density lipoprotein; MI, myocardial infarction; SBP, systolic blood pressure

*Univariate OR were calculated per unit increase for age, total cholesterol, HDL cholesterol, SBP, BMI, and for the categorical traits.

†All the variables were added into one multivariable logistic regression model.

The metabolite score, based on the 13 best predicting signals in the case-cohort study, but with weights based on the secondary study (figure 2), was associated with prevalent CHD before (OR=2.72, $p<0.001$) and after adjusting for TRFs (odds ratio [OR]=1.59, $p=0.002$). After excluding statin users (n=299), similar results were obtained (supplement I, table S8).

The proportion of variance in the metabolite score explained by age, SBP and diabetes was higher for the secondary study than for primary study, whereas the variables TC, HDL-C and BMI explained a lower proportion of the variance (figure 2). The proportion of variance explained by sex, current smoking and parental history of MI was comparable for both studies. With all TRFs combined, 32.1% of the variance of the metabolite score could be explained, which is comparable to that of the primary study.

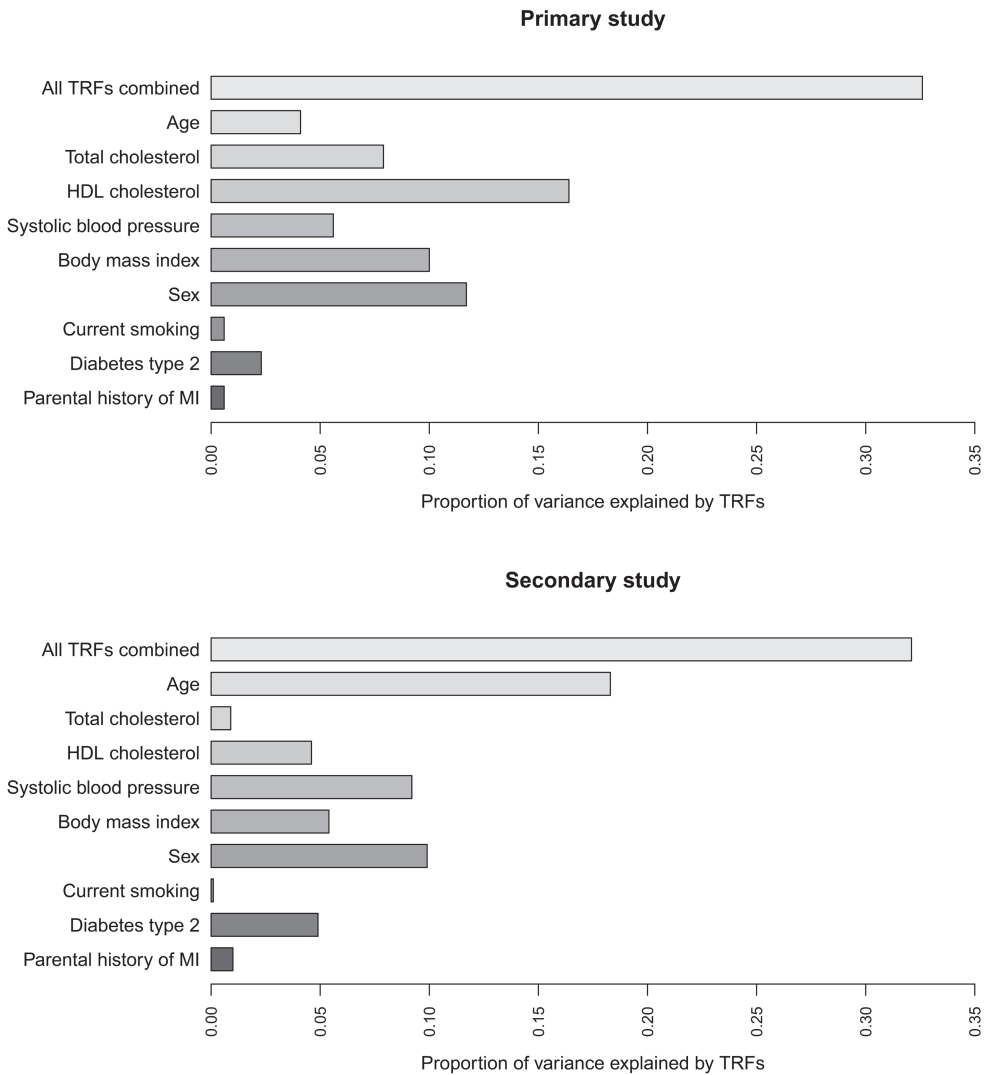


Figure 2 | The proportion of variance in the metabolite score explained by TRFs in the primary study (upper panel) and the secondary study (lower panel).

Discussion

A metabolite score, based on ¹H-NMR spectroscopy, is significantly associated with incident CHD independent of TRFs. When combined with age and sex, this score was as predictive for incident CHD as all TRFs combined. A score based on the same ¹H-NMR signals was also associated with prevalent CHD, independent of TRFs.

The observation that the metabolite score could not improve CHD risk prediction beyond TRFs in individuals free from CHD is in line with a previous study published by El Harchaoui *et al.*⁷ However, Würtz *et al.* found that metabolites measured by ¹H-NMR spectroscopy improved risk stratification for subclinical atherosclerosis in comparison to conventional lipids.⁸ In this study both ¹H-NMR determined lipoproteins and low-molecular weight metabolites were included.⁸ We only included information on low-molecular weight metabolites, whereas in the study by El Harchaoui *et al.* only information on lipoproteins was included.⁷ Perhaps the combination of both lipoproteins and low-molecular-weight metabolites results in the optimal prediction of CHD.

The metabolite score represents the metabolites valine, ornithine, glucose, 1,5-anhydrosorbitol, creatinine, an unsaturated lipid structure, glutamate and, glycoproteins, citrate, and TMAO, of which TMAO was not available in the secondary study. Most of these metabolites have been associated with CHD or CHD risk factors before.^{11,13,29-31} Valine has been associated with metabolic risk factors,³⁰ insulin resistance,³² incident type 2 diabetes,³¹ and future cardiovascular events.¹¹ Ornithine, is produced by splitting of urea from arginine, resulting in a lower bioavailability of arginine. Arginine is necessary to produce nitric oxide, which is essential for a normal endothelial function.²⁹ This pathway has been linked to CHD¹⁰ and CHD mortality.²⁹ The presence of glucose and 1,5-anhydrosorbitol, a short-term marker for glycemic control,³³ could indicate that our metabolite score marks individuals at higher risk of developing diabetes or insulin resistance and thereby CHD. Low creatinine levels are a marker for high HDL-C and low LDL cholesterol levels.³⁴ Thus the presence of creatinine and an unsaturated lipid structure could indicate that our metabolite risk score is a marker for an unfavorable lipid profile. This is confirmed by the proportion of variance explained by HDL-C and TC levels, 7.9% and 16.4%. In the secondary study these explained variances are only 0.9% and 4.6%, but this discrepancy might be caused by statin treatment, resulting in lower cholesterol levels for the cases compared to controls. A secondary explanation for the incorporation of creatinine in the metabolite score is that elevated creatinine levels may indicate kidney dysfunction, which is associated with cardiovascular disease.³⁵ Increased TMAO levels have been associated with cardiovascular risk before.^{13,36} Thus it seems that the LASSO procedure selected relevant metabolites that have been associated with CHD and CHD risk factors before.

Several issues have to be resolved before it can be concluded if a metabolomics approach is useful for CHD risk prediction. First, the known, quantifiable serum metabolome consists of 4229 metabolites,³⁷ of which only 36 (0.9%) were included in this study. Other studies that use ¹H-NMR spectroscopy also incorporated lipoproteins.^{11,12} Therefore we hope that we can achieve better in follow-up studies when incorporating H-NMR determined lipoproteins

in addition to low-molecular-weight metabolites in our analysis. Moreover, other metabolomics platforms should also be included.³⁷ Second, the 16 signals provided by our study should be measured in large prospective cohorts for replication and to determine universally applicable weights. The current study is too small for that purpose. Third, we had non-fasting samples in the primary study and fasted samples in the secondary study. However, we still found that the ¹H-NMR signals selected in the primary non-fasted study were also associated with prevalent CHD in the fasted secondary study. This indicates that we have selected ¹H-NMR signals that are robust whether fasted or non-fasted samples are used. Fourth, constructing robust prediction models constitutes a statistical challenge, especially in a high-dimensional setting. In our case, model selection by LASSO regression resulted in predictor selection that eliminated high correlations among predictors. This can lead to reduced transferability of prediction models as correlation structures of predictors can vary between studies. A wide variety of penalized regression models are available (*e.g.* elastic net, ridge regression) and further research is needed to select the appropriate methods for the application at hand.

Conclusion

A metabolite score derived from a single point metabolome measurement is associated with the risk of CHD independent of TRFs, but thus not improve risk prediction beyond TRFs. On the other hand, LASSO regression resulted in the selection of relevant metabolites, suggesting that more comprehensive metabolomics methods may be promising tools to further improve upon CHD disease understanding and risk stratification.

Acknowledgements

We thank the MORGEN-project steering committee consisting of (not mentioning co-authors of this article) HB Bueno de Mesquita, HA Smit and JC Seidell. Furthermore we thank the epidemiologists and field-workers of the Municipal Health Services in Maastricht for their contribution to baseline data collection and those involved in logistic management (A Jansen and J Steenbrink) and data management (A Blokstra, A van Kessel, P Steinberger, E den Hoedt, I Toxopeus, J van der Laan). We thank A Hilton, V Visser, P Erkens, S Philippens, J Bremen and B Bleijlevens for assistance in clinical data collection. Statistics Netherlands is acknowledged for providing data on causes of death. We thank Joris Deelen, Dennis Kremer, Ruud van der Breggen, Yingchang Lu, Mariken Tijhuis and Cecile Povel for assistance with practical work.

We thank the participants from the Genetic Research in Isolated Populations, Erasmus Rucphen Family studies, as well as the general practitioner and other clinicians who made this work possible.

Funding sources

This study was funded by a grant from the Netherlands Heart Foundation (2006B195), the Netherlands Consortium for Healthy Ageing (05060810) and by BBMRI-NL, a Research Infrastructure financed by the Dutch government (NWO 184.021.007).

The MORGEN-Project was financially supported by the Ministry of Public Health, Welfare and Sports of the Netherlands.

The ERF study was funded by the Centre for Medical Systems Biology in the framework of the Netherlands Genomics Initiatives and by the European Network of Genomic And Genetic Epidemiology consortium.

The funding sources had no involvement in the design and conduct of the study; or with the collection, management, analysis or interpretation of the data and reporting or the decision to submit the manuscript for publication.

References

- Conroy R. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *European heart journal*. 2003;24(11):987-1003.
- Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation*. May 12 1998;97(18):1837-1847.
- Ellis DI, Dunn WB, Griffin JL, Allwood JW, Goodacre R. Metabolic fingerprinting as a diagnostic tool. *Pharmacogenomics*. Sep 2007;8(9):1243-1266.
- Beckonert O, Keun HC, Ebbels TM, et al. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature protocols*. 2007;2(11):2692-2703.
- Kirschenlohr HL, Griffin JL, Clarke SC, et al. Proton NMR analysis of plasma is a weak predictor of coronary artery disease. *Nature medicine*. Jun 2006;12(6):705-710.
- Brindle JT, Antti H, Holmes E, et al. Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using ¹H-NMR-based metabolomics. *Nature medicine*. Dec 2002;8(12):1439-1444.
- El Harchaoui K, van der Steeg WA, Stroes ES, et al. Value of low-density lipoprotein particle number and size as predictors of coronary artery disease in apparently healthy men and women: the EPIC-Norfolk Prospective Population Study. *Journal of the American College of Cardiology*. Feb 6 2007;49(5):547-553.
- Wurtz P, Raiko JR, Magnussen CG, et al. High-throughput quantification of circulating metabolites improves prediction of subclinical atherosclerosis. *European heart journal*. Sep 2012;33(18):2307-2316.
- Magnusson M, Lewis GD, Ericson U, et al. A diabetes-predictive amino acid score and future cardiovascular disease. *European heart journal*. Jul 2013;34(26):1982-1989.
- Wang Z, Tang WH, Cho L, Brennan DM, Hazen SL. Targeted metabolomic evaluation of arginine methylation and cardiovascular risks: potential mechanisms beyond nitric oxide synthase inhibition. *Arteriosclerosis, thrombosis, and vascular biology*. Sep 2009;29(9):1383-1391.
- Shah SH, Bain JR, Muehlbauer MJ, et al. Association of a peripheral blood metabolic profile with coronary artery disease and risk of subsequent cardiovascular events. *Circulation. Cardiovascular genetics*. Apr 2010;3(2):207-214.
- Shah SH, Sun JL, Stevens RD, et al. Baseline metabolomic profiles predict cardiovascular events in patients at risk for coronary artery disease. *American heart journal*. May 2012;163(5):844-850 e841.
- Wang Z, Klipfell E, Bennett BJ, et al. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature*. Apr 7 2011;472(7341):57-63.
- Roussel R, Mentre F, Bouchemal N, et al. NMR-based prediction of cardiovascular risk in diabetes. *Nature medicine*. Apr 2007;13(4):399-400.
- Picavet HS, Schouten JS. Physical load in daily life and low back problems in the general population-The MORGEN study. *Preventive medicine*. Nov 2000;31(5):506-512.
- Merry AH, Boer JM, Schouten LJ, et al. Validity of coronary heart diseases and heart failure based on hospital discharge and mortality data in the Netherlands using the cardiovascular registry Maastricht cohort study. *European journal of epidemiology*. 2009;24(5):237-247.
- Lean MEJ, Han TS, Seidell JC. Impairment of health and quality of life in people with large waist circumference. *The Lancet*. 1998;351(9106):853-856.
- Van Leer EM, Seidell JC, Kromhout D. Dietary calcium, potassium, magnesium and blood pressure in the Netherlands. *Int J Epidemiol*. Dec 1995;24(6):1117-1123.
- Kattermann R, Jaworek D, Moller G, et al. Multicentre study of a new enzymatic method of cholesterol determination. *Journal of clinical chemistry and clinical biochemistry. Zeitschrift fur klinische Chemie und klinische Biochemie*. Mar 1984;22(3):245-251.
- Aulchenko YS, Heutink P, Mackay I, et al. Linkage disequilibrium in young genetically isolated Dutch population. *European journal of human genetics : EJHG*. Jul 2004;12(7):527-534.
- Sayed-Tabatabaei FA, van Rijn MJ, Schut AF, et al. Heritability of the function and structure of the arterial wall: findings of the Erasmus Rucphen Family (ERF) study. *Stroke; a journal of cerebral circulation*. Nov 2005;36(11):2351-2356.
- van Gent CM, van der Voort HA, de Bruyn AM, Klein F. Cholesterol determinations. A comparative study of methods with special reference to enzymatic procedures. *Clinica chimica acta;*

- international journal of clinical chemistry*. Mar 1 1977;75(2):243-251.
23. Prentice RL. A Case-Cohort Design for Epidemiologic Cohort Studies and Disease Prevention Trials. *Biometrika*. Apr 1986;73(1):1-11.
24. Goeman JJ. L1 penalized estimation in the Cox proportional hazards model. *Biometrical journal. Biometrische Zeitschrift*. Feb 2010;52(1):70-84.
25. Tibshirani R. The lasso method for variable selection in the Cox model. *Statistics in medicine*. Feb 28 1997;16(4):385-395.
26. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA : the journal of the American Medical Association*. May 14 1982;247(18):2543-2546.
27. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine*. Jan 30 2008;27(2):157-172; discussion 207-112.
28. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. *Bioinformatics*. May 15 2007;23(10):1294-1296.
29. Sourij H, Meinitzer A, Pilz S, et al. Arginine bioavailability ratios are associated with cardiovascular mortality in patients referred to coronary angiography. *Atherosclerosis*. Sep 2011;218(1):220-225.
30. Cheng S, Rhee EP, Larson MG, et al. Metabolite profiling identifies pathways associated with metabolic risk in humans. *Circulation*. May 8 2012;125(18):2222-2231.
31. Wang TJ, Larson MG, Vasan RS, et al. Metabolite profiles and the risk of developing diabetes. *Nature medicine*. Apr 2011;17(4):448-453.
32. Wurtz P, Soinenen P, Kangas AJ, et al. Branched-chain and aromatic amino acids are predictors of insulin resistance in young adults. *Diabetes care*. Mar 2013;36(3):648-655.
33. Dungan KM. 1,5-anhydroglucitol (GlycoMark) as a marker of short-term glycemic control and glycemic excursions. *Expert review of molecular diagnostics*. Jan 2008;8(1):9-19.
34. Bernini P, Bertini I, Luchinat C, Tenori L, Tognacini A. The cardiovascular risk of healthy individuals studied by NMR metabonomics of plasma samples. *Journal of proteome research*. Nov 4 2011;10(11):4983-4992.
35. Go AS, Chertow GM, Fan D, McCulloch CE, Hsu CY. Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization. *The New England journal of medicine*. Sep 23 2004;351(13):1296-1305.
36. Tang WH, Wang Z, Levison BS, et al. Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *The New England journal of medicine*. Apr 25 2013;368(17):1575-1584.
37. Psychogios N, Hau DD, Peng J, et al. The human serum metabolome. *PLoS one*. 2011;6(2):e16957.

Comparing a targeted and an untargeted approach for extracting information from a ^1H -NMR spectra

Anika A.M. Vaarhorst^{1*}, Aswin Verhoeven^{2*}, Stefan Böhringer³, Sibel Görler², Axel Meissner², André M. Deelder², Anton P.M. Gorgels⁴, Piet A. van den Brandt^{5,6}, Leo J. Schouten⁶, Marleen M. van Greevenbroek⁷, Audrey H.H. Merry⁵, W.M. Monique Verschuren⁸, Jolanda M.A. Boer⁸, Edith Feskens⁹, Bastiaan T. Heijmans¹, P. Eline Slagboom^{1,10}

From the departments of **1)** Molecular Epidemiology, **2)** Center for Proteomics and Metabolomics, **3)** Medical Statistics, Leiden University Medical Center, Leiden, the Netherlands; From the departments of **4)** Cardiology, **7)** Internal Medicine (CARIM School for Cardiovascular diseases), Maastricht University Medical Centre, Maastricht, the Netherlands; From the departments of **5)** Epidemiology (CAPHRI School for Public Health and Primary Care), **6)** Epidemiology (GROW School of Oncology and Developmental Biology), Maastricht University, Maastricht, the Netherlands; **8)** National Institute for Public Health and the Environment, Bilthoven, the Netherlands; **9)** Division of Human Nutrition, Wageningen University and Research Center, Wageningen, the Netherlands; **10)** Netherlands Consortium for Healthy Ageing

Abstract

Two commonly used methods for extracting information from proton nuclear magnetic resonance ($^1\text{H-NMR}$) spectra are a targeted approach, which involves determining the surface area of all resolvable signals in the $^1\text{H-NMR}$ spectrum, and an untargeted approach, by portioning the $^1\text{H-NMR}$ spectrum into discrete sections. Here we compare both approaches for predicting coronary heart disease (CHD). Using $^1\text{H-NMR}$ spectroscopy, two-dimensional J-resolved (JRES) projection spectra were obtained in plasma samples from 79 incident CHD cases (median follow-up 8.1 years) and 565 randomly selected individuals from the same cohort (N=5074). Prior to analysis the data was autoscaled and pareto scaled, respectively. For the targeted approach, LASSO regression on autoscaled data resulted in the selection of 11 signals from a total of 96. A score based on these 11 signals was associated with incident CHD independent of traditional CHD risk factors (TRFs) (HR = 1.46; 95%CI = 1.09-1.98). Pareto scaled data resulted in the selection of 6 signals. A score based on these signals was not associated with CHD independent of TRFs (HR = 1.16; 95%CI = 0.83-1.62). For the untargeted approach, LASSO regression on autoscaled data resulted in the selection of 21 bins from a total of 400. Eight bins were located in a part of the $^1\text{H-NMR}$ spectra where no peaks were observed, which could indicate overfitting. A score based on these bins was associated with incident CHD independent of TRFs (HR = 2.25; 95%CI = 1.54-3.29). After pareto scaling, 6 bins were selected. None of the bins were located in the part of the $^1\text{H-NMR}$ spectra where no peaks were observed. A score based on these 6 bins was not associated with incident CHD independent of TRFs (HR = 1.17; 95%CI = 0.81-1.69). We conclude that the untargeted approach might be more prone to overfitting compared to the targeted approach. However to test this presumption, it is necessary to validate the metabolite scores in independent prospective cohorts.

Introduction

It is expected that acquiring information about metabolite concentrations in body fluids is useful in diagnosing, understanding and predicting disease.¹ Acquiring information about metabolites in a systematic and non-biased way is the main concern of the field of metabolomics. Proton nuclear magnetic resonance ($^1\text{H-NMR}$) spectroscopy and mass spectrometry (MS) are two commonly used methods to gain information about a person's metabolites at a large scale.² However, for large population based studies, $^1\text{H-NMR}$ spectroscopy might be preferred due to faster processing times and being more robust compared to MS based methods.²

Two distinct approaches can be used to extract information from the $^1\text{H-NMR}$ spectrum: the targeted approach, by determining the surface area of all resolvable signals in the $^1\text{H-NMR}$ spectrum³ and the untargeted approach, which involves partitioning the $^1\text{H-NMR}$ spectrum into discrete sections, and determine the sum of the spectral intensity in each section ("equidistant binning").⁴ Conceptually, both techniques have their advantages and disadvantages. For example, the targeted approach treats each signal (and therefore, in principle, each metabolite) separately, as long as they can be properly deconvoluted.⁵ Any underlying uninformative background signal, can be included in the deconvolution process, removing its intensity from the final result.⁵ However, not every metabolite has signals that can easily be separated from the other signals, or reach the detection level in every individual. Therefore, the targeted approach may not only remove unwanted background signal, but also real information. Another disadvantage is that signal selection is partly a manual process which is time-consuming and prone to human error and bias.^{6,7} In contrast, the untargeted or binning approach includes all the spectral intensities in the analysis, also of as yet unidentified signals, which may occur as a result of low concentration or spectral crowding. The obvious disadvantage of this approach is that each bin can potentially contain overlapping signals or that signals are obscured by background variation in the same bin.⁵ This can complicate the interpretation of the result, or could mask the effect of a low intensity metabolite that happen to be situated in the same bin as a very high intensity metabolite.⁵ Also, variations of the baseline intensity that is uninformative will be included in the bins.⁵ This could potentially result in overfitting of the model. Overfitting occurs when variables show a difference between cases and non-cases, which are actually based on chance and are therefore not reproducible in independent cohorts.⁸ This problem increases with the inclusion of too many noise variables.⁸

When applying multivariate techniques to find $^1\text{H-NMR}$ signals or bins associated with incident CHD, $^1\text{H-NMR}$ signals or bins with high intensities can obscure the $^1\text{H-NMR}$ signals or bins with low intensities.⁹ Two commonly used methods to overcome this problem are

autoscaling and pareto scaling.¹⁰ With autoscaling or standardizing, the standard deviation is used as the scaling factor, and as a result differences in the variation amplitude of the peaks or bins are eliminated and can no longer influence the variable selection.^{9,10} With pareto scaling the square root of the standard deviation is used as the scaling factor.^{9,10} With this method, the influence of the peaks or bins with high intensities and strong variation is reduced but not completely eliminated.^{9,10}

In order to decide if the targeted approach is more suitable than the untargeted approach for predicting coronary heart disease (CHD), both approaches in combination with both scaling methods were applied to ¹H-NMR spectra obtained from EDTA plasma samples from study persons who participated in a prospective cohort study. For efficiency, a case-cohort design was used.¹¹ Data from both approaches were used to generate models to predict CHD using LASSO regression.^{12,13}

Material and Methods

study population

We conducted a prospective case-cohort study within the Monitoring project on chronic disease risk factors (MORGEN-Project) 1993-1997,¹⁴ one of the two monitoring studies that were included in the Cardiovascular Registry Maastricht study.¹⁵ For detailed information of the study population, including information on inclusion and exclusion criteria and obtaining information on traditional risk factors (TRFs, *i.e.* total cholesterol [TC], high-density lipoprotein cholesterol [HDL-C], systolic blood pressure [SBP], current smoking, body mass index [BMI], current diabetes status and a parental history of MI) see chapter 2 of this thesis.

Sub-cohort selection: As shown in figure 1 the cohort consisted of 5074 eligible cohort members. From this eligible cohort, a sub-cohort of 738 participants was randomly drawn. This took place before cardiologic follow-up. EDTA plasma was unavailable for 92 participants and ¹H-NMR analysis failed in 19 participants. For 62 participants information on TRFs was incomplete, resulting in a sub-cohort of 565 participants.

Cardiologic follow-up: The cardiologic follow-up has been described in detail earlier¹⁵ and ended on 31 December 2003 with a median follow-up of 8.1 years (range=0.2-10.9 years). During follow-up 125 participants developed CHD (acute MI n=55, UAP n=51, dead due to CHD n=19, [ICD-9 410-414 and ICD-10 I20-I25]). For 31 patients EDTA plasma for ¹H-NMR analysis was unavailable and in one patient ¹H-NMR analysis failed. For 14 patients information on TRFs was incomplete, resulting in 79 patients.

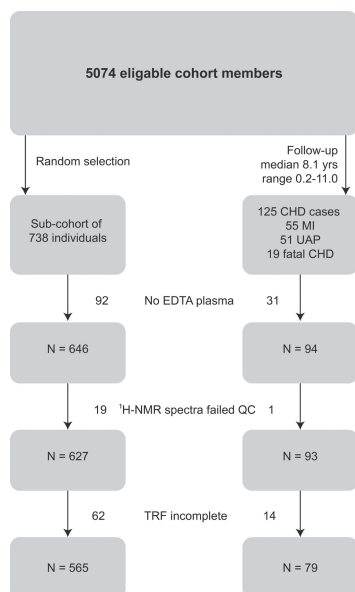


Fig 1 | Schematic representation of the case-cohort design and the inclusion criteria for the study participants

$^1\text{H-NMR}$ metabolite profiling

For detailed information on obtaining the $^1\text{H-NMR}$ spectra in the EDTA plasma samples see Material & methods from chapter 2 of this thesis.

Targeted approach - Using the procedure described in chapter 2 of this thesis, 100 signals were detected and quantified in the $^1\text{H-NMR}$ spectra of every individual. For 76 signals metabolites were assigned (see Supplement I, table S1 for all $^1\text{H-NMR}$ signals and their assigned metabolites). These signals represented 36 different compounds (*i.e.* after subtracting the two signals representing free EDTA). Signals representing calcium-EDTA and magnesium-EDTA complexes were not removed from analysis since they may give an indication of the levels of calcium and magnesium ions, respectively. In total we had 98 signals available for analysis.

The untargeted approach - For the untargeted approach, the projected JRES $^1\text{H-NMR}$ spectra were partitioned into discrete sections of 0.02 ppm, using the sum of the spectral intensity in each section ("equidistant binning").⁴

Data pretreatment

Principal component analysis (PCA) showed that three individuals had abnormal values for the signal located at 1.174 ppm representing ethanol (Supplement II, figure S1). This signal and another signal located at 3.649 ppm also representing ethanol were left out from

further analysis in all participants. We also found six individuals with high values for signals representing glucose (Supplement II, figure S1). Since these participants also had high glucose levels when measured using a conventional technique, both the participants and glucose signals were not removed from further analysis. Thus in total 96 signals representing 35 different metabolites were included in the analysis.

For the untargeted approach, after removing the water signal (from bin 4.41 to 5.09 ppm) and the two EDTA signals (from bin 3.15 to 3.25 ppm and from bin 3.53 to 3.67 ppm), the ^1H -NMR spectra were divided into 401 bins of 0.02 ppm. The PCA scores plot showed that the bin located at 1.17 ppm, containing an ethanol peak, was abnormal in the same three individuals (Supplement II, figure S1). This bin was removed from further analysis. The bin containing the other ethanol peak at 3.66 ppm, was left in because it was lower and located in a more crowded region of the spectrum. In total there were 400 bins available for further analysis.

Next the data were autoscaled and pareto scaled. Both scaling methods were combined with mean centering. Thus we now have four different datasets with information on the ^1H -NMR spectra. One for the targeted approach combined with autoscaling, one for the targeted approach combined with pareto scaling, one for the untargeted approach combined with autoscaling and one for the untargeted approach combined with pareto scaling. The effects of autoscaling and pareto scaling are visualized in figure 2.

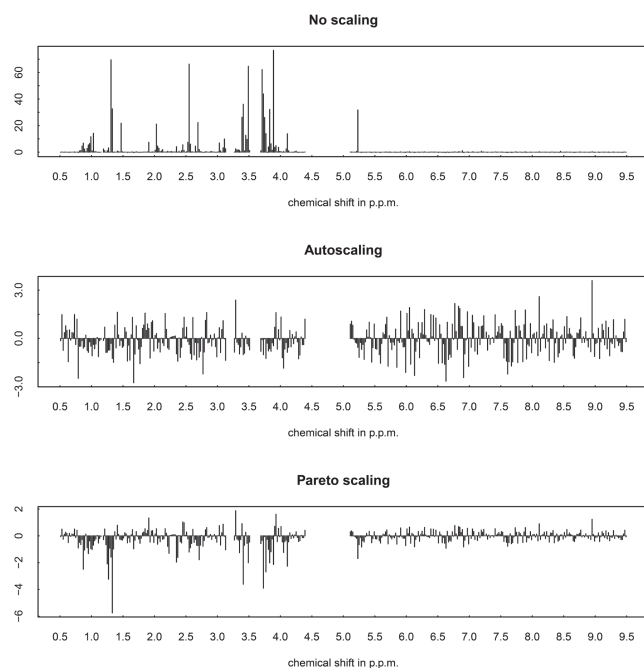


Fig 2 | The upper panel shows the ^1H -NMR spectra of a 36-year old woman after partitioning the ^1H -NMR spectra into discrete sections of 0.02 ppm. The resulting bins are visualized by vertical bars. The height of the bars corresponds to the intensities of the bins before any scaling method was applied. The middle panel shows the effects of autoscaling combined with mean centering on the intensities of the bins. The lower panel shows the effects of pareto scaling combined with mean centering on the intensities of the bins.

Statistical analysis

The first step in the data analysis was to select a subset of the most informative ^1H -NMR signals or bins for CHD risk prediction, both for the targeted and the untargeted approach. To perform this first step, we used least absolute shrinkage and selection operator (LASSO) regression.^{12,13} Twenty-fold cross-validation was used to determine tuning parameters.^{12,13} Next, to construct weighted metabolite scores based on the results from LASSO regression, the selected metabolites by LASSO regression were multiplied by their corresponding coefficients and the resulting values were summed in each individual. Cox regression according to the method of Prentice to adjust for the case-cohort design was used to see if these metabolite score was associated with incident CHD before and after adjusting for TRFs.¹¹ The hazard ratios were expressed per standard deviation increase in the metabolite scores. Age in years was used as the timescale variable.

The predictive value of both the targeted and the untargeted approach was evaluated using Harrell's concordance index (C-index)¹⁶ using Stata/SE version 11.2. For all other statistical calculations, R version 2.14.1 was used. For LASSO regression R-package glmnet (version 1.9-5) was used that was modified to perform an analysis adapted to the case-cohort design.¹⁷

Results

Targeted approach

Using LASSO regression on the autoscaled and mean-centered data, 11 signals were selected as the best prediction set (Fig. 3). These included 7 signals representing 5 individual compounds (glutamate, ornithine, 1,5-anhydrosorbitol, glucose and creatinine), as well as signals for unsaturated lipids ($\text{CH}_2\text{CH}=\text{CH}$) and glycoproteins. For 4 signals, no metabolite

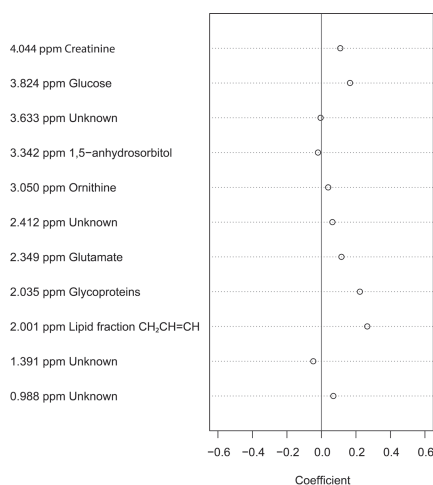


Fig 3 | The 11 signals and their corresponding coefficients selected by LASSO regression using the targeted approach in combination with autoscaling.

could be assigned. As shown in table 1, the metabolite score based on these signals and their corresponding coefficients (range = -2.22-5.53) was associated with incident CHD before (HR = 1.86; 95% CI = 1.46-2.37) and after adjusting for TRFs (HR = 1.46; 95% CI = 1.09-1.98). This metabolite score had a C-index of 0.74 (95% CI = 0.68-0.79) (see table 2).

The analysis was repeated with pareto scaled data. This resulted in the selection of 6 signals. These 6 signals represented different 3 compounds (glutamate, Trimethylamine N-oxide [TMAO] and glucose), two signals representing lipid structures (CH_3 and CH_2) and glycoproteins. A metabolite score based on these metabolites and corresponding weights (range = -2.19-4.51) was associated with incident CHD before (HR = 1.68; 95% CI = 1.32-2.14) but not after adjusting for traditional CHD risk factors (HR = 1.16; 95% CI = 0.83-1.62). As shown in table 2, this metabolite score had a C-index of 0.72, (95% CI = 0.66-0.77).

Table 1 | Metabolite scores based on variables selected by LASSO regression and elastic net using the targeted and the untargeted approach and their association with incident coronary heart disease before and after adjusting for traditional risk factors.

	Scaling method	N predictors	Univariate		Adjusted ^a	
			HR (95% CI)	P-value	HR (95% CI)	P-value
Targeted approach	Auto	11	1.86 (1.46-2.37)	<0.001	1.46 (1.09-1.98)	0.011
	Pareto	6	1.68 (1.32-2.14)	<0.001	1.16 (0.83-1.62)	0.388
Untargeted approach	Auto	21	2.65 (1.94-3.64)	<0.001	2.25 (1.54-3.29)	<0.001
	Pareto	6	1.85 (1.43-2.41)	<0.001	1.17 (0.81-1.96)	0.394

HR, hazard ratio; CI, confidence interval; CHD, coronary heart disease; TRFs, traditional risk factors

^aAdjusted for sex, total cholesterol, HDL cholesterol, systolic blood pressure, body mass index, current smoking, self-reported diabetes and parental history of MI. Age in years was used as the time scale variable.

Untargeted approach

Using LASSO regression on the autoscaled and mean-centered data, 21 bins were selected as the best prediction set (Fig. 4). As shown in figure 5, eight of the 21 selected bins were located in a part of the ^1H -NMR spectra where no peaks were observed. A metabolite score constructed from these bins and their corresponding coefficients was associated with incident CHD before (HR = 2.65; 95% CI = 1.94-3.64) and after adjusting for TRFs (HR = 2.25; 95% CI = 1.54-3.29) (Table 1). This metabolite score had a C-index of 0.81, (95% CI = 0.78-0.85), resulting in a significantly better ($p < 0.001$) discrimination than the metabolite score based on the targeted approach (Table 2).

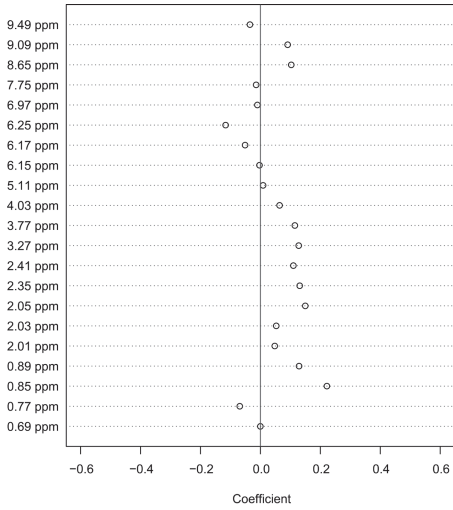


Fig 4 | The 21 bins and their corresponding coefficients selected by LASSO regression using the untargeted approach in combination with autoscaling.

When applying LASSO regression after pareto scaling 6 bins were selected as the best prediction set. A metabolite score constructed from these bins and their corresponding coefficients was associated with incident CHD before (HR = 1.85; 95% CI = 1.43-2.41), but not after adjusting for TRFs (HR = 1.17; 95% CI = 0.81-1.69). As shown in table 2, this metabolite score had a C-index of 0.73, (95% CI = 0.68-0.78), which is comparable to the C-index of the metabolite score obtained by the targeted approach after pareto scaling (C-index = 0.72; 95%CI = 0.66-0.77).

Table 2 | Discrimination of the models generated using the targeted and the untargeted approach.

Scaling Method	Targeted approach		Untargeted approach		P-value ^a
	C-index	95% CI	C-index	95% CI	
Auto	0.74	0.68-0.79	0.81	0.78-0.85	<0.001
Pareto	0.72	0.66-0.77	0.73	0.68-0.78	0.176

^aThe p-value obtained after comparing the models from the two approaches (targeted versus untargeted)

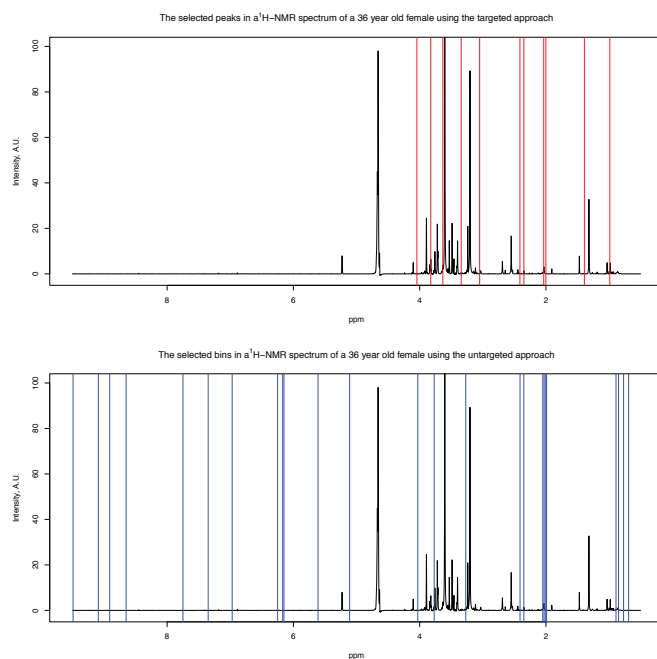


Fig. 5 | Both panels shown the same ^1H -NMR spectrum of a 36-year old women. In the upper panel, the red vertical lines indicate the location of the peaks selected by LASSO regression using the targeted approach in combination with autoscaling. In the lower panel, the blue vertical lines indicate the locations of the bins selected by LASSO regression using the untargeted approach in combination with autoscaling.

Discussion

Two approaches for extracting data from ^1H -NMR spectra obtained in EDTA plasma samples from individuals participating in a prospective case-cohort study for finding CHD risk markers were compared. A targeted approach, which involves identifying and quantifying individual signals in the ^1H -NMR spectra and an untargeted approach, for which we used data obtained by equidistant binning. For both approaches we also evaluated the use of auto scaling and paretoscaling.

When comparing the targeted versus the untargeted approach after autoscaling, we found that the metabolite scores based on both approaches were significantly associated with incident CHD before and after adjusting for TRFs. The metabolite score based on the untargeted approach showed a stronger association with incident CHD than the metabolite based on the targeted approach. Also the metabolites score based on the untargeted approach resulted in better discrimination than the metabolite score based on targeted approach. However, we observed that when using the untargeted approach, LASSO regression resulted in the selection of bins that were located in a part of the ^1H -NMR spectra where no peaks were observed. Possibly these bins represent only noise. When comparing the targeted versus the untargeted approach after pareto scaling, we found that the metabolite scores based on both approaches were associated with incident CHD, but not after adjusting for TRFs. In the univariate analysis, the association between the metabolite

score and incident CHD was slightly stronger than when the untargeted approach was used. The discriminatory capabilities for both metabolite scores were comparable.

At first sight the metabolite score based on untargeted approach combined with autoscaling as data pretreatment seems the best approach for predicting CHD. However, in a study conducted by Weljie *et al.* it was observed that the targeted approach resulted in more stable PCA clustering than when using the untargeted approach.³ The main reason for this observation is that very small and random fluctuations within the noise of a ¹H-NMR spectrum can lead to large and irrelevant variations in the PCA clustering, which was also observed by Halouska *et al.*¹⁸ In a study by Rubtsov *et al.* it was found that a targeted approach led to better results in a metabolomics study of liver toxicity response to acute phenobarbital exposure.⁵ They reasoned that the targeted approach gives a better resolution and sensitivity compared to equidistant binning because one bin can contain two or more separate peaks, thus reducing the resolution of those signals.⁵ We cannot confirm their findings since we found that after autoscaling the metabolite score based on the untargeted approach resulted in a better discrimination for incident CHD. After pareto scaling no differences between both approaches were observed.

A possible explanation for our observation that the untargeted approach in combination with autoscaling resulted in a better discrimination of incident CHD than the targeted approach could be that the inclusion of noise variables resulted in overfitting. Normally bins representing only noise have very low intensities. However autoscaling makes all bins, including the ones representing only noise, equally important.¹⁰ In contrast pareto scaling stays closer to the original measurement, thus giving less weight to the bins representing only noise.¹⁰ This could explain that after pareto scaling, the two approaches resulted in metabolite scores with comparable discriminatory capabilities. To answer the question if the good discriminatory capability of the metabolite score based on the untargeted approach combined with autoscaling is caused by overfitting, it is necessary to replicate this metabolite score in an independent prospective cohort and compare its performance with the metabolite scores based on the targeted approach. Another possibility is to split our dataset in a discovery cohort and a validation cohort.⁸ The discovery cohort can be used to construct the metabolite score, which is then tested in the validation cohort.⁸ The disadvantage of this method is that our study group might be too small for that.⁸

We conclude that for data reduction in metabolomics data the untargeted approach is possibly more prone to false positive results than the targeted approach. Therefore we hypothesize that it is better to use the targeted approach in combination with autoscaling. However to confirm this presumption, the metabolite scores need to be validated in independent prospective cohorts.

References

1. Nicholson JK, Lindon JC. Systems biology: Metabolomics. *Nature*. Oct 23 2008;455(7216):1054-1056.
2. Mayr M. Metabolomics: ready for the prime time? *Circulation. Cardiovascular genetics*. Oct 2008;1(1):58-65.
3. Weljie AM, Newton J, Mercier P, Carlson E, Slupsky CM. Targeted profiling: quantitative analysis of 1H NMR metabolomics data. *Analytical chemistry*. Jul 1 2006;78(13):4430-4442.
4. De Meyer T, Sinnaeve D, Van Gasse B, et al. Evaluation of standard and advanced preprocessing methods for the univariate analysis of blood serum 1H-NMR spectra. *Analytical and bioanalytical chemistry*. Oct 2010;398(4):1781-1790.
5. Rubtsov DV, Waterman C, Currie RA, et al. Application of a Bayesian deconvolution approach for high-resolution (1)H NMR spectra to assessing the metabolic effects of acute phenobarbital exposure in liver tissue. *Analytical chemistry*. Jun 1 2010;82(11):4479-4485.
6. Tredwell GD, Behrends V, Geier FM, Liebeke M, Bundy JG. Between-person comparison of metabolite fitting for NMR-based quantitative metabolomics. *Analytical chemistry*. Nov 15 2011;83(22):8683-8687.
7. Wishart DS. Quantitative metabolomics using NMR. *Trac-Trend Anal Chem*. Mar 2008;27(3):228-237.
8. Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. *Nature reviews. Cancer*. Apr 2004;4(4):309-314.
9. Keun HC, Ebbels TMD, Antti H, et al. Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling. *Analytica chimica acta*. Aug 25 2003;490(1-2):265-276.
10. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*. 2006;7:142.
11. Prentice RL. A Case-Cohort Design for Epidemiologic Cohort Studies and Disease Prevention Trials. *Biometrika*. Apr 1986;73(1):1-11.
12. Tibshirani R. The lasso method for variable selection in the Cox model. *Statistics in medicine*. Feb 28 1997;16(4):385-395.
13. Tibshirani R. Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B Met*. 1996;58(1):267-288.
14. Picavet HS, Schouten JS. Physical load in daily life and low back problems in the general population-The MORGEN study. *Preventive medicine*. Nov 2000;31(5):506-512.
15. Merry AH, Boer JM, Schouten LJ, et al. Validity of coronary heart diseases and heart failure based on hospital discharge and mortality data in the Netherlands using the cardiovascular registry Maastricht cohort study. *European journal of epidemiology*. 2009;24(5):237-247.
16. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA : the journal of the American Medical Association*. May 14 1982;247(18):2543-2546.
17. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*. 2010;33(1):1-22.
18. Halouska S, Powers R. Negative impact of noise on the principal component analysis of NMR data. *Journal of magnetic resonance*. Jan 2006;178(1):88-95.

Lipid metabolism in long-lived families: The Leiden Longevity Study

Anika A. M. Vaarhorst¹, Marian Beekman¹, Eka H. D. Suchiman¹, Diana van Heemst²,
Jeanine J. Houwing-Duistermaat³, Rudi G. J. Westendorp², P. Eline Slagboom^{1,4},
Bastiaan T. Heijmans¹ On behalf of the project group
and the Leiden Longevity Study (LLS) Group⁵

1. Molecular Epidemiology Section, Department Medical Statistics and Bioinformatics, Leiden University Medical Centre, The Netherlands
2. Department of Gerontology and Geriatrics, Leiden University Medical Centre, Leiden, The Netherlands
3. Medical Statistics Section, Department Medical Statistic and Bioinformatics, Leiden University Medical Centre, Leiden, The Netherlands
4. Netherlands Consortium for Healthy Ageing, Leiden, The Netherlands
5. LUMC (AAM Vaarhorst, PE Slagboom, BT Heijmans, M Beekman, JN Kok, JW Jukema), WU (Y Lu, EJM Feskens, M Muller), RIVM (JMA Boer, MET Dollé), UNIMAAS/UHM (MMJ van Greevenbroek, APM Gorgels), and the VU (DI Boomsma)

Abstract

Mechanisms underlying the variation in human life expectancy are largely unknown, but lipid metabolism and especially lipoprotein size was suggested to play an important role in longevity. We have performed comprehensive lipid phenotyping in the Leiden Longevity Study (LLS). By applying multiple logistic regression analysis we tested for the first time the effects of parameters in lipid metabolism (*i.e.*, classical serum lipids, lipoprotein particle sizes, and apolipoprotein E levels) on longevity independent of each other. Parameters in lipid metabolism were measured in offspring of nonagenarian siblings from 421 families of the LLS ($n=1,664$; mean age, 59 years) and in the partners of the offspring as population controls ($n=711$; mean age, 60 years). In the initial model, where lipoprotein particle sizes, classical serum lipids and apolipoprotein E were included, offspring had larger low-density lipoprotein (LDL) particle sizes ($p=0.017$), and lower triglyceride levels ($p=0.026$), indicating that they displayed a more beneficial lipid profile. After backwards regression only LDL size ($p=0.014$) and triglyceride levels ($p=0.05$) were associated with offspring from long-lived families. Sex-specific backwards regression analysis revealed that LDL particle sizes were associated with male longevity (increase in log odds ratio (OR) per unit=0.21; $p=0.023$). Triglyceride levels (decrease OR per unit=0.22; $p=0.01$), but not LDL particle size, were associated with female longevity. Due to the analysis of a comprehensive lipid profile, we confirmed an important role of lipid metabolism in human longevity, with LDL size and triglyceride levels as major predicting factors.

Introduction

Over the last century, life expectancy has dramatically increased in western societies;¹ nevertheless, inter-individual differences in life expectancy remain. For example, family members of nonagenarian siblings and centenarians have a 30% survival benefit for three generations compared with their birth cohort.^{2,3} Offspring of nonagenarians siblings and centenarians are characterized by a lower prevalence of several major diseases, including diabetes, hypertension, coronary artery disease (CAD),^{4,5} and strokes,⁶ implying that these differences in life expectancy are associated with a familial, possible genetic factor that influences the risk of metabolic diseases.

The mechanisms underlying the variation in human life expectancy are still largely unknown but various studies suggest involvement of lipid metabolism. Specifically, low plasma apoE levels in elderly apolipoprotein E gene (*APOE*) $\epsilon 3/\epsilon 3$ carriers are associated with a decreased mortality risk⁷ and in comparison with same-aged controls, offspring of centenarians or nonagenarian siblings had a tendency for higher plasma levels of high-density lipoprotein cholesterol (HDL-C), lower levels of low-density lipoprotein cholesterol (LDL-C),⁸ and larger lipoprotein particle sizes.^{9,10} Studies into longevity performed so far, often analysed a subset of blood lipid parameters only and were relatively small. Their findings thus require confirmation in larger studies, particularly because they could not account for the correlation between the various lipid parameters. The importance of the latter is illustrated by the disappearance of the association of HDL particle size with coronary artery disease and familial longevity after adjustment for triglyceride and/or apolipoprotein B levels.⁹⁻¹¹

To improve the understanding of the association between lipid metabolism and human longevity, we performed multiple analyses of classical blood lipid parameters, lipoprotein particle sizes and apoE levels in long-lived families from the Leiden Longevity Study. We compared 1,664 offspring of nonagenarian sibships, who were shown to have a life-long survival advantage³ and a decreased incidence of metabolic and cardiovascular disease,⁵ with 711 similarly aged partners as controls. Since men and women differ in lipid metabolism and life expectancy,¹² we also stratified for gender in our analysis.

Material and methods

Study design and subjects

The Leiden Longevity Study (Fig. 1) included 421 Caucasian families consisting of long-lived siblings together with their offspring and the spouses of the offspring.³ Long-lived families were recruited if at least two long-lived siblings were alive and willing to participate. Males were considered long-lived if 89 years or older and females 91 years or older. The mean age of the long-lived siblings was 93.4 ± 2.6 years. In 2001, less than 0.5% of the Dutch population fulfilled these criteria. The Medical Ethical Committee of the Leiden University Medical Centre approved the study and informed consent was obtained from all subjects.

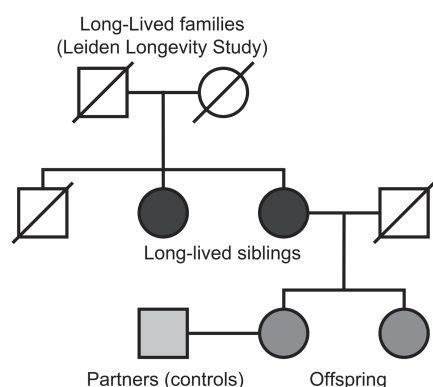


Figure 1 | Schematic representation of the Leiden Longevity Study. The Leiden longevity study consists of 982 long-lived siblings (males >89 years; females >91 years), their offspring (n=1,671), and the partners of the offspring (n=745)

Previously, lipoprotein particle sizes have been analysed in 165 families from the Leiden Longevity Study using a 400-MHz proton NMR analyser at LipoScience.¹⁰ Between June 2007 and September 2007, lipoprotein particle sizes were measured in the remaining 255 families using the same method as described in Heijmans *et al.* 2006.¹⁰ In short this technique takes advantage of the natural proton NMR spectroscopy “signatures” emitted by lipoprotein particles of different sizes. This information can be used to calculate the concentration of each lipoprotein particle subclass. Next, mean particle sizes (diameter in nanometers) were determined by weighting the relative mass percentage of each subclass according to its diameter.¹³ Together with the previous study, lipoprotein particle sizes were available for 2,416 individuals consisting of the offspring (n=1,671) of long-lived siblings and the partners of the offspring as controls (n=745). Between November 2006 and May 2008, additional information on self-reported bodily measures was collected for the offspring and the controls, and in addition information on medication use was requested from the participants’ pharmacist.⁵

Biochemical analysis

In all subjects non-fasting venous blood samples were taken. Lipid levels (triglycerides, total cholesterol, and HDL-C) were determined in serum samples using fully automated equipment (the Hitachi Modular or the Cobas Integra 800 both from Roche, Almere, the Netherlands).¹⁴ LDL-C levels were calculated using the Friedewald formula ($\text{LDL-C} = \text{total cholesterol} - \text{HDL-C} - (\text{triglycerides}/2.2)$; unit mmol/L) and set to missing if plasma triglyceride levels exceeded 4.52 mmol/L.¹⁰ ApoE was determined in serum samples using a human apoE-specific sandwich ELISA.^{7,15}

APOE ϵ 2/ ϵ 3/ ϵ 4 genotyping

DNA was isolated using standard techniques (QIAamp bloodmaxi kit, QIAGEN, (Venlo, the Netherlands) high salt extraction). For determining APOE ϵ 2/ ϵ 3/ ϵ 4 genotypes, two SNPs in the APOE ϵ 2/ ϵ 3/ ϵ 4 gene were genotyped using two Taqman SNP genotyping assays with the following assayIDs: C_904973_10 (rs7412) and C_3084793_20 (rs429358) (APPLIED BIOSYSTEMS, Foster City, USA). The assay was run on a 7900HT (APPLIED BIOSYSTEMS) according to manufacturer's specification and genotypes were called using the Sequence Detection Software version 2.2 (APPLIED BIOSYSTEMS). The genotype call rate for the APOE ϵ 2/ ϵ 3/ ϵ 4 polymorphism was 98.7%.

Statistical analysis

Continuous variables were tested for normality; if appropriate, they were logarithmically transformed to obtain a normal distribution. Differences in lipid parameters between offspring and their controls were assessed using logistic regression, adjusted for age, sex, and the interaction between them. Robust standard errors were used to adjust for dependencies within families. To determine the effects of lipid parameters in more detail and independent of each other, multiple logistic regression was used. The lipid variables LDL particle size, HDL particle size, LDL-C, HDL-C, triglyceride levels, and ApoE levels were included in the initial starting model together with the covariates age, sex, and the interaction between them. Using a backward regression procedure, the lipid parameter with the highest p-value was removed each round until a model only including lipid variables independently associating with familial longevity was obtained. We also stratified for sex. Differences between offspring and controls for genotype distributions of rs5882 and the APOE ϵ 2/ ϵ 3/ ϵ 4 polymorphism were evaluated using the chi-square test. Only subjects with complete measurements for all the biochemical variables were included in the analysis. STATA/SE version 8.0 for Windows was used for data analysis.

Results

To test the association of blood lipid parameters with human longevity, we studied individuals for whom complete measurements were available (total cholesterol, HDL-C, LDL-C, triglycerides levels, ApoE, mean HDL, and LDL particle sizes). Compared with controls (n=711), offspring of long-lived siblings (n=1,664) had lower triglyceride levels (1.67 vs. 1.75 mmol/L; $p=0.001$), larger mean HDL particle sizes (9.05 vs. 9.03 nm; $p=0.01$), larger mean LDL particle sizes (21.32 vs. 21.22 nm; $p<10^{-3}$), and a smaller total cholesterol/HDL-C ratio (4.08 vs. 4.20; $p=0.004$). These observations were independent of age, sex, and the interaction between them (Table 1). Further adjustment for self-reported body mass index (BMI), which was available in 619 controls and 1,399 offspring, did not affect these associations (triglyceride levels, $p=0.01$; HDL particle size, $p=0.03$; LDL particle size, $p<10^{-3}$; total cholesterol/HDL-C ratio, $p=0.03$). As shown in Table 1, no differences were observed for apoE levels for the complete group, even when the analysis was restricted to *APOE* $\epsilon 3/\epsilon 3$ carriers only. We observed that the genotype distribution for the *APOE* polymorphism differs between offspring and controls ($\epsilon 2/\epsilon 2=0.4$, $\epsilon 2/\epsilon 3=10.2$, $\epsilon 2/\epsilon 3=3.2$, $\epsilon 3/\epsilon 3=62.3$, $\epsilon 3/\epsilon 4=21.3$ and $\epsilon 4/\epsilon 4=2.5$; $\epsilon 2/\epsilon 2=1.0$, $\epsilon 2/\epsilon 3=13.3$, $\epsilon 2/\epsilon 3=2.3$, $\epsilon 3/\epsilon 3=60.8$, $\epsilon 3/\epsilon 4=21.2$, and $\epsilon 4/\epsilon 4=1.2$, respectively, $p=0.019$).

Sex-specific analysis revealed that female offspring had lower triglyceride levels (1.51 mmol/L vs. 1.62 mmol/L, $p=0.001$), a lower cholesterol/HDL-C ratio (3.75 vs. 3.88, $p=0.018$), larger LDL particle sizes (21.60 nm vs. 21.49 nm, $p=0.01$), and larger HDL particle sizes (9.25 nm vs. 9.21 nm, $p=0.031$) compared with controls. After adjusting for BMI, HDL particle size and the cholesterol/HDL-C ratio were no longer significant different between offspring and controls (data not shown). In males however, offspring only had larger LDL sizes (20.99 nm vs. 20.86 nm, $p=0.023$) compared to offspring and for the cholesterol/HDL-C ratio a trend was observed were offspring had a lower ratio compared to controls (4.47 vs. 4.64, $p=0.080$). Adjustment for BMI did not affect these results (data not shown).

Table 1 | Baseline characteristics of the Leiden Longevity Study population

Leiden Longevity Study	Controls Mean (SD)	Offspring Mean (SD)	P value
Number of individuals	711	1664	-
Females N (%)	409 (57.5)	898 (54.0)	
Lipid lowering medication N (%) ^a	68 (9.56)	109 (6.55)	0.011
Males	32 (10.60)	60 (7.83)	0.172
Females	36 (8.80)	49 (5.46)	0.032
Age (yr)	58.87 ± 7.45	59.40 ± 6.53	0.046
Males	61.55 ± 7.33	59.26 ± 6.50	<10 ⁻³
Females	56.88 ± 6.91	59.52 ± 6.56	<10 ⁻³
Body Mass Index ^c	25.56 ± 3.67	25.28 ± 3.46	0.055 ^d
Males	25.84 ± 3.27	25.64 ± 2.86	0.486 ^e
Females	25.34 ± 3.94	24.97 ± 3.86	0.060 ^e
Triglyceride (mmol/L) ^b	1.75 ± 0.86	1.67 ± 0.82	0.001 ^d
Males	1.93 ± 0.89	1.86 ± 0.85	0.311 ^e
Females	1.62 ± 0.81	1.51 ± 0.76	0.001 ^e
HDL-cholesterol (mmol/L)	1.43 ± 0.46	1.46 ± 0.44	0.053 ^d
Males	1.25 ± 0.36	1.29 ± 0.37	0.104 ^e
Females	1.57 ± 0.48	1.60 ± 0.45	0.240 ^e
LDL-cholesterol (mmol/L)	3.33 ± 0.92	3.34 ± 0.99	0.750 ^d
Males	3.27 ± 0.94	3.30 ± 0.94	0.997 ^e
Females	3.37 ± 0.94	3.37 ± 1.03	0.701 ^e
Total cholesterol/HDL ratio ^b	4.20 ± 1.39	4.08 ± 1.28	0.004 ^d
Males	4.64 ± 1.49	4.47 ± 1.29	0.080 ^e
Females	3.88 ± 1.21	3.75 ± 1.17	0.018 ^e
HDL particle size (nm)	9.03 ± 0.51	9.05 ± 0.50	0.010 ^d
Males	8.79 ± 0.43	8.83 ± 0.44	0.172 ^e
Females	9.21 ± 0.50	9.25 ± 0.46	0.031 ^e
LDL particle size (nm)	21.22 ± 0.83	21.32 ± 0.81	<10 ^{-3d}
Males	20.86 ± 0.76	20.99 ± 0.78	0.023 ^e
Females	21.49 ± 0.78	21.60 ± 0.73	0.010 ^e
ApoE (mg/dL) ^b	2.81 ± 0.98	2.82 ± 0.99	0.967 ^d
Males	2.72 ± 0.92	2.76 ± 0.97	0.424 ^e
Females	2.87 ± 1.02	2.87 ± 1.01	0.481 ^e
APOEε3/ε3 carriers	441 (62.03)	989 (59.44)	-
Males	198 (65.56)	443 (57.83)	-
Females	243 (59.41)	546 (60.80)	-
ApoE (mg/dL) in APOEε3/ε3 carriers	2.71 ± 0.73	2.68 ± 0.79	0.356
Males	2.66 ± 0.72	2.61 ± 0.81	0.495
Females	2.80 ± 0.74	2.73 ± 0.77	0.104

^aLipid lowering medication includes 3-hydroxy-3-methylglutaryl-coenzyme-A reductase inhibitors (statin), fibrates, niacin, bile acid sequestrants, and Niemann-Pick C1-like 1 protein inhibitors

^bLogarithmically transformed to obtain a normal distribution

^cBody Mass Index was calculated based on self-reported height and weight in 1867 participants of the LLS, consisting of 822 males and 1045 females.

^dAdjusted for medication use, age, sex and the interaction between age and sex

^eAdjusted for age

Table 2 | Pearson correlation coefficients^a between the lipid parameters in the Leiden Longevity Study.

		HDL size	TG	CHOL	HDL-C	LDL-C	Chol/HDL	ApoE
LDL size	Total population	0.793	-0.470	0.067	0.624	-0.016**	-0.606	-0.074
	Men	0.778	-0.489	-0.027**	0.633	-0.077*	-0.621	-0.141
	Women	0.734	-0.375	0.015	0.518	-0.076	-0.527	-0.077
HDL size	Total population	-	-0.550	-0.005**	0.742	-0.143	-0.754	-0.065
	Men	-	-0.519	-0.092	0.736	-0.150	-0.741	-0.169
	Women	-	-0.525	-0.075	0.683	-0.194	-0.741	-0.051**
TG	Total population	-	-	0.327	-0.372	0.191	0.614	0.233
	Men	-	-	0.378	-0.393	0.150	0.616	0.306
	Women	-	-	0.390	-0.290	0.244	0.587	0.209
CHOL	Total population	-	-	-	0.364	0.924	0.317	0.188
	Men	-	-	-	0.255	0.922	0.389	0.169
	Women	-	-	-	0.364	0.924	0.360	0.198
HDL-C	Total population	-	-	-	-	0.129	-0.743	0.021**
	Men	-	-	-	-	0.119*	-0.765	-0.087*
	Women	-	-	-	-	0.126	-0.715	0.053**
LDL-C	Total population	-	-	-	-	-	0.486	0.115
	Men	-	-	-	-	-	0.508	0.078*
	Women	-	-	-	-	-	0.559	0.103
CHOL/HDL	Total population	-	-	-	-	-	-	0.129
	Men	-	-	-	-	-	-	0.195
	Women	-	-	-	-	-	-	0.116

^aSignificant at <0.01 (2-tailed), unless otherwise indicated: *0.01< p <0.05, ** p >0.05

LDL size, mean LDL lipoprotein particle size; *HDL size* mean, HDL lipoprotein particle size; *TG*, triglycerides; *CHOL*, cholesterol; *HDL-C*, HDL-cholesterol; *LDL-C*, LDL-cholesterol; *CHOL/HDL*, ratio between total cholesterol and HDL-cholesterol; *ApoE*, apolipoprotein E

These univariate analyses do not account for the strong correlations that exist between the various lipid parameters. For example, the Pearson's correlation coefficient between LDL particle size and triglyceride levels is -0.470, and the correlation between HDL-C and LDL particle size is 0.624 (Table 2). To assess the association between longevity and the lipid parameters in more detail and independent of each other, we used multiple logistic regression where the starting model included the all the available lipid parameters, except for total cholesterol and the ratio LDL-C/HDL-C which were not included because they are highly correlated with LDL-C and HDL-C, respectively, which introduces multicollinearity. In this initial model (Table 3), the lipid parameters LDL size (increase in log OR per unit=0.21; p =0.017) and triglyceride levels (decrease in log OR per unit=0.26; p =0.026) were significantly different between offspring and controls independent from the other lipid variables. After

backwards regression, were age, sex, and the interaction between age and sex were not removed to account for the study design, the final model included LDL particle size (increase in log OR per unit=0.16; $p=0.014$) and triglyceride levels (decrease in log OR per unit=0.21; $p=0.05$).

Table 3 | Multiple variate logistic regression analysis with all measured lipid variables included.

	Total population (N=2375)			Males (N=1068)			Females (N=1307)		
	<i>beta</i> ^a	<i>SE</i>	<i>p-value</i>	<i>beta</i> ^a	<i>SE</i>	<i>p-value</i>	<i>beta</i> ^a	<i>SE</i>	<i>p-value</i>
Age (yr)	0.062	0.010	<10 ⁻³	-0.050	0.012	<10 ⁻³	0.066	0.010	<10 ⁻³
Sex	6.901	1.066	<10 ⁻³	-	-	-	-	-	-
Sex*Age	-0.112	0.018	<10 ⁻³	-	-	-	-	-	-
LDL size (nm)	0.207	0.087	0.017	0.268	0.140	0.055	0.178	0.117	0.127
HDL size (nm)	-0.124	0.201	0.537	-0.283	0.333	0.395	-0.064	0.259	0.805
LDL-C (mmol/L)	-0.001	0.053	0.992	-0.031	0.080	0.697	0.024	0.074	0.741
HDL-C (mmol/L)	0.010	0.177	0.953	0.235	0.322	0.466	-0.080	0.219	0.713
Triglycerides (mmol/L) ^b	-0.257	0.115	0.026	-0.039	0.171	0.821	-0.417	0.164	0.011
ApoE (mmol/L) ^b	0.117	0.163	0.473	0.240	0.226	0.287	0.000	0.203	1.000
Constant	-6.210	1.784	0.000	0.425	2.898	0.883	-6.092	2.391	0.011

People using lipid lowering medication were excluded from analysis. Robust standard errors were used to account for dependencies within families in all analysis

SE, standard error; *LDL size*, mean LDL lipoprotein particle size; *HDL size*, mean HDL lipoprotein particle size; *HDL-C*, HDL-cholesterol; *LDL-C*, LDL-cholesterol; *CHOL/HDL*, ratio between total cholesterol and HDL-cholesterol; *ApoE*, apolipoprotein E

^a Change in log OR per unit

^b Logarithmically transformed to obtain a normal distribution

Sex-specific multiple regression indicated differences between men and women. In the initial multivariate logistic regression model, none of the lipid parameters was significantly different in male offspring compared with controls, but for LDL size a trend was observed ($p=0.055$) (Table 3). After backwards regression, male offspring had a larger LDL particle size than controls (increase in log OR per unit=0.21; $p=0.023$), whereas no differences in triglyceride levels were observed. For female offspring, in the initial starting model, only triglyceride levels were lower among female offspring (decrease in log OR per unit=0.42; $p=0.011$) (Table 3). After backwards regression, no difference in LDL size were observed in female offspring as compared with controls, whereas triglyceride levels were significantly lower among female offspring (decrease in log OR per unit=0.22; $p=0.01$).

The valine allele of rs5882 in the cholesterol ester transfer protein gene was previously found to be associated with larger LDL and HDL particle sizes.⁹ The association of the valine allele with larger LDL size was replicated in a linear regression analysis in controls and

offspring combined ($p=0.008$; adjusted for age, sex, and triglyceride levels), but this allele was not associated with HDL size ($p=0.200$). The polymorphism, however, did not explain the difference in LDL size between controls and offspring (increase in log OR per unit= 0.16; $p=0.009$ after accounting for rs5882, as compared with increase in log OR per unit=0.17; $p=0.014$ without rs5882). Nor was the valine allele associated with the offspring group ($p=0.131$). Similar results were obtained for man and women separately.

Discussion

We investigated the association of classical serum lipid levels, apoE levels and lipoprotein particle sizes with human longevity, in a study that was sufficiently large to account for correlations among lipid parameters and to test for differences between sexes. Our main finding is that both LDL particle size and triglyceride levels are associated with human longevity independent of other lipid parameters. Further analysis indicated that LDL particle size is particularly associated with male, whereas triglyceride levels are associated with female longevity.

Our study ($N=2,375$) confirms the association of LDL particle size previously observed in a smaller study ($N=474$).⁹ The differences in the final end model were comparable with the findings from Barzilai for males (0.21 nm increase), but not for females (in the final model only triglyceride levels were significantly different between offspring and controls). Remarkable, an association of triglyceride levels with longevity was not reported before. In one study, only a trend was observed¹⁰ and in another study triglycerides were not mentioned in relation to longevity.⁹ In a very small study where fasting samples were used, no association between longevity and triglyceride levels were found.⁸ Since we used non-fasting samples and had a large study population, this could explain why we did find an association with triglyceride levels and longevity. Our results did not confirm the previously reported association of LDL and HDL cholesterol levels⁸ and HDL particle size.⁹ For HDL particle size we did observe an association in a univariate analysis, but a multiple analysis that accounted for the correlation between lipid parameters showed that this result depended on other blood lipid parameters in particular triglyceride levels.¹⁰

Because the *APOE* genotype was previously found to confound the association between ApoE level and all-cause mortality in elderly subjects,⁷ it was necessary to stratify the data according to *APOE* genotype. Still we did not find an association of apoE level with longevity. However, we did find that the beneficial *APOEε2* allele was more common in the offspring compared with controls, which might explain why offspring have a more beneficial lipid profile compared with controls.⁷

Since offspring of nonagenarian siblings from the Leiden Longevity Study have a lower prevalence of late-onset metabolic diseases,⁵ it could be hypothesized that low triglyceride levels and a larger mean LDL particle size are protective against diseases like type II diabetes and CAD. High triglyceride levels are a component of the metabolic syndrome, which is associated with a higher risk for cardiovascular diseases (CVD) and type II diabetes.¹⁶ Triglyceride levels itself are also independently associated with a higher risk for CVD.¹⁷ Elevated triglyceride levels are associated with increased levels of remnants lipoproteins, which might initiate and promote the development of atherosclerotic plaques.¹⁸ Mean LDL particle size is determined by summing the size-weighted relative contribution made by each subspecies (*i.e.*, small, medium and large LDL particles) to the total concentration of LDL particles.¹⁹ Thus, people with a small mean LDL particle size have relatively more small LDL particles as compared with people with a large mean LDL particle sizes. Small LDL particle sizes are associated with type II diabetes,²⁰ metabolic syndrome,^{21,22} insulin resistance,^{23,24} incident CVD,²⁵ and future CAD.²⁶ Possible atherosclerosis could be induced by those small LDL particles, as indicated in a study of two knockout mice strains with similar cholesterol levels, where the *Ldlr*^{-/-}*Apob100/100* mice had a larger number of smaller lipoprotein proteins and more atherosclerosis than the *Apoe*^{-/-}*Apob100/100* mice.²⁷ Thus from literature it is clear that LDL particle size and triglyceride levels have been contributing to cardiovascular diseases.

LDL particle size seems to be particularly associated with male longevity whereas triglyceride levels are associated with female longevity, indicating that the role of lipid metabolism in longevity may be different in men and women. In women it was found that LDL particle size associated with incident CVD, however in that study LDL particle size was comparable but not superior to standard lipid levels in predicting CVD.²⁵ This could support our finding that triglyceride levels eliminated the association between LDL size and longevity in women. Moreover, in a meta-analysis of population based prospective studies it was concluded that high triglyceride levels are a larger risk factor for cardiovascular disease in women than in men.²⁸ But in a more recent meta-analysis, it was found that the effect of triglyceride levels on cardiovascular disease did not differ between men and women.¹⁷ On the other hand, men are characterized by smaller and denser LDL particles and higher triglyceride levels than women,^{12,29} which is also the case in our study. In addition, men have higher hepatic lipase activity compared with women, resulting in a smaller mean LDL particle size.^{30,31} Thus different plasma concentrations of enzymes involved in intravascular lipid remodelling between men and women³¹ may explain our observed sex differences regarding LDL particle size and familial longevity. But further research on lipid enzymes difference between men and women in relation to longevity is needed to confirm this hypothesis.

This study has several strong points. Firstly, due to the large sample size it was possible to adjust for the correlations among the lipid parameters and to stratify for gender. Secondly, since controls are the partners of the offspring of long-lived siblings, they did not differ on any major indicators of lifestyle, including current smoking, self-reported BMI, and level of education.⁵ A weak point is that the NMR spectroscopy-based lipid parameters were not determined in a single measurement which introduced batch effects, for which it was necessary to adjust during analysis. We choose not to correct for the treatment of high cholesterol levels with lipid lowering medication, which mostly consist of 3-hydroxy-3-methylglutaryl-coenzyme-A reductase inhibitor (statin) treatment (93.8%), because per individual it is unknown if statin treatment reduces moderate to severe hypercholesterolemia to above or below population average lipid levels. Since the offspring group is using lesser lipid lowering medication than the controls, this might bias our results in such a way that we underestimate the effect of lipid metabolism on longevity. Also, the use of non-fasting samples introduced noise in our dataset, which may lead to an underestimation of the effects found our study.

Our study confirms the link between lipid metabolism and human longevity and identifies LDL particle size and triglyceride levels as the key lipid parameters. To gain more insight into why there are differences between males and females, it is necessary to include information on lipid enzymes, genetic and expression data of genes known to be involved in lipid metabolism in future studies investigating the relation between lipid metabolism and longevity in humans.

Acknowledgements

We wish to thank all the participants of the Leiden Longevity Study for their contribution. We are grateful to Marja Kersbergen and Margo van Schie for determining apoE plasma levels and standard lipid parameters.

This study is supported by grants from Innovation Oriented research Program IOP on genomics (SenterNovem; IGE5007), the Dutch Heart Foundation (NHS2006B195), the Netherlands Consortium for Healthy Ageing (050 60 810), and by a stimulation grant (05040202, Healthy Ageing) from the Netherlands Genomics Initiative (NGI)/Netherlands Organization for scientific

References

- Oeppen J, Vaupel JW. Demography - Broken limits to life expectancy. *Science*. May 10 2002;296(5570):1029-1031.
- Perls T, Kohler IV, Andersen S, et al. Survival of parents and siblings of supercentenarians. *The journals of gerontology. Series A, Biological sciences and medical sciences*. Sep 2007;62(9):1028-1034.
- Schoenmaker M, de Craen AJ, de Meijer PH, et al. Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study. *European journal of human genetics : EJHG*. Jan 2006;14(1):79-84.
- Adams ER, Nolan VG, Andersen SL, Perls TT, Terry DF. Centenarian offspring: start healthier and stay healthier. *Journal of the American Geriatrics Society*. Nov 2008;56(11):2089-2092.
- Westendorp RG, van Heemst D, Rosing MP, et al. Nonagenarian siblings and their offspring display lower risk of mortality and morbidity than sporadic nonagenarians: The Leiden Longevity Study. *Journal of the American Geriatrics Society*. Sep 2009;57(9):1634-1637.
- Terry DF, Wilcox MA, McCormick MA, Perls TT. Cardiovascular disease delay in centenarian offspring. *The journals of gerontology. Series A, Biological sciences and medical sciences*. Apr 2004;59(4):385-389.
- Mooijaart SP, Berbee JF, van Heemst D, et al. ApoE plasma levels and risk of cardiovascular mortality in old age. *PLoS medicine*. Jun 2006;3(6):e176.
- Barzilai N, Gabriely I, Gabriely M, Iankowitz N, Sorkin JD. Offspring of centenarians have a favorable lipid profile. *Journal of the American Geriatrics Society*. Jan 2001;49(1):76-79.
- Barzilai N, Atzmon G, Schechter C, et al. Unique lipoprotein phenotype and genotype associated with exceptional longevity. *JAMA : the journal of the American Medical Association*. Oct 15 2003;290(15):2030-2040.
- Heijmans BT, Beekman M, Houwing-Duistermaat JJ, et al. Lipoprotein particle profiles mark familial and sporadic human longevity. *PLoS medicine*. Dec 2006;3(12):e495.
- El Harchaoui K, Arsenault BJ, Franssen R, et al. High-density lipoprotein particle size and concentration and coronary risk. *Annals of internal medicine*. Jan 20 2009;150(2):84-93.
- Freedman DS, Otvos JD, Jeyarajah EJ, et al. Sex and age differences in lipoprotein subclasses measured by nuclear magnetic resonance spectroscopy: the Framingham Study. *Clinical chemistry*. Jul 2004;50(7):1189-1200.
- Otvos JD. Measurement of lipoprotein subclass profiles by nuclear magnetic resonance spectroscopy. *Clinical laboratory*. 2002;48(3-4):171-180.
- Rosing MP, Westendorp RG, Frolich M, et al. Human insulin/IGF-1 and familial longevity at middle age. *Aging*. Aug 2009;1(8):714-722.
- van Vlijmen BJ, van den Maagdenberg AM, Gijbels MJ, et al. Diet-induced hyperlipoproteinemia and atherosclerosis in apolipoprotein E3-Leiden transgenic mice. *The Journal of clinical investigation*. Apr 1994;93(4):1403-1410.
- National Cholesterol Education Program Expert Panel on Detection E, Treatment of High Blood Cholesterol in A. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation*. Dec 17 2002;106(25):3143-3421.
- Sarwar N, Danesh J, Eiriksdottir G, et al. Triglycerides and the risk of coronary heart disease: 10,158 incident cases among 262,525 participants in 29 Western prospective studies. *Circulation*. Jan 30 2007;115(4):450-458.
- Nordestgaard BG, Benn M, Schnohr P, Tybjaerg-Hansen A. Nonfasting triglycerides and risk of myocardial infarction, ischemic heart disease, and death in men and women. *JAMA : the journal of the American Medical Association*. Jul 18 2007;298(3):299-308.
- Otvos JD, Jeyarajah EJ, Bennett DW, Krauss RM. Development of a proton nuclear magnetic resonance spectroscopic method for determining plasma lipoprotein concentrations and subspecies distributions from a single, rapid measurement. *Clinical chemistry*. Sep 1992;38(9):1632-1638.
- Festa A, Williams K, Hanley AJ, et al. Nuclear magnetic resonance lipoprotein abnormalities in prediabetic subjects in the Insulin Resistance Atherosclerosis Study. *Circulation*. Jun 28 2005;111(25):3465-3472.
- Gentile M, Panico S, Jossa F, et al. Small dense LDL particles and metabolic syndrome in a sample of middle-aged women. Findings from Progetto Athena. *Clinica chimica acta; international journal of clinical chemistry*. Feb 2008;388(1-2):179-183.
- Kathiresan S, Otvos JD, Sullivan LM, et al. Increased small low-density lipoprotein particle

- number: a prominent feature of the metabolic syndrome in the Framingham Heart Study. *Circulation*. Jan 3 2006;113(1):20-29.
23. Garvey WT, Kwon S, Zheng D, et al. Effects of insulin resistance and type 2 diabetes on lipoprotein subclass particle size and concentration determined by nuclear magnetic resonance. *Diabetes*. Feb 2003;52(2):453-462.
24. Goff DC, Jr., D'Agostino RB, Jr., Haffner SM, Otvos JD. Insulin resistance and adiposity influence lipoprotein size and subclass concentrations. Results from the Insulin Resistance Atherosclerosis Study. *Metabolism: clinical and experimental*. Feb 2005;54(2):264-270.
25. 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 cardiovascular disease in women. *Circulation*. Feb 24 2009;119(7):931-939.
26. El Harchaoui K, van der Steeg WA, Stroes ES, et al. Value of low-density lipoprotein particle number and size as predictors of coronary artery disease in apparently healthy men and women: the EPIC-Norfolk Prospective Population Study. *Journal of the American College of Cardiology*. Feb 6 2007;49(5):547-553.
27. Veniant MM, Beigneux AP, Bensadoun A, Fong LG, Young SG. Lipoprotein size and susceptibility to atherosclerosis--insights from genetically modified mouse models. *Current drug targets*. Mar 2008;9(3):174-189.
28. Hokanson JE, Austin MA. Plasma triglyceride level is a risk factor for cardiovascular disease independent of high-density lipoprotein cholesterol level: a meta-analysis of population-based prospective studies. *Journal of cardiovascular risk*. Apr 1996;3(2):213-219.
29. Lemieux I, Pascot A, Lamarche B, et al. Is the gender difference in LDL size explained by the metabolic complications of visceral obesity? *European journal of clinical investigation*. Dec 2002;32(12):909-917.
30. Carr MC, Hokanson JE, Zambon A, et al. The contribution of intraabdominal fat to gender differences in hepatic lipase activity and low/high density lipoprotein heterogeneity. *The Journal of clinical endocrinology and metabolism*. Jun 2001;86(6):2831-2837.
31. Magkos F, Mohammed BS, Mittendorfer B. Plasma lipid transfer enzymes in non-diabetic lean and obese men and women. *Lipids*. May 2009;44(5):459-464.

Literature-Based Genetic Risk Scores for Coronary Heart Disease; the CAREMA Prospective-Cohort Study

Anika A.M.Vaarhorst¹, Yingchang Lu^{2,3}, Bastiaan T. Heijmans¹, Martijn E.T. Dollé³, Stefan Böhringer⁴, Hein Putter⁴, Sandra Imholz³, Audrey H.H. Merry⁵, Marleen M. van Greevenbroek⁶, J. Wouter Jukema^{7,8,9}, Anton P.M. Gorgels¹⁰, Piet A. van den Brandt^{5,11}, Michael Müller², Leo J. Schouten⁵, Edith J.M. Feskens², Jolanda M.A. Boer³, P. Eline Slagboom^{1,12}

1. Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands
2. Division of Human Nutrition, Wageningen University and Research Center, Wageningen, The Netherlands
3. National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands
4. Department of Medical Statistics, Leiden University Medical Center, Leiden, The Netherlands
5. Department of Epidemiology, CAPHRI School for Public Health and Primary Care, Maastricht University, Maastricht, The Netherlands
6. Department of Internal Medicine, CARIM School for Cardiovascular Diseases, Maastricht University Medical Centre, Maastricht, The Netherlands
7. Department of Cardiology, Leiden University Medical Center, Leiden, the Netherlands,
8. Interuniversity Cardiology Institute of the Netherlands (ICIN), Utrecht, The Netherlands
9. The Durrer Center for Cardiogenetic Research, Amsterdam, The Netherlands
10. Department of Cardiology, University Hospital Maastricht, Maastricht, The Netherlands
11. Department of Epidemiology, GROW School of Oncology and Developmental Biology, Maastricht University, Maastricht, The Netherlands
12. Netherlands Consortium for Healthy Ageing, Leiden, The Netherlands

Circulation Cardiovascular Genetics 2012; 5(2):209-209

Abstract

Background Genome-wide association studies (GWAS) have identified many single nucleotide polymorphisms (SNPs) associated with coronary heart disease (CHD) or CHD risk factors (RF). Using a case-cohort study within the prospective CAREMA cohort, we tested if genetic risk scores (GRS) based on GWAS identified SNPs are associated with and predictive for future CHD.

Methods and Results 742 incident cases, i.e. participants who developed CHD during a median follow up of 12.1 years (range 0.0-16.9 years) were compared with a randomly selected sub-cohort of 2221 participants selected from the total cohort (n=21,148). We genotyped 179 SNPs previously associated with CHD or CHD RF in GWAS as published up to May 2, 2011. The allele-count GRSs, composed of all SNPs, the 153 RF SNPs or the 29 CHD SNPs, were not associated with CHD independent of CHD RF. The weighted 29 CHD SNP GRS, with weights obtained from GWAS for every SNP, was associated with CHD independent of CHD RF (HR=1.12 per weighted risk allele; 95%-CI=1.04-1.21) and improved risk reclassification with 2.8% ($p=0.031$). As an exploratory approach to achieve weighting, we performed LASSO regression analysis on all SNPs and the CHD SNPs. The CHD LASSO GRS performed equal to the weighted CHD GRS, whereas the Overall LASSO GRS performed slightly better than the weighted CHD GRS.

Conclusions A GRS composed of CHD SNPs improves risk prediction, when adjusted for the effect sizes of the SNPs. Alternatively LASSO regression analysis may be used to achieve weighting; however validation in independent populations is required.

Introduction

Coronary heart disease (CHD) is a complex disease influenced by both lifestyle and genetic factors. During past years, genome-wide association studies (GWAS) have identified multiple common single nucleotide polymorphisms (SNPs) that are robustly associated with the risk of CHD¹⁻⁵ or CHD risk factors including blood pressure^{6,7} and plasma lipid levels.^{8,9} It was hoped that GWAS would identify SNPs that were useful in predicting CHD risk. Individually, however, these SNPs have a relatively small effect size. Greater effect sizes may be expected for genetic risk scores (GRSs) that comprise the cumulative effect of individual SNPs.^{10,11}

Two previous prospective studies investigated the association of a GRS with incident CHD.^{12,13} Paynter *et al.*¹² found that an allele-count GRS based on 101 published GWAS SNPs known to be associated with cardiovascular disease and its intermediate risk factors was associated with a cardiovascular disease in US women, but not independent from risk factors used in the ATP III or Reynolds risk score. The association was somewhat stronger for an allele-count GRS based on a subset of 12 SNPs selected for their association with clinical cardiovascular end points.¹² In a recent Finnish cohort a weighted GRS based on 13 GWAS-identified, CHD-associated SNPs was independently associated with 10-year incidence of CHD.¹³

In our study, we composed three allele-count GRS on the basis of 179 SNPs that were associated with CHD or CHD risk factors in GWAS published up to May 2, 2011; that is an Overall GRS based on all 179 SNPs, a Risk Factor GRS based on 153 SNPs associated with CHD risk factors and an allele-count CHD GRS based on 29 SNPs associated with CHD. In addition a weighted CHD GRS, with the weights obtained from 2 published GWAS, was constructed. As an exploratory analysis we constructed a least absolute shrinkage and selection operator (LASSO) GRS based on all 179 SNPs and on the 29 CHD SNPs. We investigated their predictive value for CHD during a median follow-up of 12.1 years in a population-based setting.

Methods

Study population

The Cardiovascular Registry Maastricht (CAREMA) study population has been described before.¹⁴ In short, the study participants, living in the Maastricht region, were selected from two large monitoring projects in the Netherlands: the monitoring project on cardiovascular disease risk factors (PPHVZ) 1987-1991 and the monitoring project on chronic disease risk factors (MORGEN Project) 1993-1997, including the transition year (1992) between these two projects. Each year, a random sample of people aged between 20 and 59 years were selected from municipal registries of Maastricht and surrounding communities. In total 21,662 men and women, born between 1927 and 1997, were included in this study and

21,148 participants (97.6%) had given informed consent to retrieve information from the municipal population registries and from their general practitioner and specialist.

Cardiologic follow-up

The cardiologic follow-up has been described earlier.¹⁴ In short by linking the cohort to the hospital information system of the University Hospital Maastricht (UHM), 97.6% of the cohort members could be found (n=20,632). Next, these subjects were linked to the cardiology information system of the department of cardiology. For participants who died, the cause of death was obtained from Statistics Netherlands. In addition, the CAREMA cohort was linked to the hospital discharge registry of the UHM to enlarge the completeness of the cardiologic follow-up.

All participants, who were younger than 30 years of age at baseline (n=3505), who had an acute myocardial infarction (MI), unstable angina pectoris, a coronary artery bypass graft or a percutaneous transluminal coronary angioplasty before baseline (n=187), were lost to follow-up (n=17), or were included in the transition year 1992 (n=2203), were excluded from further analysis.

Case-cohort design

During the follow-up period (median, 12.1 yrs; range, 0.0-16.9 yrs), which ended on 31 December 2003, 742 participants developed incident CHD, consisting of 368 patients with incident MI, 294 patients with unstable angina pectoris and 80 patients who have died because of CHD. No blood for DNA extraction was available for 59 patients, and in 15 cases DNA extraction was unsuccessful. Twenty-two cases were excluded because no complete information on traditional risk factors (TRFs, *i.e.* total cholesterol, high density lipoprotein [HDL] cholesterol, systolic blood pressure, current smoking, body mass index [BMI], self-reported diabetes and parental history of MI) was available. Thus the available case group consists of 646 patients.

From the full eligible CAREMA cohort, randomly a sub-cohort of 2221 participants was drawn. No blood was available for 188 participants, DNA extraction was unsuccessful in 49 participants and no complete information on TRF was available for 71 participants. Thus the available sub-cohort group consisted of 1913 participants.

Determination of risk factors

At baseline, all participants filled in a questionnaire on medical history, parental history of MI, and lifestyle factors. During a medical examination at baseline, information was collected

on blood pressure, height and weight. In addition, non-fasting blood samples were collected using EDTA tubes. The blood was centrifuged for 10 minutes at 1000 rpm, fractioned into blood plasma, white blood cells fractions (buffy coats) and erythrocytes and subsequently stored at -20°C. Within three weeks, the blood plasma samples were transported to the Lipid Reference Laboratory of the University Hospital Dijkzigt (LRL) in Rotterdam where total and HDL-cholesterol levels were determined using a CHOD-PAP method.¹⁵

SNP selection

Using the GWAS Catalog¹⁶ (www.genome.gov/gwastudies; accessed at May 2, 2011), 248 SNPs were identified that showed an association with CHD or its intermediate risk factors (i.e. blood pressure, anthropomorphic traits [BMI, waist circumference and waist to hip ratio], blood lipid levels and type 2 diabetes) in at least two GWAS or in a meta-analysis at a genome-wide significance levels ($p < 5 \times 10^{-8}$). For the SNPs that were in perfect linkage disequilibrium (LD) ($D' = 1$; $R^2 = 1$) with each other, one of them was excluded ($n = 22$). For 29 SNPs it was impossible to design primers or they did not fit in the assay designs, leaving 197 SNPs for genotyping (supplement III, table S1).

DNA extraction & genotyping

DNA was extracted from the buffy coats, using a standard procedure.¹⁷ The resulting DNA pellet was dissolved in TE buffer and DNA concentrations were determined using the Nanodrop ND1000 Spectrophotometer. SNPs were genotyped using the iPLEX Gold chemistry of Sequenom's MassARRAY platform (San Diego, CA, USA). Sequenom's MassARRAY® Assay Design 3.1 Software was used for assay designs, and Sequenom's SpectroTyper 4.0 software was used to call genotypes automatically, followed by manual review. Nine SNPs failed genotyping (e.g. no PCR product or an abnormal cluster plot), 2 SNPs were removed due to a success rate $< 90\%$, and 7 SNPs were removed because they were out of Hardy Weinberg equilibrium ($p < 0.001$) calculated in the sub-cohort, leaving 179 SNPs for analysis (supplement III, table S2). Individuals were removed when their success rate was $< 80\%$, leaving 642 cases and 1905 sub-cohort members. The overall call-rate for the remaining SNPs and individuals was 98.24%. For success rate per SNP see supplement III, table S2)

Statistical analysis

To impute missing genotypes (1.76%) a multiple imputation method was used (R packages *mi* v0.08-08). Missing genotypes were imputed, where the chance of being a major homozygote, heterozygote or minor homozygote depends on the distribution of the particular SNPs. This

process was repeated 20 times; thus, 20 new datasets with nonmissing data were created. Next, all imputed datasets were analyzed as described below and the results from every imputed dataset were combined using the R package *mitools* (version 2.0). For results on the non-imputed dataset, see supplement III, table S3)

We constructed three allele-count GRSs and one weighted GRS. The first GRS comprised all 179 SNPs associated with CHD and its intermediate risk factors (Overall GRS). The rationale behind this GRS was to combine the information of all known SNPs associated with CHD or its intermediate risk factors and test its predictive value in a prospective cohort study. The second one consisted of the 153 SNPs associated with intermediate risk factors for CHD (Risk Factor GRS). This GRS was included to evaluate to what extent such SNPs could replace TRF. The third GRS consisted of SNPs previously associated with CHD (CHD GRS). The reason for including this GRS (not diluted by SNPs with only a small effect on TRF) was to test if this GRS would associate with incident CHD and possibly improving risk prediction in addition to TRFs. A similar approach was used by Paynter *et al.*,¹² although fewer SNPs associated with CHD or its intermediate risk factors were identified at that time. The fourth GRS contained the same 29 CHD SNPs, but every SNP was weighted according to their effect size (weighted CHD GRS) to allow the prediction to be based mainly on the most informative SNPs. For estimating the weights of the individual SNPs, the effect sizes were obtained from two large GWAS on CHD (see supplement III, table S4).^{4,18} A similar approach was used by Ripatti *et al.*¹³

We tested whether the four GRSs were associated with future CHD, using Cox-proportional hazards models with robust variance estimation to adjust for the case-cohort design according to the method of Prentice.¹⁹ These models were used to estimate the 10-year risk of CHD using a base model with and without each GRS. The base model consists of these TRFs; sex, current smoking, self-reported diabetes, parental history of MI, HDL cholesterol, total cholesterol, systolic blood pressure and BMI. Age in years was used as the time-scale variable, and we adjusted for delayed entry. The same analyses were also done for every SNP separately.

Because of a change in scaling (the allele-count CHD GRS theoretically ranges from 0 to 58, whereas the weighted CHD GRS ranges from 0 to 22.14), the hazard ratio of a weighted GRS cannot be directly compared with an allele-count GRS. Thus, for comparing the effects of the different GRS, we calculated the z-scores for all GRS in all individuals and performed the above analysis on these standardized GRS.

To test if adding a GRS to the base model improved risk prediction, *c*-statistics, according the non-parametric method of DeLong *et al.*²⁰ and the net reclassification improvement (NRI) as described by Pencina *et al.*²¹ were used. The NRI was not directly applicable to our

study because of non-cases (n=550) who were censored (e.g. due to mortality from causes other than CHD during the 10-year follow up period) and cases (n=208) that got an event after 10 years of follow-up.²² Therefore, cases who developed an event after 10 years of follow-up were treated as controls that left the study at 10 years of follow-up (n=208), which resulted in 434 cases and 2012 non-cases. Next, we calculated a Kaplan-Meier curve for the sub-cohort. Based on this Kaplan-Meier curve, weights were assigned to every individual. Weights of zero were assigned to non-cases who left the study before 10 years of follow-up (n=550). Using the 10-year risk estimates, subjects were divided in the following risk categories: 0-<5%, 5-<10%, 10-<20%, ≥20%, as previously applied.^{12, 13} For all analyses we used the software R (version 2.10.1, www.r-project.org) in combination with the package survival (version 2.35-8) for the Cox-proportional hazards models according to Prentice and the Kaplan-Meier curve. For estimating the c-statistics and corresponding p-values, we used the R package pROC (version 1.4.1).

GRS based on LASSO regression

Not all GWAS-identified SNPs may contribute to the risk of future CHD in our population-based prospective study.²³ Therefore, as an exploratory analysis, we selected a subset of the most informative SNPs using LASSO regression,²⁴ in combination with 10-fold cross-validation to address the issue of internal validation.²⁴⁻²⁶ Using the LASSO approach a subset of informative SNPs was extracted from the 29 CHD SNPs (CHD LASSO GRS) and the complete SNPs set (Overall LASSO GRS). Both GRS were subject to the same statistical analysis as for the previous four GRS. For the LASSO regression analysis the R package *penalized* (version 0.9-32), was used. See supplement III, Materials and methods for more detailed information about this analysis.

Results

Baseline characteristics

The TRFs are associated with future CHD in the CAREMA study population (see table 1). When all TRFs were entered in one Cox proportional hazards model, with age as the time-scale variable, BMI was no longer significantly associated with future CHD.

Table 1 | Baseline characteristics of the CAREMA case-cohort study.

Variables	Cases(n=646)	Sub-cohort (n=1913)*	HR (95% CI)†	AdjustedHR (95% CI)‡
Age, y	50.2 ± 7.1	45.2 ± 8.4	1.08 (1.07-1.09)§	-
Tot Chol, mmol/L	6.3 ± 1.2	5.6 ± 1.1	1.47 (1.35-1.61)	1.33 (1.20-1.48)
HDL-C, mmol/L	1.0 ± 0.3	1.3 ± 0.3	0.05 (0.03-0.07)	0.14 (0.08-0.22)
SBP, mm Hg	128.9 ± 17.1	120.2 ± 15.2	1.02 (1.02-1.03)	1.02 (1.01-1.03)
BMI, kg/m ²	26.6 ± 4.0	25.0 ± 3.7	1.08 (1.05-1.10)	1.02 (0.99-1.05)
Men	483 (74.8%)	870 (45.5%)	1.41 (3.31-50.3)	2.52 (1.93-3.29)
Current smoking	373 (57.7%)	790 (41.3%)	2.03 (1.68-2.46)	1.95 (1.54-2.46)
Type 2 diabetes	24 (3.7%)	17 (0.9%)	3.60 (1.83-7.08)	3.03 (1.46-6.26)
Parental history of MI	270 (41.8%)	670 (35.0%)	1.31 (1.12-1.54)	1.41 (1.16-1.72)

CAREMA indicates Cardiovascular Registry Maastricht; HR, hazard ratio; 95% CI, 95% confidence interval; Tot Chol, total cholesterol; HDL-C, high density lipoprotein cholesterol; SBP, systolic blood pressure; BMI, body mass index; MI, myocardial infarction

Results are presented as means ± SD or N (%)

*Includes 101 cases

†Univariate analysis; age in years was used as the time-scale variable

‡All the variables were added into one multivariable Cox proportional hazards models model and age was used as the time scale-variable.

§Follow-up time in years was used as the time-scale variable

|| Current smokers compared to non-current smokers

Genetic Risk Scores

An Overall GRS was constructed of all the 179 SNPs, of which a detailed overview including their association with incident CHD before and after adjustment for TRF is presented in supplement III, table S5. The Overall GRS had a mean value (SD) of 182.9 (9.2) for all participants, with a range from 142 to 217 risk alleles. As shown in table 2, this Overall GRS was associated with CHD with a hazard ratio (HR) of 1.02 per risk allele (95% confidence interval [CI], 1.01-1.03); after adjustment for TRFs the effect size further attenuated (HR/risk allele, 1.01; 95% CI, 1.00-1.02).

As shown in Table 2, the Risk Factor GRS constructed from the 153 SNPs selected for their previous association with CHD risk factors (blood pressure, anthropomorphic traits, blood

lipid levels and type 2 diabetes) was associated with future CHD (HR/risk allele, 1.02; 95% CI, 1.01-1.03), but not after adjustment for TRFs (HR/risk allele, 1.00; 95% CI, 0.99-1.02).

Table 2 | Associations between the different GRS and CHD.

Genetic Risk Scores	N SNPs	Cases (n=642)	Sub-cohort (n=1905)	Coronary heart disease		P
				HR (95% CI)	HR Std† (95% CI)	
Overall GRS	179	184.1 ± 8.8	182.5 ± 9.2	1.02 (1.01-1.03)	1.20 (1.09-1.32)	<0.001
Adjusted for TRFs*				1.01 (1.00-1.02)	1.07 (0.99-1.16)	0.084
Risk Factor GRS	153	156.5 ± 8.3	155.3 ± 8.5	1.02 (1.01-1.03)	1.16 (1.06-1.27)	0.002
Adjusted for TRFs*				1.00 (0.99-1.02)	1.03 (0.95-1.12)	0.452
CHD GRS	29	31.2 ± 3.3	30.7 ± 3.5	1.04 (1.01-1.07)	1.16 (1.04-1.26)	0.004
Adjusted for TRFs*				1.03 (0.95-1.12)	1.12 (0.85-1.49)	0.413
Weighted CHD GRS	29	11.9 ± 1.3	11.6 ± 1.4	1.15 (1.07-1.23)	1.21 (1.10-1.33)	<0.001
Adjusted for TRFs*				1.12 (1.04-1.21)	1.17 (1.06-1.30)	0.001
LASSO CHD GRS	3	1.82 ± 0.74	1.66 ± 0.75	1.38 (1.21-1.57)	1.27 (1.16-1.40)	<0.001
Adjusted for TRFs*				1.36 (1.21-1.52)	1.26 (1.16-1.37)	<0.001
LASSO GRS	14	4.0 ± 1.0	3.7 ± 1.0	1.48 (1.31-1.67)	1.49 (1.31-1.69)	<0.001
Adjusted for TRFs*				1.39 (1.26-1.53)	1.40 (1.27-1.53)	<0.001

GRS indicates genetic risk scores; CHD, coronary heart disease; SNPs, single-nucleotide polymorphisms; HR, hazard ratio; Std, standardized; TRF, traditional risk factors; LASSO, least absolute shrinkage and selection operator

*sex, current smoking (yes/no), systolic blood pressure, total cholesterol, HDL cholesterol, self-reported diabetes (yes/no), body mass index and parental history of myocardial infarction (MI) (no MI, one parent with MI, both parents with MI). Age was used as the time scale variable

†Increase in hazard ratio per standard deviation

Next, the CHD GRS constructed from the 29 SNPs selected for their previous association with CHD was investigated. The allele-count CHD GRS was associated with future CHD with a HR of 1.04 per risk allele (95% CI, 1.01-1.07). After adjustment for TRFs this GRS was no longer associated (HR, 1.03; 95% CI, 0.95-1.12). The CHD GRS that was weighted for previously reported effect sizes of the included SNPs, however, was associated with future CHD both before (HR, 1.15; 95% CI, 1.07-1.23) and after the adjustment for TRFs (HR, 1.12; 95% CI, 1.04-1.21).

C-indexes

The discriminatory capabilities of the 4 GRS without including other variables were low (C-index = 0.547 for the Overall GRS, 0.538 for the Risk Factor GRS, 0.527 for the CHD GRS, 0.550 for the weighted CHD GRS), but all statistically significant (see Table 3). As shown in

the Figure and table 3, when using a prediction model based on all TRFs (C -index=0.816), adding any of the four GRS did not improve risk discrimination.

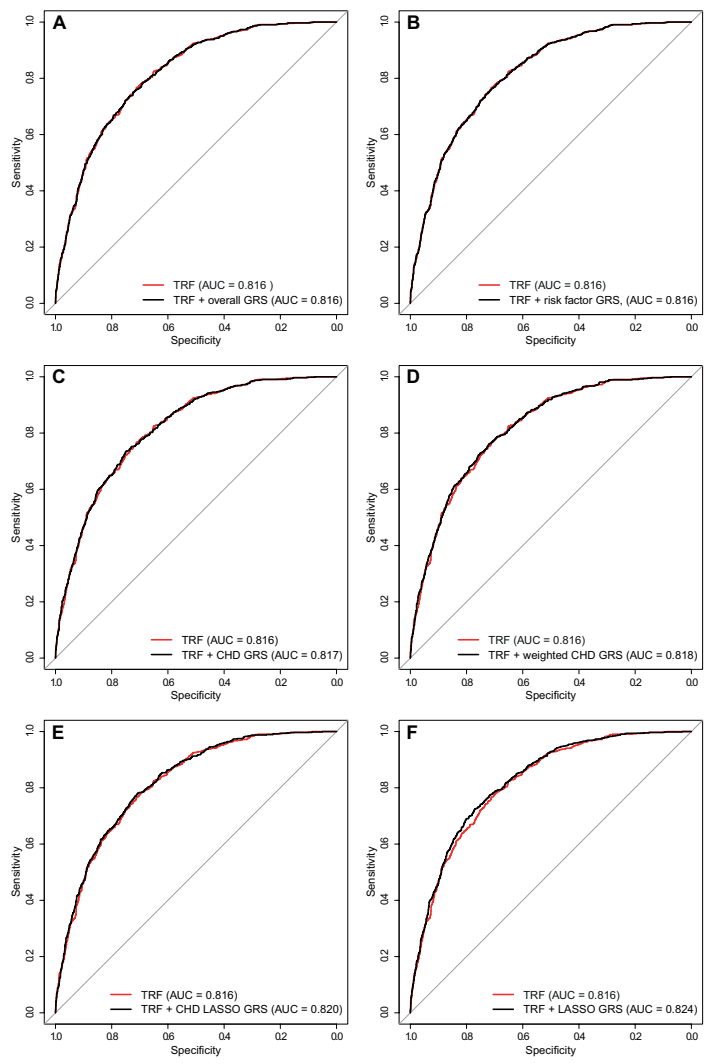


Figure | Receiver-operator characteristic curves for coronary heart disease (CHD). Curves were based on the traditional risk factors (age, sex, total cholesterol, HDL-cholesterol, systolic blood pressure, current smoking, self-reported diabetes and parental history of MI) with and without each genetic risk score (GRS). The following GRS were included: **A**, Overall GRS; **B**, Risk factor GRS; **C**, CHD GRS; **D**, weighted CHD GRS; **E**, CHD LASSO GRS; and **F**, the LASSO GRS. AUC indicates area under the curve.

Table 3 | Improvement in discrimination, as assessed with c-statistics, when the different GRSs were added to the base model (composed age, sex, current smoking plus TRF^{*}).

Model	C-index	P value
Overall GRS	0.547	<0.001
Risk Factor GRS	0.538	0.002
CHD GRS	0.527	0.021
Weighted CHD GRS	0.550	<0.001
CHD LASSO GRS	0.559	<0.001
LASSO GRS	0.599	<0.001
<i>Base model composed of TRF*</i>	0.816	
Base model and overall GRS	0.816	0.966
Base model and risk factor GRS	0.816	0.858
Base model and CHD GRS	0.817	0.417
Base model and weighted CHD GRS	0.818	0.134
Base model van CHD LASSO GRS	0.820	0.048
Base model and LASSO GRS	0.824	0.011

GRS indicates genetic risk scores; TRFs, traditional risk factors; CHD, coronary heart disease; LASSO, least absolute shrinkage and selection operator.

*Sex, current smoking, systolic blood pressure, total cholesterol, high density lipoprotein cholesterol, self-reported diabetes, body mass index and parental history of myocardial infarction. Age was used as the time scale variable

Net reclassification improvement

We found no improvement in reclassification when the Overall GRS, the Risk Factor GRS or the allele-count CHD GRS were added to the base model (Table 4, for more details see supplement III, table S6, A through C). When we used the weighted CHD GRS, 8.1% of the participants were classified into a more appropriate risk category, but for 5.3% of the population the reclassification got worse; thus, the NRI improved with 2.8% ($p=0.031$), which could be attributed to an improvement in reclassification for events (*i.e.* individuals who did have a CHD event; NRI=2.5%, $p=0.040$, see supplement III, table S6d).

GRS based on LASSO regression

As an exploratory analysis, we performed LASSO regression analysis on the 29 SNPs previously associated with CHD, extracting 3 SNPs with corresponding weights. These SNPs and their weights were used to compose a weighted GRS (see supplemental results and supplement III, table S7 for the included SNPs and their corresponding weights). This CHD LASSO GRS was associated with incident CHD (HR, 1.38; 95% CI, 1.21-1.57) and remained significantly associated after adjusting for TRFs (HR, 1.36; 95% CI, 1.21-1.52). The standardized HR of

this CHD LASSO GRS (HR/sd increase, 1.27; 95% CI, 1.16-1.40) is comparable to that of the weighted CHD GRS (HR/sd increase, 1.21; 95% CI, 1.10-1.33). Next, we performed LASSO regression analysis on all SNPs; resulting in the extraction 14 SNPs with corresponding weights (see supplement III, results and table S7 for the included SNPs and their weights). The LASSO Overall GRS based on these 14 SNPs and their weights was associated with incident CHD (HR, 1.48; 95% CI, 1.31-1.29) and remained significantly associated when adjusted for TRFs (HR, 1.39; 95% CI, 1.26-1.53). As shown in Table 2, the standardized HR of the LASSO GRS was higher than that of the weighted CHD GRS.

Table 4 | Improvement in reclassification of future CHD when using a model with different GRSs in addition to TRF* compared to a model with only TRF.

Genetic risk score		Reclassification				
		P Up	P Down	NRI	Z-score	P-value
Overall GRS	Event	0.013	0.011	0.002	0.256	0.399
	Non-event	0.006	0.009	0.003	1.020	0.154
	Total	...	...	0.005	0.604	0.273
Risk Factor GRS	Event	0.009	0.004	0.005	0.919	0.179
	Non-event	0.003	0.001	-0.001	-0.976	0.836
	Total	...	...	0.003	0.633	0.263
CHD GRS	Event	0.039	0.027	0.012	0.997	0.159
	Non-event	0.011	0.014	0.003	0.983	0.163
	Total	...	...	0.015	1.237	0.108
Weighted CHD GRS	Event	0.062	0.037	0.025	1.755	0.040
	Non-event	0.016	0.019	0.003	0.676	0.249
	Total	...	...	0.028	1.873	0.031
LASSO CHD GRS	Event	0.077	0.059	0.018	1.068	0.143
	Non-event	0.019	0.026	0.007	1.472	0.071
	Total	...	...	0.025	1.422	0.077
LASSO GRS	Event	0.142	0.083	0.059	2.721	0.003
	Non-event	0.029	0.039	0.010	1.691	0.045
	Total	...	...	0.068	3.062	0.001

CHD indicates coronary heart disease; GRS, genetic risk score; TRFs, traditional risk factors; P, proportion; NRI, net reclassification index; and LASSO, least absolute shrinkage and selection operator.

*Traditional risk factors include sex, age, total cholesterol, HDL cholesterol, systolic blood pressure, body mass index, current smoking, diabetes and a parental history of myocardial infarction.

Discussion

Over the past 5 years, the pace of identification of genetic variants underlying susceptibility to CHD has rapidly increased, leading to an interest in investigating if and how this information may be used in improving CHD risk prediction. Therefore, we evaluated the predictive “potential” of various GRS constructed on the basis of 179 SNPs associated with CHD and/or intermediate CHD risk factors in published GWAS. The weighted CHD GRS, constructed from 29 SNPs previously associated with CHD using information on effect sizes from two large GWAS on CHD,^{4,18} was associated with incident CHD, the effect being independent from age, sex, current smoking, self-reported diabetes, parental history of MI, HDL cholesterol, total cholesterol, systolic blood pressure and BMI. This weighted CHD GRS also improved risk reclassification, but not discrimination above a model using only TRFs. The three allele-count GRS (*i.e.* the Overall GRS, the Risk factors GRS and the CHD GRS) were associated with incident CHD, but not independently of TRFs; neither did they improve risk prediction above a model using only TRFs.

The lack of improvement in risk prediction may be related to the fact that most GWAS-identified SNPs have a very low effect size and only a few SNPs have a higher effect size and as a result an allele-count GRS may be diluted. We observed that weighting SNPs according to their effect sizes may overcome this problem. However, it may not always be possible to provide weights to novel identified SNPs. To overcome this limitation and as an exploratory analysis, we used LASSO regression, which can be used to achieve data selection and shrinkage at the same time. With this approach we extracted a subset of 3 SNPs from the 29 SNPs previously associated to CHD. This GRS performed equally compared to the weighted CHD GRS composed of all 29 CHD SNPs. Because it is impossible to apply weighting on the SNPs associated with intermediate CHD risk factors based on previously published studies, LASSO regression was used to select the set of best predicting SNPs from all 179 SNPs, resulting in a 14-SNP LASSO GRS. The association of this 14-SNP LASSO GRS with future CHD risk was stronger than the CHD LASSO GRS, and modestly but significantly improved both risk discrimination and reclassification. With cross-validation we have addressed the issue of internal validation.²⁷ Cross-validated LASSO regression, however is exploratory and needs to be externally validated in independent prospective cohort studies.

Our findings for the allele-count Overall GRS based on all 179 SNPs are in line with a recently published large prospective cohort study.¹² In this study, a GRS composed of 101 SNPs previously associated with cardiovascular disease (CVD) or an intermediate phenotype was not associated with CVD after adjustment for risk factors used in the ATP III or Reynolds risk score.^{12,12} This GRS also did not improve risk discrimination or reclassification.¹² The allele-count CHD GRS was not associated with CHD after adjustment for TRFs in our population,

which confirms findings in a recently published large prospective cohort study of women, where a GRS composed of 12 CVD associated SNPs was associated with CVD but not after adjusting for TRFs.¹² A weighted GRS, composed of 13 SNPs associated with CHD and sharing 8 genetic variants with our CHD GRS, was tested in a large Finnish cohort study.¹³ This GRS was associated with CVD after adjustment for TRFs, which confirms our finding for the weighted GRS. In a recently published meta-analysis of 14 case-control studies, a weighted GRS of 23 SNPs associated with CHD was constructed. This GRS was highly associated with CHD, although the outcome was not adjusted for TRFs. The currently tested weighted CHD GRS encompassing 29 SNPs was associated with incident CHD in our prospective case-cohort also after adjustment for TRFs.

From our study we can conclude the following. First, the 3 allele-count GRS composed of common SNPs previously associated with CHD and/or intermediate CHD risk factors, currently has no value in predicting incident CHD compared to measuring classical biochemical parameters and systolic blood pressure measurement at one single point in time. Second, it might be important to account for the different effect sizes of SNPs in a GRS, given the fact that the weighted CHD GRS was still associated with CHD after adjustment for TRFs and the allele-count CHD GRS was not. These results are in line with a study by Davies et al,²³ in which they show that a weighted GRS outperforms allele counting. Third, when using the weighted CHD GRS in addition to TRFs and when accounting for the fact that using a new predictor also can deteriorate reclassification, 2.8% of the participants were placed in a more appropriate risk category. Fourth, 10-fold cross-validated LASSO regression possible could be used for extracting a subset of SNPs from a larger set of SNPs, although this method must be externally validated in independent prospective cohort studies.

The SNPs that we used for constructing the GRS were detected in GWAS with large sample sizes. Our study has lower power to detect the small effect sizes, which will have contributed to our observation that most of SNPs were not significantly associated with CHD when analyzed individually. We did remove SNPs in complete LD, but not the ones in high LD, which could bias the results of the Overall GRS and the Risk factor GRS, depending on the regions that were over represented. For the allele-count and the weighted CHD GRS this is not an issue, since none of the included SNPs in the CHD GRS were in very high LD with each other. Also, for constructing the LASSO GRS, highly correlated SNPs were not a problem, since penalized logistic regression methods, such as LASSO regression, only select 1 SNP from many correlated SNPs.²⁸ We have not externally validated the LASSO analysis method in an independent population and therefore we cannot say that this exploratory method applies to other populations.

In conclusion, the weighted CHD GRS, in contrast to the allele-count CHD GRS, was associated with future CHD independent of TRFs and improves risk reclassification. We conclude that it is important to adjust for the effect sizes of the different SNPs. If information on effect sizes is not available, LASSO regression analysis may be used to estimate these effect sizes. This method must be externally validated in independent prospective cohort studies.

Clinical perspective

Genome-wide association studies (GWAS) have identified many single-nucleotide polymorphisms (SNPs) associated with coronary heart disease (CHD) or its risk factors. We tested the combined effect of those SNPs on incident CHD in a Dutch population. By summing the number of risk alleles, three genetic risk scores (GRS) were constructed (*i.e.* a GRS based on all 179 SNPs, one based on 153 risk factor-associated SNPs and one based on 29 CHD-associated SNPs). These GRS were associated with future CHD but not independent from traditional risk factors (*i.e.* age, sex, smoking, self-reported diabetes, parental history of myocardial infarction, blood lipids, blood pressure, and body mass index). Next, a weighted CHD GRS was constructed by summing the number of CHD-associated risk alleles multiplied by their effect sizes obtained from 2 GWAS. This GRS was associated with CHD independent of traditional risk factors and improved risk reclassification. Effect sizes are sometimes difficult to obtain; therefore, as an exploratory analysis, least absolute shrinkage and selection operator (LASSO) regression analysis was used to obtain weighting. Using this method on the 29 CHD SNPs, 3 SNPs and corresponding weights were selected for a GRS. This GRS performed equally to the weighted CHD GRS. When applied on all 179 SNPs, 14 SNPs and corresponding weights were selected for a GRS, which performed slightly better than the weighted CHD GRS. Thus, weighting is important, and, when weights are not available, LASSO regression might be used to construct a weighted GRS; however, this method must be tested in independent populations.

Acknowledgements

The authors wish to thank D. Kromhout for supervision of the PPHVZ project and the MORGEN-project steering committee consisting of (not mentioning co-authors of this article) H. B. Bueno de Mesquita, H. A. Smit, W.M.M Verschuren and J. C. Seidell (project leader). Furthermore the authors thank epidemiologists and field-workers of the Municipal Health Services in Maastricht, Amsterdam, and Doetinchem for their contribution to baseline data collection and those involved in the logistic management (A. Jansen and J. Steenbrink) and the data management (A. Blokstra, A. van Kessel, P. Steinberger, E. den Hoedt, I. Toxopeus,

J. van der Laan). The authors further wish to thank A. Hilton, P. Erkens, J. Bremen and B. Bleijlevens for assistance in clinical data collection. Statistics Netherlands is acknowledged for providing data on causes of death.

Source of funding

This study was funded by a grant from the Netherlands Heart foundation (2006B195) and the Netherlands Consortium for Healthy Ageing (050 60 810). The funding source had no involvement in the design and conduct of the study; or with the collection, management, analysis or interpretation of the data and reporting or the decision to submit the manuscript for publication.

The Monitoring Project on Cardiovascular Disease Risk Factors (PPHVZ) and the MORGEN study were financially supported by the Ministry of Public Health, Welfare and Sports of the Netherlands.

References

1. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447:661-678.
2. Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, Jonasdottir A, et al. A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science*. 2007;316:1491-1493.
3. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, et al. Genomewide association analysis of coronary artery disease. *N Engl J Med*. 2007;357:443-453.
4. Schunkert H, König IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet*. 2011;43:333-338.
5. Erdmann J, Grosshennig A, Braund PS, König IR, Hengstenberg C, Hall AS, et al. New susceptibility locus for coronary artery disease on chromosome 3q22.3. *Nat Genet*. 2009;41:280-282.
6. Levy D, Ehret GB, Rice K, Verwoert GC, Launer LJ, Dehghan A, et al. Genome-wide association study of blood pressure and hypertension. *Nat Genet*. 2009;41:677-687.
7. Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, Coin L, et al. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet*. 2009;41:666-676.
8. Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM, Pramstaller PP, et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat Genet*. 2009;41:47-55.
9. Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, Schadt EE, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet*. 2009;41:56-65.
10. Cornelis MC, Lu Q, Zhang C, Kraft P, Manson J, Cai T, et al. Joint Effects of Common Genetic Variants on the Risk for Type 2 Diabetes in U.S. Men and Women of European Ancestry. *Ann Intern Med*. 2009;150:541-550.
11. van Heemst D, Beekman M, Mooijaart SP, Heijmans BT, Brandt BW, Zwaan BJ, et al. Slagboom PE, Westendorp RGJ. Reduced insulin/IGF-1 signalling and human longevity. *Aging cell*. 2005;4:79-85.
12. Paynter NP, Chasman DI, Paré G, Buring JE, Cook NR, Miletich JP, et al. Association between a literature-based genetic risk score and cardiovascular events in women. *JAMA*. 2010;303:631-637.
13. Ripatti S, Tikkanen E, Orho-Melander M, Havulinna AS, Silander K, Sharma A, et al. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *Lancet*. 2010;376:1393-1400.
14. Merry AHH, Boer JM, Schouten LJ, Feskens EJM, Verschuren WMM, Gorgels APM, et al. Validity of coronary heart diseases and heart failure based on hospital discharge and mortality data in the Netherlands using the cardiovascular registry Maastricht cohort study. *Eur J Epidemiol*. 2009;24:237-247.
15. Katterman R, Jaworek D, Moller G, Assmann G, Bjorkhem I, Svensson L. Multicentre study of a new enzymatic method of cholesterol determination. *J.Clin.Chem.Clin.Biochem*. 1984;22:245-251.
16. Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*. 2009;106:9362-9367.
17. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res*. 1988; 16:1215.
18. Peden JF, Hopewell JC, Saleheen D, Chambers JC, Hager J, Soranzo N, et al. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nat Genet*. 2011;43:339-344.
19. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986;73:1-11.
20. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837-845.
21. Pencina MJ, D'Agostino Sr. RB, D'Agostino Jr. RB, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27:57-172; discussion 207-112.
22. Pencina MJ, D'Agostino Sr. RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med*. 2010;30:11-21.
23. Davies RW, Dandona S, Stewart AFR, Chen L, Ellis SG, Tang WHW, et al. Improved Prediction of Cardiovascular Disease Based on a Panel of SNPs

- Identified Through Genome Wide Association Studies. *Circ. Cardiovasc. Genet.* 2010;3:468-474
24. Goeman JJ. L1 penalized estimation in the Cox proportional hazards model. *Biom. J.* 2010; 52:70-84.
 25. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *J.R.Statist. Soc.B.* 1996;58:267-288.
 26. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat. Med.* 1997;16:385-95.
 27. Steyerberg EW, Harrell FE, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol.* 2001;54:774-781.
 28. Ayers KL, Cordell HJ. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet Epidemiol.* 2010;34:879-891.

Markers of endogenous desaturase activity and risk of coronary heart disease in the CAREMA cohort study

Yingchang Lu^{1,2}, Anika A.M. Vaarhorst³, Audrey H. H. Merry⁴, Martijn E. T. Dollé², Robert Hovenier¹, Sandra Imholz², Leo J. Schouten⁵, Bastiaan T. Heijmans³, Michael Müller¹, P. Eline Slagboom³, Piet A. van den Brandt^{4,5}, Anton P. M. Gorgels⁶, Jolanda M. A. Boer², Edith J. M. Feskens¹

1. Division of Human Nutrition, Wageningen University and Research Center, Wageningen, The Netherlands;
2. National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands
3. Department of Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands
4. Department of Epidemiology, CAPHRI School for Public Health and Primary Care, Maastricht University, Maastricht, The Netherlands
5. Department of Epidemiology, GROW School of Oncology and Developmental Biology, Maastricht University, Maastricht, The Netherlands;
6. Department of Cardiology, University Hospital Maastricht, Maastricht, the Netherlands

Plos One 2012; 7(7):e41681

Abstract

Background Intakes of n-3 polyunsaturated fatty acids (PUFAs), especially EPA (C20:5n-3) and DHA (C22:6n-3), are known to prevent fatal coronary heart disease (CHD). The effects of n-6 PUFAs including arachidonic acid (C20:4n-6), however, remain unclear. δ -5 and δ -6 desaturases are rate-limiting enzymes for synthesizing long-chain n-3 and n-6 PUFAs. C20:4n-6 to C20:3n-6 and C18:3n-6 to C18:2n-6 ratios are markers of endogenous δ -5 and δ -6 desaturase activities, but have never been studied in relation to incident CHD. Therefore, the aim of this study was to investigate the relation between these ratios as well as genotypes of FADS1 rs174547 and CHD incidence.

Methods We applied a case-cohort design within the CAREMA cohort, a large prospective study among the general Dutch population followed up for a median of 12.1 years. Fatty acid profile in plasma cholesteryl esters and FADS1 genotype at baseline were measured in a random subcohort (n=1323) and incident CHD cases (n=537). Main outcome measures were hazard ratios (HRs) of incident CHD adjusted for major CHD risk factors.

Results The AA genotype of rs174547 was associated with increased plasma levels of C20:4n-6, C20:5n-3 and C22:6n-3 and increased δ -5 and δ -6 desaturase activities, but not with CHD risk. In multivariable adjusted models, high baseline δ -5 desaturase activity was associated with reduced CHD risk (P for trend=0.02), especially among those carrying the high desaturase activity genotype (AA): HR (95% CI)=0.35 (0.15–0.81) for comparing the extreme quintiles. High plasma DHA levels were also associated with reduced CHD risk.

Conclusion In this prospective cohort study, we observed a reduced CHD risk with an increased C20:4n-6 to C20:3n-6 ratio, suggesting that δ -5 desaturase activity plays a role in CHD etiology. This should be investigated further in other independent studies.

Introduction

Polyunsaturated fatty acids (PUFAs) are generally believed to reduce coronary heart disease (CHD) risk.¹⁻⁴ Intakes of n-3 PUFAs, especially eicosapentaenoic acid (EPA, C20:5n-3) and docosahexaenoic acid (DHA, C22:6n-3) present in fish oil, are confirmed to prevent fatal CHD and sudden cardiac death in both observational studies and large-scale randomized controlled trials (RCTs).^{1,3} However, direct evidence for the preventive effect of n-3 PUFAs on non-fatal CHD was only recently observed in some, but not all, large-scale RCTs.⁵⁻⁷ The replacement of saturated fatty acids by n-6 PUFAs protected against incident CHD in a recent meta-analysis including 8 RCTs.⁸ As some of these RCTs also included n-3 PUFAs in addition to n-6 PUFAs,^{2,8} the effects specific to n-6 PUFAs, however, remain unclear.

The fatty acid profile of various biological tissues is often used as a biomarker of dietary fatty acid intake. Adipose tissue reflects the intake of past months to years, while erythrocyte membranes, and plasma or serum phospholipids or cholesteryl esters reflect the intake of several weeks.⁹⁻¹¹ However, the PUFA profile in biological tissues does not only reflect dietary intake, but is also strongly dependent on the endogenous metabolism of PUFAs.^{10,12} Therefore, PUFA biomarkers in biological tissues mirror the internal PUFA exposure that may be biologically more relevant. Several PUFAs can be endogenously synthesized by a series of alternate desaturation and elongation processes.^{12,13} The δ -5 desaturase and δ -6 desaturase are rate-limiting enzymes for synthesizing long-chain n-3 and n-6 PUFAs (Figure 1).^{12,14-16} They are encoded by the FADS1 and FADS2 genes on chromosome 11 (11q12-13.1), respectively.^{12,17} Potential functional genetic variants in these genes have been identified including rs174547,¹⁸ and confirmed in recent genome-wide association studies.¹⁹⁻²¹ They have an impact on FADS1 mRNA abundance,²²⁻²⁶ and, as a result, on desaturase activity, plasma PUFA levels, and endogenous PUFA pools.^{18-21,26-29} Since it is impractical to directly assay the enzyme activities of δ -5 and δ -6 desaturase in humans,^{12,14,15,29} especially in large-scale epidemiological studies, their activities have traditionally been estimated by using PUFA product-to-precursor ratios.^{11,21,27,28}

Although few prospective cohort studies have investigated PUFA biomarkers in relation to the incidence of CHD,³⁰ the relation with PUFA product-to-precursor ratios as markers of desaturase activities has, to the best of our knowledge, never been evaluated. In this prospective cohort study, we therefore aim to investigate whether C20:4n-6 to C20:3n-6 and C18:3n-6 to C18:2n-6 ratios, as respective markers of δ -5 and δ -6 desaturase activity, influence CHD risk.

Materials and methods

Study population

We conducted a case-cohort study within the Monitoring Project on Cardiovascular Disease Risk Factors 1987–1991,³¹ one of the two monitoring studies that were included in the Cardiovascular Registry Maastricht (CAREMA) study. The CAREMA study was described in detail before.^{32,33} In total, 12,486 men and women, born between 1927 and 1967 and living in the Maastricht area, participated in the Monitoring Project on Cardiovascular Disease Risk Factors and had given informed consent to retrieve information from the municipal population registries and from the general practitioner and specialist. The Medical Ethics Committee of the Netherlands Organization for Applied Scientific Research (TNO) approved the study protocol and all participants signed an informed consent form.

Cardiologic follow-up

The cardiologic follow-up has been described before.³² In brief, 97.6% of the CAREMA members could be found by linking the cohort to the hospital information system of University Hospital Maastricht (UHM). They were linked to the cardiology information system of the Department of Cardiology to obtain information about the occurrence of myocardial infarction (MI), unstable angina pectoris (UAP), coronary artery bypass grafting (CABG), or percutaneous transluminal coronary angioplasty surgery (PTCA). For participants who died, the cause of death was obtained from Statistics Netherlands. In addition, the CAREMA cohort was linked to the hospital discharge registry of the UHM to increase the completeness of the cardiologic follow-up. Follow-up ended on 31 December 2003 with a median follow-up of 12.1 yrs (range: 0.0–16.9 yrs).

Subcohort and incident CHD selection for case-cohort design

For the present study, participants who were younger than 30 years at baseline (n=2204), had a history of MI, UAP, CABG, or PTCA before baseline (n=118), or were lost to follow-up (n=2) were excluded. Thus, the eligible cohort consisted of 10,164 participants. All 620 participants who developed incident CHD during follow-up (315 MIs, 244 UAPs and 61 CHD deaths) were included in the case-cohort study. From the eligible cohort, 1483 participants were randomly drawn as a subcohort.³⁴ By randomly selecting a subcohort and using the specific statistics for this type of research design, the results are expected to be extrapolated to the entire cohort without the need of biomarker measurements in the entire cohort.^{11,34–36}

Risk factor determination

At baseline, all participants filled in a questionnaire on life-style characteristics, medical history, and parental history of MI. During a medical examination, information was collected on blood pressure, height, and weight. In addition, non-fasting blood samples were collected using EDTA tubes. The blood was centrifuged for 10 minutes at 1000 rpm and fractioned into blood plasma, white blood cells and erythrocytes and subsequently stored at -20°C. Within three weeks, the plasma samples were transported to the Lipid Reference Laboratory of the University Hospital Dijkzigt (LRL) in Rotterdam where the total and HDL-cholesterol levels were determined using a CHOD-PAP method.³⁷ The LRL in Rotterdam is a permanent member of the International Cholesterol Reference Method Laboratory Network.

Fatty acid determination

Fatty acids from plasma cholesteryl esters were quantified by gas-liquid chromatography between 2010 and 2011 at the Department of Human Nutrition of Wageningen University. The case and non-case samples were evenly distributed among the different batches and the assay sequence within each batch was random. The solid-phase extraction method was used to separate the cholesteryl ester fraction from total plasma lipid extracts. Fatty acid methyl esters were prepared by incubating isolated cholesteryl esters with acidified methanol. Peak retention times and area percentages of total fatty acids were identified by using known cholesteryl ester standards (mixture of FAME components from Sigma (MO) and NuChek (MN)) and analyzed with the Agilent Technologies ChemStation software (Agilent, Amstelveen, The Netherlands). For certain fatty acids, the values were too low to be reliably detected in some subjects, and “0” was assigned to their values. Interassay coefficients of variance in fatty acids in plasma cholesteryl esters were 1.68% for C16:0, 1.01% for C18:2n-6, 1.88% for C20:4n-6, and 5.02% for C22:6n-3, respectively. Fatty acid product-to-precursor ratios were calculated, i.e. C20:4n-6 to C20:3n-6 to reflect δ -5 desaturase activity, and C18:3n-6 to C18:2n-6 to reflect δ -6 desaturase activity (Figure 1). The 20 subjects with a “0” value for C20:3n-6 were not included in the analyses for the C20:4n-6 to C20:3n-6 ratio, reflecting δ -5 desaturase activity. Information on plasma fatty acids was available on 1323 subcohort members and 537 CHD cases.

DNA extraction and genotyping

DNA was extracted from the white blood cell fraction (buffy coats), using a standard procedure.³⁸ The resulting DNA pellet was dissolved in TE buffer and DNA concentrations were determined using the Nanodrop ND1000 Spectrophotometer. The single nucleotide polymorphism (SNP) of rs174547 in the FADS1 gene was selected based on its association

with blood cholesterol and triglyceride levels in a genome-wide association study.²³ This SNP is in high linkage disequilibrium ($D'=1$ and $r^2=0.8$) with several other SNPs around the FADS1 and FADS2 gene region, which have an impact on mRNA abundance of FADS1,²²⁻²⁵ desaturase activity, plasma PUFA levels, and endogenous PUFA pools.^{18-21,26-29} Rs174547 was genotyped entirely independent of case and non-case status using the iPLEX Gold chemistry of Sequenom's MassARRAY platform (San Diego, CA) at the Leiden University Medical Center. Sequenom's MassARRAYH Assay Design 3.1 Software was used for SNP assay design, and Sequenom's SpectroTyper 4.0 software was used to call genotypes automatically, followed by manual review. The total genotyping success rate was 93%. Among the subjects who were measured for plasma fatty acid levels, information on rs174547 genotype was available for 1246 subcohort members and 492 CHD cases. The genotype distribution was consistent with Hardy-Weinberg equilibrium expectations.

Statistical Analysis

Generalized linear models adjusted for age and sex were used to study the relations of rs174547 genotypes with PUFAs and PUFA ratios. Cox proportional hazards models adapted for the case-cohort design according to the Prentice's method³⁵ were used to calculate hazard ratios (HRs) as measures for relative risk.³⁶ All the major predictors satisfied the proportional hazard assumption (data not shown). We estimated hazard ratios for quintiles of fatty acids (expressed as the percentage of total fatty acids present in the chromatogram) and ratios of C20:4n-6 to C20:3n-6 and C18:3n-6 to C18:2n-6 based on subcohort distributions, and the respective lowest quintile was used as reference. The base models included age and sex. Additional models were further adjusted for covariates from the Third Report of the National Cholesterol Education Program Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (ATP III) risk score based on the Framingham cohort (current smoking, systolic blood pressure, hypertensive medication use, total and HDL cholesterol levels) with the addition of a history of diabetes.³⁹ The models were also further adjusted for the total percentage of n-3 PUFAs or n-6 PUFAs in plasma cholesteryl esters where necessary. Additional covariates studied were parental history of MI, alcohol use and physical activity. The significance of a linear trend across quintiles of fatty acids and ratios of C20:4n-6 to C20:3n-6 and C18:3n-6 to C18:2n-6 was examined by including the exposure as a continuous variable in the model. Potential interactions between continuous ratios of C20:4n-6 to C20:3n-6 and C18:3n-6 to C18:2n-6 and dichotomized rs174547 genotype (homozygous major allele carriers vs. minor allele carriers) were tested by including interaction terms into the model. Statistical significance was considered to be

met with a P value < 0.05 and all testing was 2-sided. All statistical analyses were performed with SAS version 9.1 software (SAS Institute, Cary, NC).

Results

The general characteristics of the study population by subcohort-case status are shown in Table 1. As expected, cases were older, more frequently male, had higher blood pressure and total cholesterol levels, lower HDL cholesterol levels, smoked more often, and more often reported to have diabetes and a parental history of MI.

Table 1 | Baseline characteristics of sub-cohort subjects and cases of incident coronary heart disease in the CAREMA cohort study.¹

	Subcohort (N=1323) ²	Cases (N=537)	Crude HR (95% CI) ³	Adjusted HR (95% CI) ⁴
Age (y)	45.2 ± 8.5	49.7 ± 7.3	1.07 (1.06-1.09)	1.05 (1.04-1.07)
Male sex	608 (46.0%)	392 (73.0%)	3.34 (2.69-4.15)	2.22 (1.66-2.99)
Total cholesterol (mmol/L)	5.7 ± 1.1	6.4 ± 1.2	1.71 (1.56-1.87)	1.42 (1.26-1.60)
HDL cholesterol (mmol/L)	1.2 ± 0.3	1.0 ± 0.2	0.04 (0.03-0.06)	0.09 (0.05-0.16)
Systolic blood pressure (mmHG)	119.2 ± 14.9	128.0 ± 16.9	1.03 (1.02-1.04)	1.02 (1.01-1.03)
Hypertensive medication use	67 (5.1%)	58 (10.8%)	2.34 (1.63-3.35)	1.27 (0.79-2.05)
Diabetes mellitus	13 (1.0%)	20 (3.7%)	5.33 (2.74-10.36)	2.83 (1.39-5.78)
Current smoking	551 (41.8%)	304 (56.7%)	1.81 (1.49-2.21)	1.72 (1.33-2.22)
Parental history of MI	452 (34.3%)	228 (42.5%)	1.40 (1.14-1.71)	1.51 (1.16-1.95)

¹Data are expressed as mean ± SD or n (%) unless otherwise indicated. HDL: high-density lipoprotein; MI: myocardial infarction; and HR (95% CI): hazard ratio and 95% confidence interval.

²Including 84 cases.

³Hazard ratios were calculated per unit increase in total cholesterol, HDL cholesterol, and systolic blood pressure, and for the presence of the categorical traits.

⁴All variables were added into one multivariable Cox proportional hazards model.

Carrying the minor G allele of rs174547 was associated with higher levels of substrates for desaturases (C18:2n-6, C20:3n-6 and C18:3n-3) and lower levels of products from desaturases (C18:3n-6, C20:4n-6, C20:5n-3 and C22:6n-3) in the plasma cholesteryl esters. Consequently, lower C20:4n-6 to C20:3n-6 and C18:3n-6 to C18:2n-6 ratios, as markers of δ -5 and δ -6 desaturase activity, respectively, were observed in carriers of the G allele as compared to those with the AA genotype (Table 2 and Figure 1). A high baseline C20:4n-6 to C20:3n-6 ratio was associated with reduced CHD risk (Table 3). A 30% reduction in CHD risk was observed among the participants in the second, third, fourth and fifth quintile of C20:4n-6 to C20:3n-6 ratio compared with those in the first quintile after adjustment for age, sex, systolic blood pressure, hypertensive medication use, current smoking, diabetes,

total cholesterol and high-density lipoprotein cholesterol (P for trend = 0.02). Although the statistical interaction between rs174547 and δ -5 desaturase activity was not significant (P = 0.56), the protective effect of high δ -5 desaturase activity was mainly confined to subjects with the AA genotype (Supplement IV, Table S1). In this group, the effect was stronger with a 65% risk reduction for the subjects in the fifth quintile compared with the first quintile (P for trend = 0.02). Rs174547 itself was not associated with CHD risk, the age- and sex-adjusted HR per G-allele being 0.99 (95% CI 0.84–1.16, Supplement IV, Table S2).

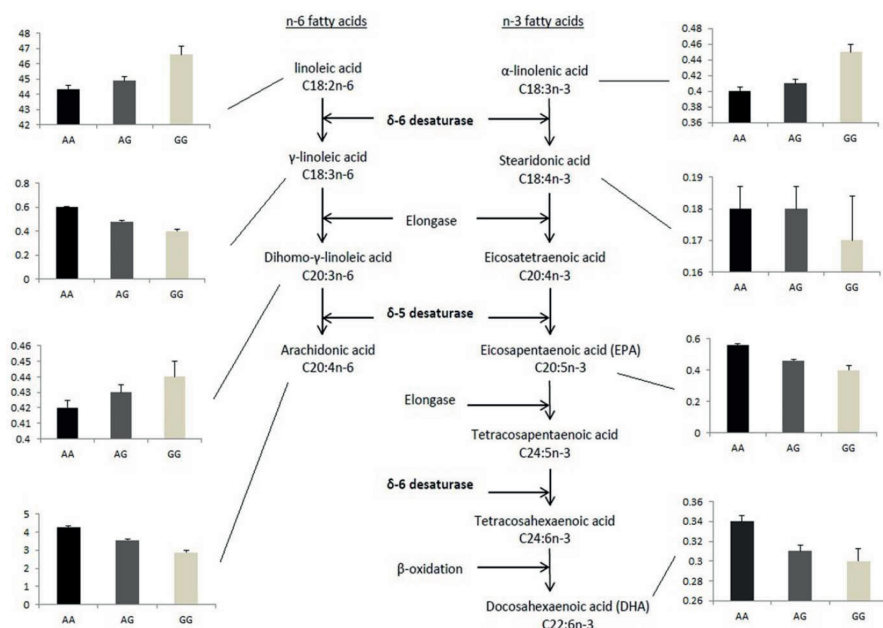


Figure 1 | Effect of genotypes of rs174547 on synthesis of PUFAs in the n-3 and n-6 pathways. Measurements of n-3 and n-6 polyunsaturated fatty acid (PUFA) levels in plasma cholesteryl esters in the sub-cohort of CAREMA study (n = 1246, Table 2). The three bars in each of the smaller plots represent levels of fatty acids (%) in individuals who carry AA, AG and GG genotypes of rs174547, respectively.

No association was observed between δ -6 desaturase activity and CHD risk (Table 3), also not after stratification by rs174547 genotype (data not shown).

The results for the four n-6 PUFAs that determine δ -5 and δ -6 desaturase activity are shown in Supplement IV, Table S3. No associations with CHD were observed for the C20 precursor (C20:3n-6) and product (C20:4n-6, arachidonic acid) of δ -5 desaturase (Figure 1), or for the C18 precursor (C18:2n-6, linoleic acid) and product (C18:3n-6) of δ -6 desaturase (Figure 1) after adjustment for age, sex, systolic blood pressure, hypertensive medication use, current smoking, diabetes, total cholesterol, and high-density lipoprotein cholesterol (P for trend > 0.16).

Regarding the n-3 PUFAs affected by desaturases, a significant inverse association was observed between C22:6n-3 (DHA) and CHD risk. This association became stronger after adjustment for plasma total and HDL cholesterol levels, and the percentages of n-6 PUFA in plasma cholesteryl esters (P for trend = 0.027, supplement IV, table S4). The proportion of plasma C20:5n-3 (EPA) was not associated with incident CHD (P for trend = 0.724, supplement IV, table S4). No association was observed between C18:3n-3 (α -linolenic acid) and CHD risk (data not shown). To explore whether there is any independent effect of C20:4n-6 to C20:3n-6 ratio on CHD beyond DHA, we additionally adjusted the models in Table 3 for percentages of DHA. The association between C20:4n-6 to C20:3n-6 ratio and CHD risk attenuated, but remained highly significant, especially among the AA carriers of rs174547 (HR = 0.44; 95% CI = 0.44:0.19–1.04 for comparing the extreme quintiles, supplement IV, table S1).

Additional adjustment for parental history of MI, alcohol use or physical activity did not materially change any of the aforementioned associations (data not shown).

Table 2 | Association of rs174547 in FADS1 with baseline PUFAs in plasma cholesteryl esters and desaturase activities in the sub-cohort (n=1246).¹

		RS174547			P-value ²
PUFA		AA (545)	AG (569)	GG (132)	
n-6 PUFA	C18:2n-6 (%)	44.30±0.27 ²	4.88±0.26	46.60±0.54	7.48x10 ⁻⁴
	C18:3n-6 (%)	0.60±0.009	0.48±0.009	0.40±0.019	6.87x10 ⁻²⁸
	C20:3n-6 (%)	0.42±0.005	0.43±0.005	0.44±0.010	0.051
	C20:4n-6 (%)	4.29±0.05	3.56±0.05	2.89±0.09	3.92x10 ⁻⁴⁶
n-3 PUFA	C18:3n-3 (%)	0.40±0.005	0.41±0.005	0.45±0.010	3.28x10 ⁻⁴
	C18:4n-3 (%) ³	0.18±0.007	0.18±0.007	0.17±0.014	0.708
	C20:5n-3 (%)	0.56±0.01	0.46±0.01	0.40±0.03	8.71x10 ⁻⁸
	C22:6n-3 (%)	0.34±0.006	0.31±0.006	0.30±0.013	0.005
Desaturase activity	δ -5 ⁴	10.65±0.09	8.59±0.09	6.86±0.19	6.40x10 ⁻⁸⁵
	δ -6 ⁴	0.014±0.0002	0.011±0.0002	0.009±0.0005	2.51x10 ⁻²⁷

¹77 subjects in the subcohort had missing values for rs174547. PUFAs: polyunsaturated fatty acids.

²General linear models were used, and all values are mean $\pm$ SEM, adjusted for age and sex.

³Only few subjects were successfully measured (AA=161, AG=185, and GG=42).

⁴ δ -5 and δ -6 desaturase activities were assessed by the ratio of C20:4n-6 to C20:3n-6 and C18:3n-6 to C18:2n-6 in plasma cholesteryl esters, respectively.

Table 3 | Association between baseline δ -5 and δ -6 desaturase activity and incident coronary heart disease (CHD).

		Quintile of δ -5 desaturase activity ¹					<i>P</i> for trend ²
		First (6.45)	Second (7.93)	Third (9.07)	Fourth (10.32)	Fifth (12.52)	
Incident CHD, n		155	117	94	93	67	
Model 1 ³	1		0.70 (0.51–0.97)	0.60 (0.42–0.83)	0.60 (0.43–0.83)	0.49 (0.34–0.70)	<0.0001
Model 2 ⁴	1		0.75 (0.54–1.06)	0.66 (0.46–0.94)	0.57 (0.39–0.82)	0.51 (0.35–0.75)	<0.0001
Model 3 ⁵	1		0.68 (0.47–0.98)	0.66 (0.45–0.96)	0.69 (0.46–1.01)	0.68 (0.45–1.02)	0.0249
Model 4 ⁶	1		0.71 (0.49–1.03)	0.70 (0.48–1.04)	0.74 (0.50–1.09)	0.77 (0.50–1.18)	0.1114
		Quintile of δ -6 desaturase activity ¹					<i>P</i> for trend ²
		First (0.0055)	Second (0.0084)	Third (0.0104)	Fourth (0.0132)	Fifth (0.019)	
Incident CHD, n		92	99	93	122	131	
Model 1 ³	1		0.99 (0.69–1.42)	0.87 (0.60–1.25)	1.09 (0.76–1.55)	1.03 (0.73–1.45)	0.606
Model 2 ⁴	1		1.03 (0.70–1.51)	0.89 (0.61–1.31)	1.07 (0.73–1.58)	0.93 (0.63–1.36)	0.627
Model 3 ⁵	1		1.07 (0.71–1.63)	0.86 (0.55–1.33)	1.11 (0.73–1.69)	0.96 (0.63–1.47)	0.867

¹ δ -5 and δ -6 desaturase activity were assessed by the ratio of C20:4n-6 to C20:3n-6 and the ratio of C18:3n-6 to C18:2n-6 in plasma cholesteryl esters, respectively and median ratios in each quintile are listed between brackets.

² From models with desaturase activity included as a continuous variable.

³ Model 1 was adjusted for age and sex.

⁴ Model 2 was adjusted for age, sex, systolic blood pressure, hypertensive medication use, current smoking, and diabetes.

⁵ Model 3 was adjusted for all covariates in model 2, total cholesterol, and high-density lipoprotein cholesterol.

⁶ Model 4 was adjusted for all covariates in model 3 and percentages of C22:6n-3 (DHA) in plasma cholesteryl esters.

Discussion

In this prospective cohort study, we observed an inverse association between C20:4n-6 to C20:3n-6 ratio, as the marker of δ -5 desaturase activity, and incident CHD risk, but no association with C18:3n-6 to C18:2n-6 ratio, as the marker of δ -6 desaturase activity. This association was partly mediated by DHA. Furthermore we confirmed associations of rs174547 in the *FADS1* gene with plasma PUFA levels and C20:4n-6 to C20:3n-6 ratio.^{18–21,27,28}

Consistent with the established cardiovascular protective effects of n-3 PUFAs,^{1,3} and especially tissue DHA,^{4,30} high DHA in plasma cholesteryl esters was associated with a reduced CHD risk. However, no association was observed between arachidonic acid or other n-6 PUFAs related to δ -5 or δ -6 desaturase activity in plasma cholesteryl esters and CHD risk.

Common genetic variants (including rs174547) in the *FADS* gene region have been associated with plasma lipid levels (total, LDL and HDL cholesterol, triglycerides, phospholipids and sphingolipids),^{19,21,23,40,41} glycemic traits (fasting glucose and beta-cell function),²⁶ and resting heart rate⁴² in recent genome-wide association studies. However, none of them have been

associated with CHD risk directly.^{40,43} This was also the case in our relatively large prospective study. In contrast, when using the estimated δ -5 desaturase activity based on the fatty acid proportion in plasma cholesteryl esters, we found a significant inverse association with incident CHD. This seems contradictory, as a strong association between rs174547 genotypes and estimated δ -5 desaturase activities was observed. However, the reduced risk was already observed with relatively low δ -5 desaturase activities (the second quintile) and remained constant over the following quintiles. Therefore, the majority of the participants with the GG genotype of rs174547 might have sufficient δ -5 desaturase activity to protect them from CHD. This might explain why no association between rs174547 genotypes and CHD risk was found. Both rs174547 genotypes and C20:4n-6 to C20:3n-6 ratio reflect endogenous δ -5 desaturase activity, but from two different perspectives. The former can be regarded as the desaturase effect conferred by a single common genetic variant in the FADS1 gene,^{20,26-29} and the latter as an approximate estimation of full desaturase activity.^{21,27,28} Their combination might provide the most accurate estimate of δ -5 desaturase activity. This might explain the stronger CHD risk reduction with high δ -5 desaturase activity in the subjects who inherited the AA genotype.

The exact biological mechanisms that link δ -5 desaturase activity with CHD risk are still not well understood. Arachidonic acid, EPA, and DHA are currently considered to be potentially involved directly in the pathogenesis of CHD through thrombotic, inflammatory, arrhythmic and/or lipid regulatory pathways.^{1,3,12,13,44-46} δ -5 Desaturase is the key enzyme synthesizing these PUFAs, while δ -6 desaturase is important at the beginning of the n-3 and n-6 PUFA synthetic pathways.^{14,15} Therefore, it is biologically plausible that CHD risk could be influenced by δ -5 desaturase activity, but not by δ -6 desaturase activity^{12,13} as was shown in our data. The non-significance of δ -6 desaturase activity on CHD risk is perhaps, also compatible with the reported normal viability and life span of δ -6 desaturase knockout mice.⁴⁷ Increased δ -5 desaturase activity might contribute to the intracellular increase of EPA and especially arachidonic acid levels.¹⁶ In non-fish eating populations, arachidonic acid is the predominant tissue very-long-chain PUFA, reaching 80% of the total very-long-chain PUFA.^{30,44} Despite the potential pro-coagulant and pro-inflammatory effects of increased exposures to arachidonic acid and its derived eicosanoid metabolites,^{2,13,44-46,48,49} there is no evidence of increased CHD risk with $\approx$ 5–7 times habitual arachidonic acid intake based on short-term small-scale controlled feeding studies.^{2,50-54} Tissue arachidonic acid levels are generally not associated with CHD risk.³⁰ This was supported by our finding based on the fatty acid profile in plasma cholesteryl esters, which suggests that arachidonic acid does not mediate the observed association between C20:4n-6 to C20:3n-6 ratio, as the marker of δ -5 desaturase activity, and CHD risk.

Increased δ -5 desaturase activity (C20:4n-6 to C20:3n-6 ratio) was associated with increased plasma levels of EPA and DHA. Our results showed that a possible protective effect of increased δ -5 desaturase activity on CHD may partly be mediated by increased endogenous exposure to DHA. The observation that increased DHA levels associated with increased δ -5 desaturase activity protect against CHD is consistent with the established cardiovascular protective effect of increased n-3 PUFA exposure (EPA and/or DHA).^{1,3} Accumulating evidence from observational studies suggests that DHA might be more protective for CHD than EPA,^{4,30} which is consistent with our findings. However, EPA and DHA are usually correlated with each other in tissues, and their potential effects cannot be easily discerned. More research on this issue is therefore warranted. In addition to blood triglyceride lowering and HDL cholesterol increasing effects of EPA and DHA, n-3 PUFAs have long been observed to have anti-thrombotic, anti-inflammatory, anti-arrhythmic, and blood pressure-lowering effects in humans even though the underlying mechanisms for these effects are incompletely understood.^{1,3,12,13,46} Interestingly, the protective effects on fatal CHD and sudden cardiac death have been shown to level off with a modest intake of EPA and/or DHA (250 mg/day), and little additional benefit was observed with higher intakes.¹ This is also consistent with our results for C20:4n-6 to C20:3n-6 ratio as the marker of δ -5 desaturase activity. Nevertheless, there might be other unidentified pleiotropic cardiovascular protective effects of increased δ -5 desaturase activity. For example, these desaturases are also significantly expressed in immune cells^{55,56} that play important roles in atherosclerotic CHD progression.

Our results should be interpreted in the context of several limitations. First, our analyses were based on a single baseline measurement of fatty acid levels in plasma cholesteryl esters that may not accurately reflect long-term fatty acid exposures. However, we did detect the established protective effect of DHA against CHD.^{1,3,4,12,13,30} Second, we estimated δ -5 and δ -6 desaturase activities based on n-6 PUFAs, while δ -5 and δ -6 desaturases are not only involved in n-6 PUFA conversion, but also in n-3 PUFA conversion. However, in comparison to n-6 PUFA conversion, the amount of n-3 PUFA conversion is relatively small,¹⁶ which should not have affected our results. Third, other potential unmeasured environmental or physiological factors could have confounded the observed associations. However, the relatively large magnitude of the protective effect of increased δ -5 desaturase activity relative to the effect of other risk factors for CHD makes confounding with either unknown risk factors unlikely. Finally, our models that included total and HDL cholesterol may have been over-adjusted, as these are probably intermediates in the metabolic pathway between desaturase and CHD risk (supplement IV, note S1).

In conclusion, in this prospective cohort study, we observed a reduced CHD risk with increased C20:4n-6 to C20:3n-6 ratio that was partly mediated by DHA. These results suggest that δ -5 desaturase activity plays a role protecting us from CHD.

Acknowledgments

The authors wish to thank D. Kromhout for supervision of the project and the project steering committee consisting of (not mentioning co-authors of this article) H. B. Bueno de Mesquita, H. A. Smit, and J. C. Seidell (project leader). Furthermore, the authors thank epidemiologists and field-workers of the Municipal Health Services in Maastricht for their contribution to baseline data collection and those involved in the logistic management (A. Jansen and J. Steenbrink) and the data management (A. Blokstra, A. van Kessel, P. Steinberger, E. den Hoedt, I. Toxopeus, J. van der Laan). The authors further wish to thank D. Jaspers, A. Hilton, V. Visser, P. Erkens, S. Philippens, J. Bremen, B. Bleijlevens, T. van Moergastel, S. van de Crommert for assistance in clinical data collection, and E. Vijge and A. Engelen for preparing the plasma samples. Statistics Netherlands is acknowledged for providing data on causes of death.

Funding

This study was supported by grant 2006B195 of the Netherlands Heart Foundation. The monitoring project on cardiovascular disease risk factors was financially supported by the Ministry of Public Health, Welfare and Sports of the Netherlands. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Mozaffarian D. Fish and n-3 fatty acids for the prevention of fatal coronary heart disease and sudden cardiac death. *Am J Clin Nutr* 2008;87:1991S-1996S.
2. Harris WS, Mozaffarian D, Rimm E, et al. Omega-6 fatty acids and risk for cardiovascular disease: a science advisory from the American Heart Association Nutrition Subcommittee of the Council on Nutrition, Physical Activity, and Metabolism; Council on Cardiovascular Nursing; and Council on Epidemiology and Prevention. *Circulation* 2009;119: 902-907.
3. De Caterina R. n-3 fatty acids in cardiovascular disease. *N Engl J Med* 2011;364:2439-2450
4. Joensen AM, Overvad K, Dethlefsen C, et al. Marine n-3 Polyunsaturated Fatty Acids in Adipose Tissue and the Risk of Acute Coronary Syndrome. *Circulation* 2011;124:1232-1238.
5. Yokoyama M, Origasa H, Matsuzaki M, et al. Effects of eicosapentaenoic acid on major coronary events in hypercholesterolaemic patients (JELIS): a randomised open-label, blinded endpoint analysis. *Lancet* 2007;369:1090-1098.
6. Kromhout D, Giltay EJ, Geleijnse JM. n-3 fatty acids and cardiovascular events after myocardial infarction. *N Engl J Med* 2010;363:2015-2026.
7. Rauch B, Schiele R, Schneider S, et al. OMEGA, a randomized, placebo-controlled trial to test the effect of highly purified omega-3 fatty acids on top of modern guideline-adjusted therapy after myocardial infarction. *Circulation* 2010;122:2152-2159.
8. Mozaffarian D, Micha R, Wallace S. Effects on coronary heart disease of increasing polyunsaturated fat in place of saturated fat: a systematic review and meta-analysis of randomized controlled trials. *PLoS medicine* 2010;7:e1000252.
9. Baylin A, Campos H. The use of fatty acid biomarkers to reflect dietary intake. *Curr Opin Lipidol* 2006;17:22-27.
10. Raatz SK, Bibus D, Thomas W, Kris-Etherton P. Total fat intake modifies plasma fatty acid composition in humans. *J Nutr* 2001;131:231-234.
11. Kroger J, Zietemann V, Enzenbach C, et al. Erythrocyte membrane phospholipid fatty acids, desaturase activity, and dietary fatty acids in relation to risk of type 2 diabetes in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam Study. *Am J Clin Nutr* 2011;93:127-142.
12. Nakamura MT, Nara TY. Structure, function, and dietary regulation of delta6, delta5, and delta9 desaturases. *Annu Rev Nutr* 2004;24:345-376.
13. Schmitz G, Ecker J. The opposing effects of n-3 and n-6 fatty acids. *Prog Lipid Res* 2008;47:147-155.
14. Cho HP, Nakamura MT, Clarke SD. Cloning, expression, and nutritional regulation of the mammalian Delta-6 desaturase. *J Biol Chem* 1999;274:471-477.
15. Cho HP, Nakamura M, Clarke SD. Cloning, expression, and fatty acid regulation of the human delta-5 desaturase. *J Biol Chem* 1999;274:37335-37339.
16. Burdge G. Alpha-linolenic acid metabolism in men and women: nutritional and biological implications. *Curr Opin Clin Nutr Metab Care* 2004;7:137-144.
17. Marquardt A, Stohr H, White K, Weber BH. cDNA cloning, genomic structure, and chromosomal localization of three members of the human fatty acid desaturase family. *Genomics* 2000;66:175-183.
18. Schaeffer L, Gohlke H, Muller M, et al. Common genetic variants of the FADS1 FADS2 gene cluster and their reconstructed haplotypes are associated with the fatty acid composition in phospholipids. *Hum Mol Genet* 2006;15:1745-1756.
19. Tanaka T, Shen J, Abecasis GR, et al. Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study. *PLoS Genet* 2009;5:e1000338.
20. Lemaitre RN, Tanaka T, Tang W, et al. Genetic Loci Associated with Plasma Phospholipid n-3 Fatty Acids: A Meta-Analysis of Genome-Wide Association Studies from the CHARGE Consortium. *PLoS genetics* 2011;7:e1002193.
21. Suhre K, Shin SY, Petersen AK, et al. Human metabolic individuality in biomedical and pharmaceutical research. *Nature* 2011;477:54-60.
22. Dixon AL, Liang L, Moffatt MF, et al. A genome-wide association study of global gene expression. *Nat Genet* 2007;39:1202-1207.
23. Kathiresan S, Willer CJ, Peloso GM, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet* 2009;41:56-65.
24. Schadt EE, Molony C, Chudin E, et al. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* 2008;6:e107.
25. Plaisier CL, Horvath S, Huertas-Vazquez A, et al. A systems genetics approach implicates USF1, FADS3, and other causal candidate genes for

- familial combined hyperlipidemia. *PLoS Genet* 2009;5:e1000642.
26. Dupuis J, Langenberg C, Prokopenko I, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* 2010;42:105-116.
27. Bokor S, Dumont J, Spinneker A, et al. Single nucleotide polymorphisms in the FADS gene cluster are associated with delta-5 and delta-6 desaturase activities estimated by serum fatty acid ratios. *J Lipid Res* 2010;51:2325-2333.
28. Mathias RA, Vergara C, Gao L, et al. FADS genetic variants and omega-6 polyunsaturated fatty acid metabolism in a homogeneous island population. *J Lipid Res* 2010;51:2766-2774.
29. Lu Y, Feskens EJ, Dolle ME, et al. Dietary n-3 and n-6 polyunsaturated fatty acid intake interacts with FADS1 genetic variation to affect total and HDL-cholesterol concentrations in the Doetinchem Cohort Study. *Am J Clin Nutr* 2010;92:258-265.
30. Harris WS, Poston WC, Haddock CK. Tissue n-3 and n-6 fatty acids and risk for coronary heart disease events. *Atherosclerosis* 2007;193:1-10.
31. Verschuren W, Leer E, Blokstra A, et al. Cardiovascular disease risk factors in The Netherlands. *Neth J Cardiol* 1993;6:205-210.
32. Merry AH, Boer JM, Schouten LJ, et al. Validity of coronary heart diseases and heart failure based on hospital discharge and mortality data in the Netherlands using the cardiovascular registry Maastricht cohort study. *Eur J Epidemiol* 2009;24:237-247.
33. Vaarhorst AA, Lu Y, Heijmans BT, et al. Literature-Based Genetic Risk Scores for Coronary Heart Disease; The CAREMA Prospective-Cohort Study. *Circulation Cardiovascular genetics* 2012;5:202-209.
34. Cai J, Zeng D. Sample size/power calculation for case-cohort studies. *Biometrics* 2004;60:1015-1024.
35. Barlow WE, Ichikawa L, Rosner D, Izumi S. Analysis of case-cohort designs. *J Clin Epidemiol* 1999;52:1165-1172.
36. Langholz B, Jiao J. Computational methods for case-cohort studies. *Computational Statistics & Data Analysis* 2007;51:3737-3748.
37. Kattermann R, Jaworek D, Moller G, et al. Multicentre study of a new enzymatic method of cholesterol determination. *Journal of clinical chemistry and clinical biochemistry Zeitschrift fur klinische Chemie und klinische Biochemie* 1984;22:245-251.
38. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* 1988;16:1215.
39. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation* 2002;106:3143-3421.
40. Teslovich TM, Musunuru K, Smith AV, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 2010;466:707-713.
41. Demirkan A, van Duijn CM, Ugocsai P, et al. Genome-Wide Association Study Identifies Novel Loci Associated with Circulating Phospho- and Sphingolipid Concentrations. *PLoS genetics* 2012;8:e1002490.
42. Eijgelsheim M, Newton-Cheh C, Sotoodehnia N, et al. Genome-wide association analysis identifies multiple loci related to resting heart rate. *Human molecular genetics* 2010;19:3885-3894.
43. Schunkert H, König IR, Kathiresan S, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet* 2011;43:333-338.
44. Calder PC. Dietary arachidonic acid: harmful, harmless or helpful? *Br J Nutr* 2007;98:451-453.
45. Calder PC. The American Heart Association advisory on n-6 fatty acids: evidence based or biased evidence? *Br J Nutr* 2010;104:1575-1576.
46. Serhan CN, Chiang N, Van Dyke TE. Resolving inflammation: dual anti-inflammatory and pro-resolution lipid mediators. *Nature reviews Immunology* 2008;8:349-361.
47. Stoffel W, Holz B, Jenke B, et al. Delta6- desaturase (FADS2) deficiency unveils the role of omega3- and omega6-polyunsaturated fatty acids. *EMBO J* 2008;27:2281-2292.
48. Seyberth HW, Oelz O, Kennedy T, et al. Increased arachidonate in lipids after administration to man: effects on prostaglandin biosynthesis. *Clinical pharmacology and therapeutics* 1975;18:521-529.
49. Simopoulos AP. Evolutionary aspects of diet, the omega-6/omega-3 ratio and genetic variation: nutritional implications for chronic diseases. *Biomed Pharmacother* 2006;60: 502-507.
50. Kusumoto A, Ishikura Y, Kawashima H, et al. Effects of arachidonate-enriched triacylglycerol supplementation on serum fatty acids and plate-

- let aggregation in healthy male subjects with a fish diet. *Br J Nutr* 2007;98:626-635.
51. Nelson GJ, Schmidt PC, Bartolini G, Kelley DS, Kyle D. The effect of dietary arachidonic acid on platelet function, platelet fatty acid composition, and blood coagulation in humans. *Lipids* 1997;32:421-425.
 52. Nelson GJ, Schmidt PC, Bartolini G, et al. The effect of dietary arachidonic acid on plasma lipoprotein distributions, apoproteins, blood lipid levels, and tissue fatty acid composition in humans. *Lipids* 1997;32:427-433.
 53. Ferretti A, Nelson GJ, Schmidt PC, et al. Increased dietary arachidonic acid enhances the synthesis of vasoactive eicosanoids in humans. *Lipids* 1997;32:435-439.
 54. Kelley DS, Taylor PC, Nelson GJ, et al. Effects of dietary arachidonic acid on human immune response. *Lipids* 1997;32:449-456.
 55. Biogps website. Available: <http://biogps.org/#got-o=genereport&id=3992>. Accessed 2011 August 15.
 56. Fairfax BP, Makino S, Radhakrishnan J, et al. (2012) Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat Genet* 2012;44:502-10.

7

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).^{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 ¹H-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 ¹H-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.

Main findings

In **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 ¹H-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 ¹H-NMR spectra representing 36 different low-molecular-weight metabolites. To select the ¹H-NMR signals associated with incident CHD, Least Absolute Shrinkage and Selection Operator (LASSO) regression was applied.^{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.^{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-anhydrosorbitol, 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

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.⁵ For scores based on traditional CHD risk factors, like the Framingham risk score, the *C*-index ranges between 0.71 to 0.84.⁶ 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 ¹H-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 ¹H-NMR spectra into discrete sections, and uses the sum of the spectral intensity in each section.⁷ This process is also known as equidistant binning.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.⁸ The ¹H-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 ¹H-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 ¹H-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 ¹H-NMR spectroscopy,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.¹⁰ The offspring therefore represent individuals with a better metabolic health as compared with controls of comparable age, sex and BMI representing the general population.¹⁰ We investigated whether lipoprotein particle sizes obtained by ¹H-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.

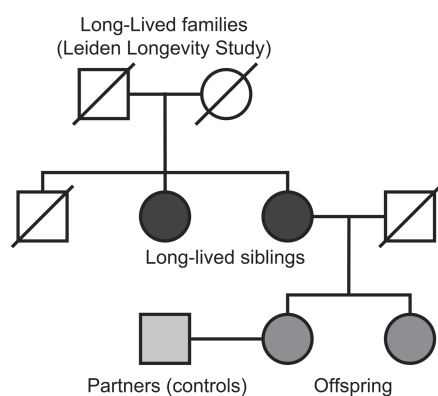


Figure 1 | Schematic representation of the Leiden Longevity Study.

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.^{11,12} 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-bonds.²¹ Mono unsaturated fatty acids (MUFAs) have one double bond and polyunsaturated fatty acids (PUFAs) are characterized by multiple double bonds.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, where 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 where 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).

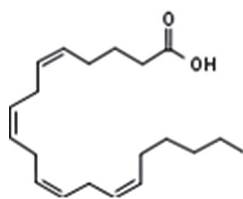


Figure | The chemical structure of arachidonic acid (C20:4 ω -6). This fatty acid consists of a carboxylic acid with a long aliphatic tail. There are 20 carbon atoms present and 4 double bounds (C20:4). The first double bound is located at the sixth carbon from the CH₃ end of the fatty acid (ω -6).

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

could not improve risk prediction beyond traditional CHD risk factors. However the 100 ^1H -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 ^1H -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 ^1H -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 ^1H -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 ^1H -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 ^1H -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 ^1H -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.³¹ 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.³²⁻³⁸ In general, when using a GRS based on SNPs associated with CHD risk factors, no improvement in CHD risk prediction was observed,³⁵⁻³⁷ 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,³⁶ whereas three other studies showed that a GRS based on CHD associated SNPs improved both risk discrimination and/or reclassification.^{32,34,37} 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,³⁴ 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.³⁷ 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.³¹

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.*³⁹ 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.*, where it was found that 12 CHD associated SNPs combined could explain 25.15% of

the variance in heritability of coronary artery disease.⁴⁰ This means that SNPs discovered in GWAS so far only explain a small proportion the heritability found in CHD.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.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.⁴³ Recently studies have been published that favour the infinitesimal model.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.⁴⁴ Other studies also found that common SNPs explain a significant proportion of the heritability of traits like height,^{45,46} BMI,⁴⁶ rheumatoid arthritis,⁴⁷ and other traits.⁴⁸ These results might favour a polygene approach for disease risk prediction, which already has been proved to work for schizophrenia.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).³⁷ These SNPs were combined into a weighted polygenic GRS.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.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.*⁵⁰ In this study a comprehensive lipid profile, consisting of 128 different lipid species, was obtained in the participants of the LLS.⁵⁰ In males no differences between offspring from long-lived siblings and controls were found.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.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,⁵¹ higher insulin sensitivity,⁵¹ lower free T3 levels,⁵² lower vitamin D levels,⁵³ and a lower RPTOR gene expression in blood.⁵⁴ Since offspring of long-lived siblings are characterized by a lower prevalence of MI and hypertension,¹⁰ 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.⁵⁵ 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,¹⁹ 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.⁵⁶ 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.⁵⁶ Individually these SNPs were not significantly associated with heart failure, but when combined into a risk score these SNPs predict heart failure.⁵⁶ 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, where only a minority of the included SNPs are truly associated with the disease or phenotype of interest.⁴⁹

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.³⁷ 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 ^1H -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 be done in females and males separately.

In a study conducted by Pencina *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 life, 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.⁶¹ This pill contains a statin to lower cholesterol levels, three blood pressure lowering drugs, folic acid and aspirin.⁶¹ It was estimated that such a pill would prevent 88% of heart attacks and 80% of strokes after two years of treatment.⁶¹ Several randomized clinical trials have investigated the efficacy of this polypill and were summarized in a meta-analysis.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 *et al.*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.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 *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%.⁶³ Thus the combination of screening for cardiovascular risk with prescribing a polypill if this risk exceeds 5% might be an effective strategy.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).⁵⁹

References

1. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. *Circulation*. May 12 1998;97(18):1837-1847.
2. National Cholesterol Education Program Expert Panel on Detection E, Treatment of High Blood Cholesterol in A. Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) final report. *Circulation*. Dec 17 2002;106(25):3143-3421.
3. Tibshirani R. The lasso method for variable selection in the Cox model. *Statistics in medicine*. Feb 28 1997;16(4):385-395.
4. Tibshirani R. Regression shrinkage and selection via the Lasso. *J Roy Stat Soc B Met*. 1996;58(1):267-288.
5. Harrell FE, Jr., Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA : the journal of the American Medical Association*. May 14 1982;247(18):2543-2546.
6. Conroy R. Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *European heart journal*. 2003;24(11):987-1003.
7. De Meyer T, Sinnaeve D, Van Gasse B, et al. Evaluation of standard and advanced preprocessing methods for the univariate analysis of blood serum 1H-NMR spectra. *Analytical and bioanalytical chemistry*. Oct 2010;398(4):1781-1790.
8. Jacob D, Deborde C, Moing A. An efficient spectra processing method for metabolite identification from 1H-NMR metabolomics data. *Analytical and bioanalytical chemistry*. Jun 2013;405(15):5049-5061.
9. Heijmans BT, Beekman M, Houwing-Duistermaat JJ, et al. Lipoprotein particle profiles mark familial and sporadic human longevity. *PLoS medicine*. Dec 2006;3(12):e495.
10. Westendorp RG, van Heemst D, Rozing MP, et al. Nonagenarian siblings and their offspring display lower risk of mortality and morbidity than sporadic nonagenarians: The Leiden Longevity Study. *Journal of the American Geriatrics Society*. Sep 2009;57(9):1634-1637.
11. Schunkert H, König IR, Kathiresan S, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics*. Apr 2011;43(4):333-338.
12. Coronary Artery Disease Genetics C. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nature genetics*. Apr 2011;43(4):339-344.
13. Kroger J, Zietemann V, Enzenbach C, et al. Erythrocyte membrane phospholipid fatty acids, desaturase activity, and dietary fatty acids in relation to risk of type 2 diabetes in the European Prospective Investigation into Cancer and Nutrition (EPIC)-Potsdam Study. *The American journal of clinical nutrition*. Jan 2011;93(1):127-142.
14. Bokor S, Dumont J, Spinneker A, et al. Single nucleotide polymorphisms in the FADS gene cluster are associated with delta-5 and delta-6 desaturase activities estimated by serum fatty acid ratios. *Journal of lipid research*. Aug 2010;51(8):2325-2333.
15. Suhre K, Shin SY, Petersen AK, et al. Human metabolic individuality in biomedical and pharmaceutical research. *Nature*. Sep 1 2011;477(7362):54-60.
16. Myocardial Infarction Genetics C, Kathiresan S, Voight BF, et al. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nature genetics*. Mar 2009;41(3):334-341.
17. Teslovich TM, Musunuru K, Smith AV, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*. Aug 5 2010;466(7307):707-713.
18. Sabatti C, Service SK, Hartikainen AL, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics*. Jan 2009;41(1):35-46.
19. Kettunen J, Tukiainen T, Sarin AP, et al. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature genetics*. Mar 2012;44(3):269-276.
20. Eijgelsheim M, Newton-Cheh C, Sotoodehnia N, et al. Genome-wide association analysis identifies multiple loci related to resting heart rate. *Human molecular genetics*. Oct 1 2010;19(19):3885-3894.
21. Wikipedia. Fatty acid. 2014; http://en.wikipedia.org/wiki/Fatty_acid, 2014.
22. Psychogios N, Hau DD, Peng J, et al. The human serum metabolome. *PLoS one*. 2011;6(2):e16957.
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

- cardiovascular disease in women. *Circulation*. Feb 24 2009;119(7):931-939.
24. Wurtz P, Raiko JR, Magnussen CG, et al. High-throughput quantification of circulating metabolites improves prediction of subclinical atherosclerosis. *European heart journal*. Sep 2012;33(18):2307-2316.
 25. Shah SH, Bain JR, Muehlbauer MJ, et al. Association of a peripheral blood metabolic profile with coronary artery disease and risk of subsequent cardiovascular events. *Circulation. Cardiovascular genetics*. Apr 2010;3(2):207-214.
 26. Wang Z, Klipfell E, Bennett BJ, et al. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature*. Apr 7 2011;472(7341):57-63.
 27. Rubtsov DV, Waterman C, Currie RA, et al. Application of a Bayesian deconvolution approach for high-resolution (1)H NMR spectra to assessing the metabolic effects of acute phenobarbital exposure in liver tissue. *Analytical chemistry*. Jun 1 2010;82(11):4479-4485.
 28. Astle W, De Iorio M, Richardson S, Stephens D, Ebbels T. A Bayesian Model of NMR Spectra for the Deconvolution and Quantification of Metabolites in Complex Biological Mixtures. *Journal of the American Statistical Association*. 2012;107(500):1259-1271.
 29. Yeboah J, McClelland RL, Polonsky TS, et al. Comparison of novel risk markers for improvement in cardiovascular risk assessment in intermediate-risk individuals. *JAMA : the journal of the American Medical Association*. Aug 22 2012;308(8):788-795.
 30. Hlatky MA, Greenland P, Arnett DK, et al. Criteria for evaluation of novel markers of cardiovascular risk: a scientific statement from the American Heart Association. *Circulation*. May 5 2009;119(17):2408-2416.
 31. Emerging Risk Factors Collaboration, Kaptoge S, Di Angelantonio E, et al. C-reactive protein, fibrinogen, and cardiovascular disease prediction. *The New England journal of medicine*. Oct 4 2012;367(14):1310-1320.
 32. Brautbar A, Pompeii LA, Dehghan A, et al. A genetic risk score based on direct associations with coronary heart disease improves coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC), but not in the Rotterdam and Framingham Offspring, Studies. *Atherosclerosis*. Aug 2012;223(2):421-426.
 33. Anderson JL, Horne BD, Camp NJ, et al. Joint effects of common genetic variants from multiple genes and pathways on the risk of premature coronary artery disease. *American heart journal*. Aug 2010;160(2):250-256 e253.
 34. Hughes MF, Saarela O, Stritzke J, et al. Genetic markers enhance coronary risk prediction in men: the MORGAM prospective cohorts. *PLoS one*. 2012;7(7):e40922.
 35. Aulchenko YS, Ripatti S, Lindqvist I, et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nature genetics*. Jan 2009;41(1):47-55.
 36. Paynter NP, Chasman DI, Pare G, et al. Association between a literature-based genetic risk score and cardiovascular events in women. *JAMA : the journal of the American Medical Association*. Feb 17 2010;303(7):631-637.
 37. Ganna A, Magnusson PK, Pedersen NL, et al. Multilocus genetic risk scores for coronary heart disease prediction. *Arteriosclerosis, thrombosis, and vascular biology*. Sep 2013;33(9):2267-2272.
 38. Tikkanen E, Havulinna AS, Palotie A, Salomaa V, Ripatti S. Genetic risk prediction and a 2-stage risk screening strategy for coronary heart disease. *Arteriosclerosis, thrombosis, and vascular biology*. Sep 2013;33(9):2261-2266.
 39. Do CB, Hinds DA, Francke U, Eriksson N. Comparison of Family History and SNPs for Predicting Risk of Complex Disease. *PLoS genetics*. Oct 2012;8(10).
 40. So HC, Gui AH, Cherny SS, Sham PC. Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. *Genetic epidemiology*. Jul 2011;35(5):310-317.
 41. Wienke A, Holm NV, Skytthe A, Yashin AI. The heritability of mortality due to heart diseases: a correlated frailty model applied to Danish twins. *Twin research : the official journal of the International Society for Twin Studies*. Aug 2001;4(4):266-274.
 42. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. *Nature reviews. Genetics*. Apr 2008;9(4):255-266.
 43. Gibson G. Rare and common variants: twenty arguments. *Nature reviews. Genetics*. Feb 2011;13(2):135-145.
 44. Morrison AC, Voorman A, Johnson AD, et al. Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nature genetics*. Aug 2013;45(8):899-901.
 45. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of the herita-

- bility for human height. *Nature genetics*. Jul 2010;42(7):565-569.
46. Yang J, Manolio TA, Pasquale LR, et al. Genome partitioning of genetic variation for complex traits using common SNPs. *Nature genetics*. Jun 2011;43(6):519-525.
47. Stahl EA, Wegmann D, Trynka G, et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature genetics*. May 2012;44(5):483-489.
48. Yang J, Lee T, Kim J, et al. Ubiquitous polygenicity of human complex traits: genome-wide analysis of 49 traits in Koreans. *PLoS genetics*. 2013;9(3):e1003355.
49. International Schizophrenia C, Purcell SM, Wray NR, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. Aug 6 2009;460(7256):748-752.
50. Gonzalez-Covarrubias V, Beekman M, Uh HW, et al. Lipidomics of familial longevity. *Aging cell*. Jun 2013;12(3):426-434.
51. Rozing MP, Westendorp RG, Frolich M, et al. Human insulin/IGF-1 and familial longevity at middle age. *Aging*. Aug 2009;1(8):714-722.
52. Rozing MP, Westendorp RG, de Craen AJ, et al. Low serum free triiodothyronine levels mark familial longevity: the Leiden Longevity Study. *The journals of gerontology. Series A, Biological sciences and medical sciences*. Apr 2010;65(4):365-368.
53. Noordam R, de Craen AJ, Pedram P, et al. Levels of 25-hydroxyvitamin D in familial longevity: the Leiden Longevity Study. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*. Dec 11 2012;184(18):E963-968.
54. Passtoors WM, Beekman M, Deelen J, et al. Gene expression analysis of mTOR pathway: association with human longevity. *Aging cell*. Feb 2013;12(1):24-31.
55. Kolovou G, Kolovou V, Vasiliadis I, Wierzbicki AS, Mikhailidis DP. Ideal lipid profile and genes for an extended life span. *Current opinion in cardiology*. Jul 2011;26(4):348-355.
56. Yu B, Zheng Y, Alexander D, et al. Genome-Wide Association Study of a Heart Failure Related Metabolomic Profile Among African Americans in the Atherosclerosis Risk in Communities (ARIC) Study. *Genetic epidemiology*. Aug 11 2013.
57. Poortvliet RK, de Ruijter W, de Craen AJ, et al. Blood pressure trends and mortality: the Leiden 85-plus Study. *Journal of hypertension*. Jan 2013;31(1):63-70.
58. Sabayan B, van Vliet P, de Ruijter W, Gussekloo J, de Craen AJ, Westendorp RG. High blood pressure, physical and cognitive function, and risk of stroke in the oldest old: the Leiden 85-plus Study. *Stroke; a journal of cerebral circulation*. Jan 2013;44(1):15-20.
59. Pencina MJ, D'Agostino RB, Sr., Larson MG, Massaro JM, Vasan RS. Predicting the 30-year risk of cardiovascular disease: the framingham heart study. *Circulation*. Jun 23 2009;119(24):3078-3084.
60. Merry AH, Boer JM, Schouten LJ, et al. Validity of coronary heart diseases and heart failure based on hospital discharge and mortality data in the Netherlands using the cardiovascular registry Maastricht cohort study. *European journal of epidemiology*. 2009;24(5):237-247.
61. Wald NJ, Law MR. A strategy to reduce cardiovascular disease by more than 80%. *Bmj*. Jun 28 2003;326(7404):1419.
62. Elley CR, Gupta AK, Webster R, et al. The efficacy and tolerability of 'polypills': meta-analysis of randomised controlled trials. *PLoS one*. 2012;7(12):e52145.
63. van Gils PF, Over EA, Hamberg-van Reenen HH, et al. The polypill in the primary prevention of cardiovascular disease: cost-effectiveness in the Dutch population. *BMJ open*. 2011;1(2):e000363.

8

Nederlandse samenvatting

Introductie

Een hartinfarct ontstaat door atherosclerose in de kransslagaders. Door de ophoping van vetachtige stoffen in de vaatwanden worden atherosclerotische plaques gevormd die leiden tot vernauwing van de kransslagaders. Deze atherosclerotische plaques zijn kwetsbaar en kunnen openscheuren. Zodra dit gebeurt, vormt er zich een stolsel en wordt de kransslagader geblokkeerd. Hierdoor kan het achterliggende hartweefsel niet meer worden voorzien van zuurstof en voedingsstoffen. Dit leidt tot pijn op de borst die kan uitstralen naar de nek, de rug of de armen. Er is dan sprake van een acuut hartinfarct. Als deze blokkade tijdelijk is en niet leidt tot versterving van het hartweefsel dan is er sprake van een instabiele angina pectoris. Soms is het onderscheid tussen een acuut hartinfarct en een instabiele angina pectoris moeilijk te maken, beide aandoeningen worden daarom ingedeeld onder de noemer coronaire hartziekten.

De kans op het krijgen van coronaire hartziekten neemt toe met de leeftijd, het hebben van een hoge bloeddruk, een hoog cholesterol gehalte, het hebben van overgewicht en een familie geschiedenis van coronaire hartziekten. Daarnaast lopen ook rokers en mannen een groter risico op het krijgen van coronaire hartziekten. Op basis van deze zogenaamde “traditionele risicofactoren” is het mogelijk om mensen met een hoog risico op coronaire hartziekten te identificeren. Deze mensen krijgen een leefstijl advies, en/of worden behandeld met medicatie tegen een hoge bloeddruk of een te hoog cholesterol om zodoende coronaire hartziekten te voorkomen. Het doel van dit proefschrift is om na te gaan of deze risicoschatting verbeterd kan worden door het gebruik van nieuwe technieken waarmee patronen van veel metabolieten in het bloed kunnen worden gemeten. Daarnaast wordt er ook bekeken of genetische informatie bij de risicobepaling betrokken kan worden. Metabolieten zijn de tussen- en eindproducten die ontstaan nadat het menselijk lichaam een bepaalde stof heeft verwerkt. Voorbeelden zijn glucose, aminozuren, antioxidanten, vrije vetzuren. Samen geven deze metabolieten een idee van de metabole staat van het lichaam op dat moment. De term metabolomics duidt op het identificeren en het kwantificeren van veel metabolieten tegelijkertijd in lichaamsvloeistoffen hetgeen in dit proefschrift is gedaan door middel van één soort ‘metabolomics’ techniek. Om te kijken of informatie over erfelijke achtergrond de risicoschattingen kan verbeteren maken wij gebruik bekende veel voorkomende variaties tussen mensen in het DNA. We richten ons op de variaties van één enkele nucleotide (in het Engels: single nucleotide polymorphisms [SNPs]) die zijn geassocieerd met coronaire hartziekten en met risicofactoren voor deze ziekten. Daarnaast hebben wij ook bekeken of een uitgebreid lipiden profiel inzicht kan geven in factoren die een markering zijn voor metabole gezondheid.

Studie populaties

Om te onderzoeken of metaboliëten en genetische variaties bijdragen aan een betere voorspelling voor coronaire hartziekten is er gebruikt gemaakt van de Cardiovasculaire Registratie Maastricht (CAREMA) studie. Tussen 1987 en 1997 is er aan mensen, woonachtig in Maastricht en omstreken, gevraagd of ze wilden deelnemen aan deze studie. In totaal waren er 21.148 deelnemers die toestemming hadden gegeven om mee te werken. Van deze mensen waren er 14.720 geschikt voor deelname omdat ze tussen de 30 en 59 jaar oud waren en nog nooit eerder coronaire hartziekten hadden gehad. Deze mensen zijn tot en met 31 december 2003 gevolgd. Tijdens die periode waren er 742 personen die coronaire hartziekten ontwikkelden of hieraan zijn overleden. Dit zijn de “patiënten”. Vanuit het CAREMA cohort werd er een aselechte steekproef getrokken van 2221 deelnemers. Deze groep fungeerden als “controles”. Het belangrijkste voordeel van hiervan is dat de metingen alleen in de patiënten en de subcohort leden uitgevoerd moeten worden, en niet in alle 14.720 deelnemers.

Om te kijken of metaboliëten waren geassocieerd met bestaande gevallen van coronaire hartziekten is er gebruik gemaakt van de Erasmus Rucphen Family (ERF) studie. Deelnemers werden geclassificeerd als patiënt wanneer ze hadden vermeld ooit een acuut hartinfarct te hebben gehad, als ze ooit een bypass operatie hadden ondergaan, of als ze angina symptomen hadden gerapporteerd tijdens het interview met de arts. Van elke deelnemer werd ook een hartfilmpje gemaakt. Als hieruit bleek dat ze ooit een acuut myocard infarct hadden gehad werden ze ook ingedeeld bij de patiënten. In deze studie zaten 2415 deelnemers, waarvan 170 patiënten, 2157 deelnemers werden als controles gebruikt.

Om te beantwoorden welke factoren zijn geassocieerd met metabole gezondheid hebben wij gebruik gemaakt van de Leiden Langleven Studie. Aan deze studie doen 450 families mee, die door erfelijke en andere familiale factoren beschermd zijn tegen coronaire hartziekten, hoge bloeddruk en diabetes type 2. Daarnaast leven deze families al enkele generaties langer dan andere mensen in de bevolking. De 450 langlevende families werden uit de Nederlandse bevolking geselecteerd als er in een familie minimaal twee broer/zusters van minimaal 90 jaar oud wilden meewerken met het onderzoek. Omdat deze oude mensen niet meer vergeleken konden worden met hun inmiddels overleden generatiegenoten, werden ook de 1671 kinderen van deze langlevende mensen in de Leiden Langleven studie geïnccludeerd. Deze kinderen, van middelbare leeftijd overigens, werden vergeleken met hun leeftijdsgenoten en partners als controles (n=745). Leefstijl, omgevingsfactoren en gewicht zijn in deze twee groepen gelijk. Al het voorgaande onderzoek heeft echter laten zien dat de intrinsieke metabole gezondheid bij de langlevende familieleden groter is dan bij de controles. Bij het onderzoek naar markers van metabole gezondheid op middelbare

leeftijd wordt steeds gekeken of het onderscheid tussen de twee groepen door een marker verbetert ten opzichte van traditionele variabelen.

Metabolieten en het risico op coronaire hartziekten

Om te onderzoeken of metabolieten bruikbaar zijn voor de predictie van coronaire hartziekten hebben wij in een deel van de CAREMA studie (79 patiënten en 565 controles) metabolieten gemeten door middel van waterstof kernspin resonantie (in het Engels: *proton nuclear magnetic resonance* [^1H -NMR]) spectroscopie. Deze techniek leverde 100 bruikbare signalen op en deze 100 signalen representeerden 36 verschillende metabolieten (sommige metabolieten geven meerdere signalen, en aan andere signalen kon geen metaboliet toegekend worden). Door middel van een statistische test (LASSO regressie) vonden wij 16 verschillende signalen die, wanneer gecombineerd in één score konden voorspellen wie een verhoogd risico liep op coronaire hartziekten. Deze score in combinatie met leeftijd en geslacht presteerde ongeveer even goed als alle traditionele risicofactoren gecombineerd. Wanneer deze score werd toegevoegd aan een score bestaande uit alleen traditionele risicofactoren zagen wij geen verbetering in de risicopredictie. Door gebruik te maken van de ERF studie konden wij zien dat de geselecteerde pieken ook belangrijk waren voor het onderscheid tussen mensen met en zonder coronaire hartziekten.

Ook hebben wij bekeken wat de beste manier is om ^1H -NMR data te analyseren. Om informatie uit een ^1H -NMR spectrum te halen hebben wij twee verschillende methodes met elkaar vergeleken. Ten eerste een ongerichte aanpak waarbij het gehele ^1H -NMR spectrum in kleine stukjes wordt opgedeeld. Deze stukjes worden bins genoemd. Ten tweede een gerichte aanpak, zoals beschreven in de paragraaf hierboven, waarbij we ons concentreren op de metabolieten die een duidelijk signaal in het ^1H -NMR spectrum laten zien. Bij de gerichte aanpak vonden wij elf signalen die wanneer gecombineerd in één score waren geassocieerd met coronaire hartziekten, ook na correctie voor traditionele risicofactoren. Bij de ongerichte aanpak vonden wij 21 bins in het ^1H -NMR spectrum, die wanneer gecombineerd in één score geassocieerd waren met coronaire hartziekten, ook na correctie voor traditionele risicofactoren. Deze score leek een beter onderscheidend vermogen te hebben dan de score die was gebaseerd op de gerichte aanpak. Maar het bleek dat een groot deel van de geselecteerde bins in een regio lagen waar alleen ruis aanwezig was. Dit kan duiden op overaanpassing, beter bekend als “*overfitting*” in de wetenschappelijk literatuur, van het model. Het gevolg hiervan is dat een score gebaseerd op data met veel ruis mogelijk weinig voorspellende waarde heeft in een onafhankelijke populatie. Om hiervan zeker te zijn moeten wij deze score testen in een onafhankelijke populatie.

Op dit moment kan de risicopredictie van coronaire hartziekten niet verbeterd worden met de beperkte metabolomics meting die wij hebben verricht. Maar van de andere kant, een score gebaseerd op metaboliëten (een gemakkelijk meting in bloed op één moment), gecombineerd met leeftijd en geslacht liet een even goede predictie zien als alle traditionele risicofactoren gecombineerd. Daarnaast is maar een klein deel van alle mogelijke metaboliëten getest. Daarom is het erg interessant om in de toekomst te kijken of metaboliëten gemeten met behulp van andere platformen kunnen bijdragen aan de risicopredictie van coronaire hartziekten. Hierbij zullen de resultaten gerepliceerd moeten worden in onafhankelijke studies.

Genetica en het risico op coronaire hartziekten

Om te onderzoeken of genetische informatie gebruikt kan worden bij de risicopredictie van coronaire hartziekten zijn er 179 SNPs gemeten in de CAREMA studie (642 patiënten en 1905 controles). Deze 179 SNPs waren uitgekozen omdat ze volgens genoom-wijde associatie studies waren geassocieerd met coronaire hartziekten (29 SNPs) of met risicofactoren voor coronaire hartziekten (153 SNPs). Op basis van deze SNPs werden genetische risico scores (GRSs) samengesteld door het aantal risico allelen te tellen in elke persoon. Per SNP heb je maximaal twee risico allelen, dus voor een GRS bestaande uit 29 SNPs is de minimum score 0 en de maximum score 58. Dus hoe hoger de score van GRS, hoe groter de kans op ziekte. De GRSs (één gebaseerd op alle SNPs, één gebaseerd op de risicofactoren SNPs en één gebaseerd op de coronaire hartziekten SNPs) die op die manier werden samengesteld bleken te associëren met coronaire hartziekten, maar niet na correctie voor traditionele risicofactoren. Daarom hebben wij een GRS samengesteld op basis van de SNPs geassocieerd met coronaire hartziekten, maar aan een SNP met groter effect werd een groter gewicht toegekend. Deze gewogen GRS was geassocieerd met coronaire hartziekten, ook na correctie voor traditionele risicofactoren. Deze GRS zorgde ook voor een kleine verbetering in de risicopredictie. Als laatste probeerden wij een combinatie van SNPs te vinden die samen de beste risicopredictie laten zien, door gebruik te maken van een statische test waarmee de combinatie van de best voorspellende SNPs kan worden geselecteerd (LASSO regressie). Met deze test werden er 14 van de 179 SNPs geselecteerd. De GRS gebaseerd op deze SNPs was geassocieerd met coronaire hartziekten na correctie voor traditionele risicofactoren en liet een kleine verbetering in de risicopredictie zien. Om met zekerheid te kunnen zeggen of deze GRS de risicopredictie verbetert zal deze GRS nog getest moeten worden in een onafhankelijke populatie.

Op basis van het hierboven beschreven onderzoek kan geconcludeerd worden dat een gewogen GRS gebaseerd op SNPs geassocieerd met coronaire hartziekten een heel kleine

verbetering laat zien in de risico predictie voor coronaire hartziekten. Maar waarschijnlijk is deze verbetering klinisch niet relevant. Dit wordt ook ondersteund door andere onafhankelijk gepubliceerde studies.

Een gezonde stofwisseling en familiere langlevendheid

Door middel van een uitgebreid lipiden profiel hebben wij bekeken of wij mensen kunnen classificeren op grond van hun metabole gezondheid. Hiervoor hebben wij gebruik gemaakt van de Leiden Langlevende Studie. In deze studie zijn lipoproteïne deeltjes grootte gemeten met behulp van waterstof kernspin resonantie. Daarnaast hadden wij ook informatie over totaal cholesterol, LDL-cholesterol, HDL-cholesterol, triglyceriden, apolipoproteïne E, leeftijd, geslacht en body massa index beschikbaar. Uit onze analyses bleek dat de nakomelingen van langlevende mensen gemiddeld genomen grotere LDL en HDL deeltjes grootte hadden dan hun partners. Daarnaast hadden ze lagere triglyceriden waarden en een hogere totaal cholesterol/HDL-cholesterol ratio. Het bleek dat LDL deeltjes grootte bij mannen en triglyceriden niveaus bij vrouwen de meest informatieve markers waren voor familiere langlevendheid.

Hieruit kan geconcludeerd worden dat een gezonde stofwisseling en familiere langlevendheid worden gekenmerkt door grotere LDL deeltjes in mannen en lagere triglyceriden niveaus in vrouwen. Het is zou interessant zijn om te kijken of deze markers voor een gezonde stofwisseling en familiere langlevendheid ook gebruikt kunnen worden voor de risicopredictie van coronaire hartziekten. Omdat deze markers geslacht specifiek zijn zal dit apart in mannen en in vrouwen onderzocht moeten worden.

Het combineren van metaboliëten en genetische data

In de CAREMA studie (in 537 patiënten en 1323 controles) is ook gekeken naar de invloed van de vrije vetzuren op het ontstaan van coronaire hartziekten. Hierbij lag de focus op linolzuur, gamma-linoleenzuur, dihomogammalinoleenzuur en arachidonzuur. Linolzuur wordt in het lichaam omgezet in gamma-linoleenzuur door het enzym delta-6 desaturase. Vervolgens wordt dit vetzuur omgezet in dihomogammalinoleenzuur door het enzym elongase. Daarna wordt dihomogammalinoleenzuur omgezet in arachidonzuur door het enzym delta-5 desaturase. De verhouding tussen gamma-linoleenzuur en linolzuur is een maat voor de activiteit van delta-6 desaturase. Hoe hoger dit getal, hoe actiever dit enzym. Zo geeft ook de verhouding tussen arachidonzuur en dihomogammalinoleenzuur de activiteit van delta-5 desaturase weer. Het bleek een hogere delta-5 desaturase activiteit was geassocieerd met een lagere kans op coronaire hartziekten. Delta-6 desaturase activiteit

was niet geassocieerd met coronaire hartziekten. Uit eerdere studies is bekend dat een SNP (rs174547) gelegen in het enzym FADS1 is geassocieerd met delta-5 desaturase activiteit, maar niet met coronaire hartziekten. In de CAREMA studie was deze SNP ook geassocieerd met een hogere delta-5 desaturase activiteit, maar niet met coronaire hartziekten. Het bleek wel dat het beschermende effect van een hoge delta-5 desaturase activiteit alleen werd gevonden in dragers van het AA genotype van FADS1.

Hieruit kan geconcludeerd worden dat een hoge delta-5 desaturase activiteit beschermt tegen de ontwikkeling van coronaire hartziekten. Omdat het beschermende effect van een hoog delta-5 desaturase activiteit met name werd gevonden in dragers van het AA-genotype van het FADS1 gen, zou kunnen zijn dat bepaalde markers alleen nuttig zijn voor risicopredictie in mensen met specifieke genotypen. Dit pleit voor het combineren van metabolieten met genetische data.

Conclusie en toekomstig onderzoek

De huidige scores om het risico op coronaire hartziekten te voorspellen maken geen gebruik van metabolieten en genetische data. Daarom hebben wij de bruikbaarheid van dit soort data voor risicopredictie van coronaire hartziekten geëvalueerd. Op dit moment kunnen wij nog niet met zekerheid zeggen of metabolieten en/of genetische data de risicopredictie voor coronaire hartziekten kunnen verbeteren. Gebaseerd op de resultaten beschreven in dit proefschrift denken wij ook dat het belangrijk om te evalueren of markers voor gezond oud worden ook van nut kunnen zijn bij de predictie van coronaire hartziekten. Dit zal in mannen en vrouwen apart onderzocht moeten worden. Daarnaast is het ook belangrijk om genetische en metabolieten data met elkaar te combineren.

De meeste scores zijn ontwikkeld om het risico op coronaire hartziekten te voorspellen op middelbare leeftijd en kunnen niet gebruikt worden voor oudere individuen (ouder dan 85 jaar). Zo is bijvoorbeeld onder gezonde 85 plussers een wat verhoogde bloeddruk juist geassocieerd met een lagere mortaliteit. Dit houdt dat er nog veel onderzoek gedaan moet worden naar risicopredictie voor hart- en vaatziekten in de oudere bevolking. Het gebruik van metabolieten data kan hierbij een belangrijke rol spelen.

Bij een relatief jonge populatie leidt het gebruik van een 10-jarige risicoscore tot een onderschatting van het lange termijn risico op coronaire hartziekten. Juist bij een jonge populatie kan er veel gezondheidswinst worden behaald met preventieve strategieën. Daarom zijn cohorten nodig met jongere deelnemers en langere follow-up tijden dan gebruikelijk voor onderzoek naar coronaire hartziekten. Omdat genetische variatie stabiel is gedurende het hele leven, zouden juist deze markers geschikt kunnen zijn om het risico op coronaire hartziekten te voorspellen in een jongere populatie dan tot nu toe gebruikelijk.

Om een beter begrip te krijgen van coronaire hartziekten is het noodzakelijk om genetische markers en informatie over metaboliëten met elkaar te combineren, samen met omgevings- en leefstijlfactoren. Deze informatie moet worden verzameld in grote longitudinale cohorten waarbij bloed wordt verzameld op het moment dat mensen deelnemen aan de studie, maar ook tijdens follow-up. Dit maakt het mogelijk om de effecten van bijvoorbeeld stoppen met roken te evalueren. Het is ook belangrijk om relatief jonge personen en oudere personen te includeren, omdat deze groepen ondervertegenwoordigd zijn in de meeste onderzoeken naar het risico op coronaire hartziekten.

Supplement I

Supplementary material & methods

Acquisition and processing of ^1H -NMR spectra

All proton nuclear magnetic resonance (^1H -NMR) experiments were acquired on a 600 MHz Bruker Avance II spectrometer (Bruker BioSpin, Karlsruhe, Germany) equipped with a 5 mm TCI cryogenic probe head with Z-gradient system and automatic tuning and matching. All experiments were recorded at 310K. Temperature calibration was done prior to each batch of measurements using the method of Findeisen *et al.*¹ The duration of the $\pi/2$ pulses were automatically calibrated for each individual sample using a homonuclear-gated nutation experiment on the locked and shimmed samples after automatic tuning and matching of the probe head.²

The stored EDTA plasma and serum samples were thawed at 4°C and were mixed by inverting the tubes 10 times. Next, samples (300 μL) were mixed with 300 μL 75 mM disodium phosphate buffer in $\text{H}_2\text{O}/\text{D}_2\text{O}$ (80/20) with a pH of 7.4 containing 6.15 mM NaN_3 and 4.64 mM sodium 3-[trimethylsilyl] d_4 -propionate (TSP) using a Gilson 215 liquid handler in combination with a Bruker SampleTrack system. Samples were transferred into 5mm SampleJet NMR tubes in 96 tube racks using a modified Gilson 215 tube filling station and kept at 6°C on a SampleJet sample changer while queued for acquisition.

For water suppression presaturation of the water resonance with an effective field of $\gamma B_1 = 25$ Hz was applied during the relaxation delay.³ J-resolved spectra (JRES)⁴ were recorded with a relaxation delay of 2 s and a total of one scan for each increment in the indirect dimension. A data matrix of 40 x 12,288 data points was collected covering a sweep width of 78 x 10,000 Hz. A sine-shaped window function was applied and the data was zero-filled to 256 x 16,384 complex data points prior to Fourier transformation. The resulting data matrix was tilted along the rows by shifting each row (k) by $0.4992 \cdot (128 - k)$ points and symmetrised about the central horizontal lines in order to compensate for the skew of the multiplets in the F1 dimension. For T2-filtered ^1H NMR spectra a standard 1D CPMG (Carr–Purcell–Meiboom–Gill) pulse sequence,^{5,6} was used with a relaxation delay of 4 seconds. A pulse train of 128 refocusing pulses with individual spin echo delays of 0.6 ms were applied resulting in a total T2 filtering delay of 78 ms. 73,728 data points covering a spectral width of 12,019 Hz were collected using 16 scans. The FID was zero-filled to 131,072 complex data points and an exponential window function was applied with a line broadening factor of 1.0 Hz prior to Fourier transformation. The spectra were automatically phase and baseline corrected.

Quality control, scaling and calibration of the ^1H -NMR spectra

Further data processing was performed in Matlab® (R2009a, The Mathworks Inc., Natick, MA, USA). The spectra and associated data were converted into Matlab files using in-house code. First, the spectra were combined into one file while removing superfluous information. For CPMG this included dropping the imaginary part of the spectrum, while for the JRES spectra the sum projection along the indirect dimension was taken. Quality control (QC) on the set of ^1H -NMR spectra was carried out by examining a set of spectroscopic parameters such as shim values and intensity of the water signal, and subsequently visually inspecting the spectra. The spectra that failed the QC were not included for further analysis. The spectra were then scaled with respect to the sensitivity of the receiver coil, as determined from the pulse length that was automatically calibrated for each sample.⁷ After subtracting a constant value as a simple baseline correction, the spectra were calibrated with respect to the anomeric resonance of α -D-glucose ($\delta = 5.23$ ppm).⁸ Since there are small deviations of the peak position in the different ^1H -NMR spectra, alignment was performed using the correlation optimized warping algorithm by Tomasi *et al.*⁹ This was performed actively for the CPMG spectra, after which the same warping was applied to the JRES projection. Peaks in the JRES projection were picked by finding the signals that were above the surrounding spectral area by more than the estimated noise level. Peaks in different spectra were grouped according to similarity in peak position and intensity. The intensity and the position give good initial guesses for the deconvolution of the peaks by fitting groups of mixed Gaussian-Lorentzian line shapes to isolated spectral areas using a Simplex optimization algorithm. As the fitting algorithm incidentally fails to converge properly, values further from the median than 3 times the interquartile range are discarded. Using PLS regression, the remaining peak intensities were used to build a linear model that predicts all intensities directly from the non-warped spectrum, yielding also values for the cases where the peak picking failed or the deconvolution values were discarded, and eliminating the problem of faulty warping.

For 76 signals, metabolites were assigned using information from previously reported plasma/serum metabolites,^{6,8,10,11} the Human Metabolome Database (HMDB)¹² and the Pearson correlation coefficients between the signal intensities.

References

1. Findeisen M, Brand T, Berger S. A ^1H -NMR thermometer suitable for cryoprobes. *Magn Reson Chem* 2007;45(2):175-8.
2. Wu PS, Otting G. Rapid pulse length determination in high-resolution NMR. *J Magn Reson* 2005;176(1):115-9.
3. Price WS. Water signal suppression in NMR spectroscopy. *Annual Reports on Nmr Spectroscopy*, Vol 38 1999;38:289-354.
4. Aue WP. Homonuclear broad band decoupling and two-dimensional J-resolved NMR spectroscopy. *The Journal of Chemical Physics* 1976;64(10):4226.
5. Meiboom S, Gill D. Modified Spin-Echo Method for Measuring Nuclear Relaxation Times. *Review of Scientific Instruments* 1958;29(8):688.
6. Nicholson JK, Foxall PJ, Spraul M, et al. 750 MHz ^1H and ^1H - ^{13}C NMR spectroscopy of human blood plasma. *Anal Chem* 1995;67(5):793-811.
7. Wider G, Dreier L. Measuring protein concentrations by NMR spectroscopy. *J Am Chem Soc* 2006;128(8):2571-6.
8. MacIntyre DA, Jimenez B, Lewintre EJ, et al. Serum metabolome analysis by ^1H -NMR reveals differences between chronic lymphocytic leukaemia molecular subgroups. *Leukemia* 2010;24(4):788-97.
9. Tomasi G, van den Berg F, Andersson C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics* 2004;18(5):231-241.
10. Wevers RA, Engelke U, Heerschap A. High-resolution ^1H -NMR spectroscopy of blood plasma for metabolic studies. *Clin Chem* 1994;40(7 Pt 1):1245-50.
11. Wishart DS. Quantitative metabolomics using NMR. *Trac-Trends in Analytical Chemistry* 2008;27(3):228-237.
12. Wishart DS, Knox C, Guo AC, et al. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res* 2009;37(Database issue):D603-10.

Table S1 | An overview of all peaks identified in the 2 dimensional J-resolved ¹H-NMR spectrum and their association with incident CHD in the primary prospective case-cohort study.

Chemical shift	Assignment	Cases		Sub-cohort		Association with CHD	
		Mean (sd)	Min : Max	Mean (sd)	Min : Max	HR (95% CI)	P-value
0.873 ppm*	lipids (CH3) [†]	0.45 (1.48)	-0.93 : 6.67	-0.05 (0.94)	-0.97 : 6.67	1.40 (1.13-1.74)	0.002
0.929 ppm*	isoleucine	0.27 (0.99)	-1.60 : 2.89	-0.04 (1.00)	-2.02 : 5.08	1.14 (0.90-1.45)	0.282
0.947 ppm*	leucine	0.30 (1.05)	-1.67 : 3.83	-0.04 (1.00)	-1.72 : 4.98	1.23 (0.98-1.55)	0.074
0.952 ppm*	unknown	0.01 (1.24)	-3.14 : 3.00	0.00 (0.97)	-4.29 : 3.21	0.99 (0.77-1.26)	0.921
0.957 ppm*	leucine	0.24 (1.00)	-1.78 : 2.79	-0.04 (0.99)	-2.10 : 4.07	1.17 (0.92-1.50)	0.207
0.965 ppm	unknown	0.07 (1.13)	-3.56 : 2.45	-0.02 (1.00)	-3.56 : 4.76	0.97 (0.75-1.24)	0.780
0.973 ppm*	unknown	0.03 (1.00)	-2.35 : 2.93	-0.01 (1.00)	-3.13 : 3.77	0.93 (0.71-1.20)	0.568
0.981 ppm*	valine	0.33 (0.98)	-1.65 : 3.26	-0.05 (0.99)	-2.37 : 4.43	1.23 (0.97-1.57)	0.090
0.988 ppm*	unknown	0.24 (0.87)	-1.88 : 2.69	-0.03 (1.01)	-7.36 : 4.19	1.22 (0.96-1.55)	0.102
1.000 ppm*	isoleucine	0.27 (0.96)	-1.79 : 2.48	-0.03 (1.00)	-2.07 : 4.31	1.17 (0.92-1.49)	0.214
1.032 ppm*	valine	0.29 (1.02)	-1.88 : 3.26	-0.04 (0.99)	-2.35 : 4.36	1.18 (0.92-1.51)	0.193
1.054 ppm	unknown	0.26 (1.38)	-1.94 : 5.02	-0.04 (0.93)	-1.66 : 3.99	1.21 (0.95-1.55)	0.131
1.063 ppm*	ketoisovalerate	0.12 (1.27)	-1.89 : 7.24	-0.01 (0.96)	-1.90 : 4.89	1.09 (0.84-1.41)	0.509
1.163 ppm*	unknown	0.14 (1.77)	-0.18 : 15.48	-0.02 (0.83)	-0.21 : 14.04	1.11 (0.94-1.31)	0.235
1.174 ppm*	ethanol	0.14 (1.63)	-0.13 : 14.20	-0.02 (0.87)	-0.13 : 14.52	1.10 (0.93-1.29)	0.280
1.191 ppm*	3-hydroxybutyrate	0.08 (0.96)	-0.57 : 6.77	-0.02 (1.00)	-0.58 : 9.75	1.10 (0.90-1.34)	0.362
1.212 ppm*	unknown	0.15 (1.44)	-1.33 : 10.34	-0.02 (0.92)	-2.03 : 5.48	1.17 (0.91-1.52)	0.228
1.264 ppm*	lipids (CH2) [†]	0.42 (1.39)	-0.91 : 5.79	-0.05 (0.95)	-0.94 : 6.23	1.36 (1.10-1.69)	0.005
1.319 ppm*	lactate	0.20 (1.06)	-1.51 : 4.43	-0.02 (0.99)	-1.90 : 4.41	1.22 (0.95-1.56)	0.120
1.391 ppm*	unknown	-0.14 (1.02)	-1.99 : 2.69	0.02 (1.00)	-2.60 : 4.35	0.78 (0.60-1.00)	0.053
1.407 ppm*	unknown	0.01 (1.00)	-1.78 : 2.19	0.00 (1.00)	-2.53 : 3.24	1.02 (0.79-1.32)	0.904
1.425 ppm*	unknown	0.17 (1.06)	-1.93 : 2.78	-0.02 (1.00)	-1.98 : 3.51	1.18 (0.92-1.52)	0.198
1.471 ppm*	alanine	0.22 (0.93)	-1.72 : 2.54	-0.03 (1.00)	-2.35 : 3.45	1.13 (0.88-1.45)	0.344
1.706 ppm*	unknown	0.24 (1.20)	-1.77 : 4.62	-0.03 (0.98)	-2.67 : 4.62	1.21 (0.92-1.58)	0.172
1.908 ppm*	acetate	0.13 (0.88)	-0.77 : 3.84	-0.02 (1.01)	-0.81 : 16.58	1.04 (0.87-1.24)	0.656
2.001 ppm*	lipids (CH*2CH=CH) [†]	0.49 (1.72)	-1.01 : 8.67	-0.05 (0.91)	-1.19 : 8.67	1.37 (1.13-1.67)	0.001
2.035 ppm*	glycoproteins	0.52 (1.18)	-1.51 : 3.85	-0.06 (0.96)	-2.05 : 4.88	1.32 (1.03-1.69)	0.028
2.066 ppm*	glycoproteins	0.32 (1.12)	-1.59 : 4.33	-0.04 (0.99)	-2.51 : 4.33	1.20 (0.93-1.55)	0.159
2.099 ppm*	glutamine/ glutamate	-0.01 (0.88)	-1.85 : 2.77	-0.01 (1.01)	-2.43 : 3.58	0.97 (0.76-1.22)	0.773
2.125 ppm*	glutamine/ glutamate	0.00 (1.05)	-2.41 : 3.66	0.00 (1.01)	-2.48 : 3.66	1.03 (0.81-1.32)	0.791
2.221 ppm*	lipids (CH2CO) [†]	0.40 (1.55)	-0.76 : 7.24	-0.05 (0.93)	-0.84 : 7.24	1.35 (1.10-1.66)	0.004
2.260 ppm*	valine	0.11 (1.15)	-2.10 : 2.71	-0.02 (0.98)	-2.11 : 4.41	1.01 (0.78-1.30)	0.954
2.301 ppm*	3-hydroxybutyrate	0.01 (0.63)	-0.55 : 2.71	0.00 (1.03)	-0.64 : 11.41	1.04 (0.82-1.30)	0.766
2.349 ppm*	glutamate	0.50 (1.21)	-0.91 : 4.24	-0.06 (0.96)	-1.68 : 4.24	1.38 (1.09-1.75)	0.007
2.364 ppm*	pyruvate	-0.19 (0.99)	-1.47 : 3.13	0.02 (1.00)	-1.66 : 3.91	0.96 (0.73-1.26)	0.768

Chemical shift	Assignment	Cases		Sub-cohort		Association with CHD	
		Mean (sd)	Min : Max	Mean (sd)	Min : Max	HR (95% CI)	P-value
2.393 ppm*	3-hydroxybutyrate	0.16 (0.86)	-0.94 : 3.82	-0.02 (1.01)	-0.96 : 10.08	1.13 (0.92-1.39)	0.242
2.412 ppm	unknown	0.32 (0.99)	-2.37 : 2.67	-0.04 (0.99)	-2.64 : 3.11	1.22 (0.96-1.55)	0.110
2.430 ppm*	glutamine	-0.22 (1.04)	-1.69 : 2.10	0.02 (0.99)	-1.79 : 3.30	0.81 (0.63-1.05)	0.107
2.530 ppm*	citrate	-0.10 (0.81)	-1.52 : 2.13	0.01 (1.02)	-2.56 : 4.43	0.77 (0.59-0.99)	0.045
2.645 ppm*	citrate	-0.19 (0.76)	-1.81 : 1.84	0.02 (1.02)	-2.53 : 4.56	0.73 (0.56-0.95)	0.017
2.690 ppm	Mg-EDTA	0.11 (1.03)	-1.47 : 2.79	-0.01 (0.99)	-1.80 : 3.88	1.05 (0.82-1.33)	0.710
2.737 ppm	sarcosine	0.44 (1.57)	-1.45 : 9.23	-0.05 (0.96)	-1.45 : 9.23	1.32 (1.08-1.61)	0.007
2.801 ppm	aspartate	-0.04 (1.00)	-1.70 : 4.01	0.01 (1.00)	-2.36 : 4.75	0.84 (0.65-1.09)	0.191
2.858 ppm	asparagine	0.13 (0.98)	-1.50 : 5.23	-0.02 (1.00)	-2.95 : 10.18	1.18 (0.94-1.47)	0.148
2.914 ppm*	dimethylglycine	0.17 (1.04)	-2.07 : 2.53	-0.02 (0.99)	-2.42 : 3.57	1.02 (0.79-1.32)	0.875
3.024 ppm*	lysine	-0.08 (0.83)	-2.04 : 1.85	0.01 (1.02)	-2.24 : 3.89	0.81 (0.63-1.06)	0.119
3.034 ppm*	creatine + creatinine	0.17 (1.18)	-1.72 : 6.82	-0.03 (0.97)	-2.05 : 4.07	0.96 (0.75-1.23)	0.754
3.050 ppm*	ornithine	0.24 (1.11)	-2.03 : 2.55	-0.03 (0.99)	-2.49 : 3.42	1.07 (0.82-1.40)	0.614
3.090 ppm	Ca-EDTA	0.07 (1.13)	-2.09 : 4.19	0.00 (1.01)	-2.25 : 4.19	0.95 (0.75-1.22)	0.700
3.117 ppm	Ca-EDTA	0.08 (1.12)	-3.84 : 4.04	-0.02 (0.99)	-3.84 : 4.55	1.01 (0.79-1.28)	0.957
3.145 ppm*	unknown	0.02 (1.15)	-2.57 : 5.57	-0.01 (0.98)	-2.57 : 6.04	0.88 (0.66-1.16)	0.348
3.200 ppm	EDTA	0.11 (0.84)	-2.88 : 2.04	-0.01 (1.03)	-3.39 : 9.57	1.16 (0.91-1.48)	0.228
3.240 ppm*	glucose	0.32 (1.49)	-1.38 : 9.56	-0.05 (0.90)	-2.24 : 7.29	1.20 (0.98-1.47)	0.078
3.252 ppm	TMAO	0.32 (0.93)	-2.60 : 3.46	-0.04 (1.00)	-3.26 : 6.43	1.17 (0.93-1.47)	0.174
3.267 ppm*	1,5-anhydrosorbitol	-0.15 (1.07)	-5.08 : 1.98	0.01 (0.99)	-4.19 : 3.42	1.01 (0.80-1.28)	0.927
3.312 ppm	unknown	0.20 (0.88)	-1.80 : 3.71	-0.02 (1.01)	-2.00 : 11.23	1.09 (0.82-1.44)	0.555
3.333 ppm*	proline	0.17 (0.91)	-1.44 : 2.67	-0.03 (1.01)	-1.80 : 4.80	1.02 (0.80-1.29)	0.902
3.342 ppm*	1,5-anhydrosorbitol	-0.16 (0.92)	-1.96 : 2.10	0.01 (1.01)	-1.85 : 3.53	0.86 (0.66-1.13)	0.287
3.348 ppm	unknown	-0.02 (0.90)	-1.75 : 3.21	0.01 (1.01)	-2.67 : 3.57	1.00 (0.79-1.27)	0.978
3.354 ppm*	unknown	0.11 (1.53)	-1.15 : 10.50	-0.02 (0.90)	-1.15 : 9.50	1.21 (0.95-1.53)	0.124
3.362 ppm	unknown	0.17 (1.31)	-2.79 : 4.62	-0.02 (0.95)	-2.79 : 3.72	1.21 (0.93-1.56)	0.154
3.401 ppm*	glucose	0.38 (1.48)	-1.69 : 9.93	-0.05 (0.90)	-1.95 : 7.82	1.23 (1.01-1.49)	0.036
3.458 ppm	glucose	0.30 (1.47)	-1.53 : 9.22	-0.04 (0.91)	-1.71 : 6.45	1.26 (1.01-1.57)	0.040
3.487 ppm*	glucose	0.37 (1.43)	-1.64 : 9.52	-0.05 (0.91)	-2.03 : 7.82	1.22 (1.01-1.47)	0.042
3.531 ppm*	glucose	0.39 (1.48)	-1.64 : 9.76	-0.05 (0.90)	-1.83 : 7.82	1.23 (1.02-1.49)	0.033
3.551 ppm*	unknown	-0.18 (0.90)	-1.53 : 3.77	0.02 (1.01)	-4.31 : 4.53	0.81 (0.62-1.05)	0.111
3.599 ppm	EDTA	0.12 (0.84)	-2.86 : 2.52	-0.01 (1.02)	-3.41 : 9.90	1.18 (0.93-1.50)	0.175
3.633 ppm	unknown	0.00 (0.51)	-2.47 : 0.65	0.00 (1.05)	-22.4 : 2.58	0.97 (0.81-1.16)	0.755
3.649 ppm*	ethanol	0.13 (1.59)	-0.41 : 13.79	-0.02 (0.88)	-0.41 : 14.69	1.09 (0.92-1.29)	0.303
3.712 ppm*	glucose	0.33 (1.46)	-1.71 : 9.30	-0.05 (0.91)	-1.76 : 7.51	1.19 (0.98-1.45)	0.078
3.721 ppm*	glucose	0.36 (1.45)	-1.68 : 9.67	-0.05 (0.91)	-2.31 : 7.73	1.23 (1.01-1.49)	0.040
3.735 ppm	3-phosphoglycerate	0.22 (1.12)	-1.53 : 3.37	-0.04 (0.98)	-1.53 : 2.95	1.19 (0.94-1.52)	0.156
3.747 ppm*	unknown	0.12 (1.02)	-2.19 : 2.28	-0.01 (1.00)	-3.09 : 2.77	1.09 (0.85-1.39)	0.509

Chemical shift	Assignment	Cases		Sub-cohort		Association with CHD	
		Mean (sd)	Min : Max	Mean (sd)	Min : Max	HR (95% CI)	P-value
3.759 ppm*	glucose	0.32 (1.46)	-1.47 : 9.83	-0.04 (0.91)	-2.29 : 7.89	1.21 (0.99-1.47)	0.061
3.765 ppm	3-phosphoglycerate	0.25 (1.01)	-2.35 : 2.15	-0.04 (0.99)	-3.13 : 4.02	1.24 (0.97-1.58)	0.087
3.772 ppm	3-phosphoglycerate	0.23 (1.07)	-1.33 : 2.90	-0.03 (0.98)	-1.35 : 3.08	1.21 (0.95-1.53)	0.119
3.779 ppm*	alanine	-0.01 (0.97)	-2.80 : 2.26	0.01 (1.01)	-3.58 : 3.92	1.00 (0.78-1.27)	0.977
3.818 ppm*	unknown	0.04 (1.05)	-1.82 : 4.81	0.00 (0.99)	-2.09 : 6.76	1.00 (0.79-1.28)	0.983
3.824 ppm*	glucose	0.39 (1.40)	-2.49 : 7.92	-0.05 (0.92)	-1.89 : 6.32	1.29 (1.04-1.59)	0.019
3.838 ppm*	glucose	0.37 (1.5)	-1.69 : 9.94	-0.05 (0.89)	-2.07 : 7.37	1.25 (1.02-1.52)	0.031
3.876 ppm*	1,5-anhydrosorbitol	0.00 (1.23)	-5.59 : 2.69	0.00 (0.96)	-5.37 : 5.01	1.02 (0.80-1.29)	0.888
3.888 ppm	unknown	0.03 (1.10)	-2.28 : 4.80	-0.01 (0.99)	-2.50 : 3.68	0.96 (0.75-1.23)	0.745
3.893 ppm*	glucose	0.36 (1.47)	-1.57 : 9.91	-0.05 (0.90)	-1.96 : 7.83	1.22 (1.01-1.49)	0.042
3.924 ppm*	unknown	0.00 (0.93)	-2.30 : 2.62	0.00 (1.01)	-2.64 : 3.36	0.78 (0.60-1.01)	0.062
3.959 ppm*	unknown	-0.04 (0.77)	-1.65 : 1.42	0.01 (1.02)	-2.11 : 9.13	0.93 (0.71-1.21)	0.570
3.974 ppm*	1,5-anhydrosorbitol	-0.10 (1.02)	-2.12 : 2.85	0.01 (1.00)	-2.85 : 2.98	1.02 (0.78-1.32)	0.897
3.981 ppm	Multiple metabolites	-0.16 (0.90)	-1.87 : 2.73	0.02 (1.01)	-2.46 : 5.23	0.86 (0.66-1.12)	0.259
4.044 ppm*	creatinine	0.38 (1.47)	-1.85 : 10.37	-0.05 (0.90)	-2.51 : 4.16	1.35 (1.04-1.76)	0.027
4.103 ppm*	lactate	0.20 (1.07)	-1.41 : 4.57	-0.02 (0.99)	-1.82 : 4.60	1.23 (0.96-1.57)	0.100
4.121 ppm*	proline	0.13 (0.93)	-2.01 : 2.20	-0.02 (1.01)	-2.20 : 4.39	0.97 (0.76-1.25)	0.831
4.167 ppm	phosphorylcholine	-0.02 (0.87)	-1.60 : 2.65	0.00 (1.02)	-2.68 : 6.10	0.82 (0.61-1.10)	0.179
4.223 ppm	3-phosphoglycerate	-0.01 (0.96)	-1.02 : 4.42	0.00 (1.00)	-1.20 : 4.84	0.98 (0.76-1.27)	0.894
4.239 ppm*	unknown	-0.12 (0.88)	-1.62 : 2.64	0.02 (1.02)	-2.00 : 4.14	0.92 (0.72-1.18)	0.519
4.503 ppm*	glucose	0.26 (1.14)	-1.52 : 5.96	-0.03 (0.97)	-2.54 : 4.61	1.23 (0.98-1.55)	0.074
5.180 ppm*	unknown	0.29 (1.42)	-2.20 : 7.36	-0.03 (0.94)	-2.20 : 5.14	1.27 (1.01-1.60)	0.046
5.230 ppm*	glucose	0.38 (1.47)	-1.79 : 9.75	-0.05 (0.90)	-1.98 : 7.97	1.23 (1.02-1.48)	0.035
5.299 ppm*	lipids (CH=CH) [†]	0.41 (1.65)	-0.61 : 7.82	-0.04 (0.91)	-0.76 : 7.82	1.37 (1.12-1.67)	0.002
6.890 ppm*	tyrosine	0.03 (0.90)	-1.74 : 2.89	-0.01 (1.02)	-2.38 : 5.14	0.85 (0.66-1.10)	0.212
7.186 ppm*	tyrosine	0.05 (1.01)	-1.50 : 2.95	-0.01 (1.00)	-2.15 : 4.98	0.91 (0.70-1.17)	0.444
8.451 ppm*	formate	0.02 (1.03)	-1.49 : 4.45	0.00 (1.00)	-1.30 : 9.41	0.90 (0.65-1.23)	0.495

Mean is expressed as area under the curve

*Detected and quantified in the confirmation study

[†]The term in parenthesis indicates the structural feature of the lipid measured by ¹H-NMR spectroscopy

Table S2 | Baseline characteristics of the traditional CHD risk factors in the primary study before removing participants with missing data.

	Cases (N=125)		Sub-cohort (N=738)*	
	N missing	Mean±SD/%(N)	N missing	Mean±SD/%(N)
Age (years)	0	50.49 ± 7.11	0	44.57 ± 8.23
Total cholesterol (mmol/L)	2	5.91 ± 0.99	13	5.33 ± 0.98
HDL cholesterol (mmol/L)	2	1.11 ± 0.28	13	1.33 ± 0.36
SBP (mmHg)	0	133.18 ± 17.83	1	121.23 ± 14.61
BMI (kg/m ²)	0	27.69 ± 4.45	0	25.48 ± 3.87
Men	0	95 (76.0%)	0	323 (43.8%)
Current smoking	2	69 (56.1%)	5	283 (38.6%)
Diabetes	0	6 (4.8%)	0	5 (0.7%)
Parental history of MI	14	58 (52.3%)	68	264 (39.4%)

HDL, high-density lipoprotein; *MI*, myocardial infarction; *SBP*, systolic blood pressure; *BMI*, body mass index

* Including 10 cases

Table S3 | The association between the metabolite score based on 16 ¹H-NMR signals and the metabolite score based on the 13 ¹H-NMR signals and incident CHD in the primary study.

	Univariate analysis*		Adjusted for TRF†	
	HR (95% CI)	P-value	HR (95% CI)	P-value
Metabolite score based on 16 ¹ H-NMR signals	1.93 (1.50-2.48)	<0.001	1.58 (1.18-2.12)	0.002
Metabolite score based on 13 ¹ H-NMR signals‡	1.91 (1.50-2.44)	<0.001	1.50 (1.12-2.01)	0.007

HR, hazard ration; *TRF*, traditional risk factors

*Univariate HR were calculated per unit increase for the metabolite scores. Age in years was used as the time-scale variable.

†The TRF include total cholesterol, HDL cholesterol, SBP, BMI, gender, current smoking, self-reported diabetes and parental history of MI. Age in years was used as the time-scale variable.

‡This metabolite score is based on the 13 ¹H-NMR signals that were also available in the secondary study.

Table S4 | Reclassification of participants when the metabolite score was used in addition to a risk score based on traditional risk factors (*i.e.* age, sex, current smoking, total cholesterol, HDL cholesterol, BMI, parental history of MI and self-reported diabetes). For reclassification the following risk categories were used: 0-<5%, 5-<10%, 10-<20%, ≥20%.

Model with TRF*	Model with TRF + metabolite score				Total
	<5%	5-<10%	10-<20%	≥20%	
Incident cases					
<5%	39	3	1	0	43
5-<10%	5	13	3	1	22
10-<20%	0	2	4	0	6
≥20%	0	0	1	4	5
Total	44	18	9	5	76
Non-cases					
<5%	515	7	1	0	523
5-<10%	4	15	2	0	21
10-<20%	0	3	7	0	10
≥20%	0	0	0	0	0
Total	519	25	10	0	554

TRF, traditional risk factors

*The TRF include age in years, total cholesterol, HDL cholesterol, SBP, BMI, gender, current smoking, self-reported diabetes and parental history of MI.

Table S5 | Improvement in reclassification of incident CHD when using the metabolite score in addition to traditional risk factors (*i.e.* age, sex, current smoking, total cholesterol, HDL cholesterol, BMI, parental history of MI and self-reported diabetes). For calculating the NRI the following risk categories were used: 0-<5%, 5-<10%, 10-<20%, ≥20%.

	Reclassification				
	P Up	P Down	NRI	Z-score	P-value
Event	0.140	0.096	0.044	0.924	0.178
Non-event	0.018	0.013	-0.005	-1.593	0.944
Total	...	...	0.038	0.809	0.209

P Up, proportion of participants placed into a higher risk category; P Down, proportion of participants placed into a lower risk category; NRI, net reclassification index.

Table S6a | Comparing the predictive power of the single traditional risk factors, all traditional risk factors combined with and without the metabolite score in the primary case-cohort study using Harrell's C index.

	TRFs only	TRF + MS		
	C-index (95% CI)	C-index (95% CI)	Diff (SE)	P-value
Age	0.72 (0.66-0.77)	0.79 (0.74-0.84)	0.08 (0.02)	<0.001
Total cholesterol	0.66 (0.61-0.72)	0.77 (0.72-0.81)	0.10 (0.03)	<0.001
HDL cholesterol	0.67 (0.60-0.73)	0.74 (0.68-0.79)	0.07 (0.02)	<0.001
Systolic blood pressure	0.71 (0.65-0.77)	0.79 (0.74-0.84)	0.08 (0.02)	<0.001
Body mass index	0.64 (0.58-0.71)	0.74 (0.69-0.80)	0.10 (0.03)	<0.001
Male sex	0.66 (0.61-0.71)	0.76 (0.71-0.81)	0.10 (0.02)	<0.001
Current smoking	0.57 (0.51-0.62)	0.74 (0.69-0.80)	0.18 (0.03)	<0.001
Self-reported diabetes	0.51 (0.49-0.54)	0.76 (0.71-0.81)	0.25 (0.03)	<0.001
Parental history of MI	0.55 (0.49-0.60)	0.75 (0.70-0.80)	0.20 (0.03)	<0.001
<i>All TRFs combined</i>	<i>0.82 (0.78-0.87)</i>	<i>0.84 (0.80-0.87)</i>	<i>0.01 (0.01)</i>	<i>0.107</i>

TRFs, traditional risk factors; MS, metabolite score; Diff, Difference in C-indices; SE, standard error; HDL, high density lipoprotein; MI, myocardial infarction

Table S6b | Comparing the predictive power of the metabolite score combined with and without single traditional risk factors and all traditional risk factors combined in the primary case-cohort study using Harrell's Concordance index.

	C-index (95% CI)	Diff (SE)	P-value
<i>Metabolite Score only</i>	<i>0.75 (0.74-0.84)</i>	-	-
Age + MS	0.79 (0.74-0.84)	0.038 (0.022)	0.090
Total cholesterol + MS	0.77 (0.72-0.81)	0.011 (0.008)	0.194
HDL cholesterol +MS	0.74 (0.68-0.79)	0.017 (0.021)	0.412
Systolic blood pressure +MS	0.79 (0.74-0.84)	0.035 (0.019)	0.064
Body mass index +MS	0.74 (0.69-0.80)	0.011 (0.010)	0.308
Male sex +MS	0.76 (0.71-0.81)	0.007 (0.020)	0.720
Current smoking +MS	0.74 (0.69-0.80)	0.011 (0.015)	0.476
Self-reported diabetes +MS	0.76 (0.71-0.81)	0.002 (0.003)	0.453
Parental history of MI +MS	0.75 (0.70-0.80)	0.008 (0.008)	0.970
<i>All TRFs combined</i>	<i>0.84 (0.80-0.87)</i>	<i>0.081 (0.023)</i>	<i><0.001</i>

Diff, Difference in C-indices between the metabolite score only and the traditional risk factor(s) and the metabolite score combined; SE, standard error; MS, metabolite score; TRFs, traditional risk factors

Table S7 | Baseline characteristics of the traditional CHD risk factors in the secondary study before removing participants with missing data.

	Cases (N=174)		Controls (N=2170)*	
	N missing	Mean±SD/%(N)	N missing	Mean±SD/%(N)
Age (years)	0	60.98 (11.73)	0	47.70 (14.02)
Total cholesterol (mmol/L)	1	5.00 (1.07)	12	5.60 (1.06)
HDL cholesterol (mmol/L)	1	1.15 (0.32)	12	1.29 (0.37)
SBP (mmHg)	1	148.31 (22.91)	30	139.18 (19.75)
BMI (kg/m ²)	3	28.14 (4.43)	35	26.79 (4.61)
Men	0	108 (62.1%)	0	924 (42.6%)
Current smoking	1	62 (35.8%)	3	844 (38.9%)
Diabetes	0	25 (14.4%)	0	101 (4.7%)
Parental history of MI	69	54 (31.0%)	729	430 (19.9%)

HDL, high-density lipoprotein; *MI*, myocardial infarction; *SBP*, systolic blood pressure; *BMI*, body mass index

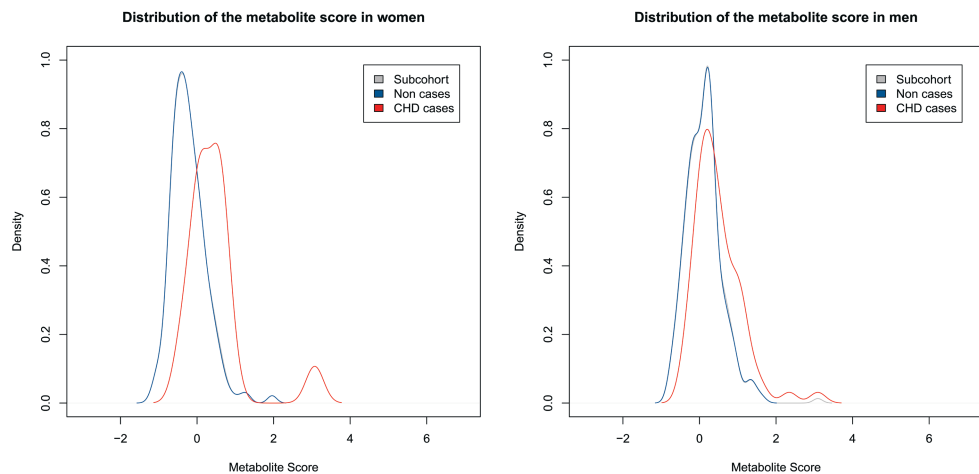
Table S8 | The association between the metabolite score and coronary heart disease in the total confirmation study (170 cases and 2157 controls) and after excluding statin users (n=299; 77 cases and 1946 controls left for analysis).

Metabolite score	Crude OR (95% CI)	P-value	Adjusted OR (95%CI)*	P-value
Total population	2.72 (2.28-3.25)	2.0x10 ⁻²⁸	1.59 (1.19-2.13)	0.002
Non statin users	2.78 (2.16-3.57)	1.2x10 ⁻¹⁵	2.02 (1.33-3.08)	0.001

OR, odds ratio; CI, confidence interval

Odds ratios were calculated per standard deviation increase in metabolite score

*The metabolite score was adjusted for age, total cholesterol, HDL cholesterol, systolic blood pressure, body mass index, sex, smoking, diabetes and parental history of myocardial infarction.

**Figure S1** | Kernel density plots showing the distribution of the metabolite score in the sub-cohort and in the cases of the primary case-cohort study.

Supplement II

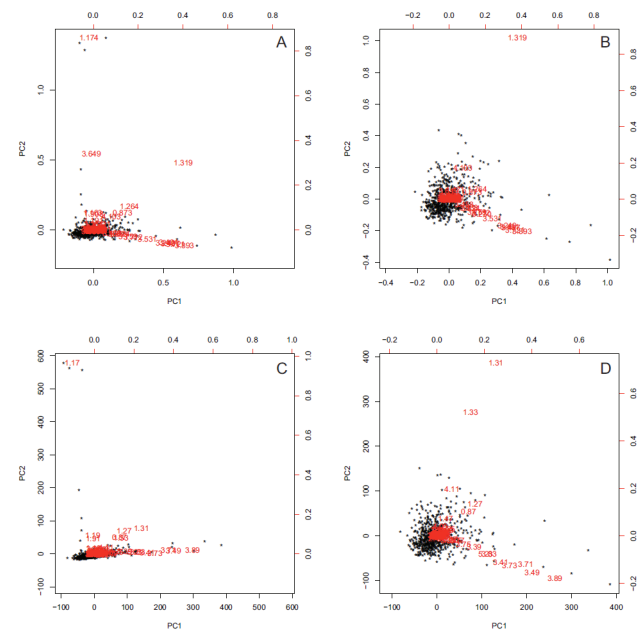


Figure S1 | Results from the principal component analysis on the 98 ¹H-NMR peaks (a/b) and on the 401 ¹H-NMR bins (c/d). The black dots represent the participants and the red numbers represent the peaks or the bins. There are three individuals with aberrant values for the peak located at 1.174 ppm representing ethanol (a). After removing this peak, there were some individuals with high values for peaks representing glucose (b). A bin located at 1.17 was aberrant for the same three individuals (c). After removing this bin, there were six individuals with high values for bins located near peaks representing glucose (d). The participants with high values for peaks or bins representing glucose also had high glucose levels according to a conventional method to determine glucose.

Supplement III

Material and methods

Statistical analysis for LASSO GRSs

Not all GWAS-identified SNPs may contribute to the risk of future CHD in our population-based prospective study. Therefore, we selected a sub-set of the most informative SNPs from all GWAS identified SNPs using least absolute shrinkage and selection operator (LASSO) regression. With LASSO regression it is possible to perform variable selection and shrinkage at the same time, which makes it an effective method for finding prediction rules based in high dimensional data.¹⁻³ The penalty parameter, which determines the number of variables selected for the model, was set by the 10-fold cross-validated partial likelihood. With 10-fold cross-validation we have addressed the issue of internal validation.⁴ The resulting SNPs and their weights, based on the coefficients resulting from 10-fold cross-validated LASSO regression, were then used to construct the LASSO GRS. This approach was used to select subsets of informative SNPs from the 29 CHD SNPs (CHD LASSO GRS) and all 179 SNPs (LASSO GRS). For this analysis the R package `penalized` (version 0.9-32), was used.

References

1. Goeman JJ. L1 penalized estimation in the cox proportional hazards model. *Biom.J.* 2010;52:70-84
2. Tibshirani R. Regression shrinkage and selection via the lasso. *J.R.Statist.Soc.B.* 1996;58:267-288
3. Tibshirani R. The lasso method for variable selection in the cox model. *Stat.Med.* 1997;16:385-395

Supplementary results

Construction the LASSO GRSs and their associations with incident CHD

By performing 10-fold cross-validated LASSO regression on the 29 CHD SNPs, a subset of 3 SNPs (see supplementary table S7) was selected as the optimal predictor for incident CHD in our data-set, namely rs10757278, rs11556924 and rs646776 (see supplementary table S7 for locus and nearby genes). When performing LASSO regression on all 179 SNPs, a subset of 14 SNPs was selected (see supplementary table S7). Three SNPs (rs10757278, rs11556924 and rs646776) were also selected for the CHD LASSO GRS, the remaining 11 SNPs were: rs2925979, rs6882076, rs2954029, rs6987702, rs10889352, rs2972146, rs514230, rs8050136, rs181362, rs925946 and rs2000999 (for locus and nearby genes see table S7).

Supplementary Tables

Supplementary table S1 can be found online at:

<http://circgenetics.ahajournals.org/content/5/2/202/suppl/DC1>

Supplementary table S1 | Candidate SNPs (based on the association between CHD and intermediate risk factors from previously published GWAS) for genotyping in the CAREMA study.

Table S2 | Success rates (in the total population, and in the cases and the sub-cohort separately) and the Hardy-Weinberg equilibrium (only in sub-cohort) for every genotyped SNP in the CAREMA cohort.

SNP	CAREMA N=2548				Subcohort N=1983				Cases N=667				HWE in subcohort				P-value	Remarks
	N miss	Success rate	N miss	Success rate	N miss	Success rate	N miss	Success rate	A1/A2	GENO	Obs/Exp	na	na	na				
SNPs selected for their association with coronary artery diseases																		
RS46522	8	1.00	7	1.00	1	1.00	1	1.00	C/T	403/1014/559	0.51/0.50	0.160						
RS216172	12	1.00	8	1.00	5	0.99	5	0.99	G/C	242/892/841	0.45/0.45	0.843						
RS579459	26	0.99	16	0.99	11	0.98	11	0.98	C/T	113/689/1165	0.35/0.36	0.412						
RS599839	na	na	na	na	na	na	na	na	na	na	na	na	na	na	na	No PCR product		
RS646776	23	0.99	21	0.99	3	1.00	3	1.00	G/A	86/665/1211	0.34/0.34	0.687						
RS964184	31	0.99	28	0.99	5	0.99	5	0.99	C/G	41/469/1445	0.24/0.24	0.708						
RS974819	8	1.00	7	1.00	1	1.00	1	1.00	A/G	189/860/927	0.44/0.43	0.638						
RS1122608	24	0.99	21	0.99	4	0.99	4	0.99	T/G	104/730/1128	0.37/0.36	0.352						
RS1412444	14	0.99	11	0.99	3	1.00	3	1.00	A/G	237/849/886	0.43/0.45	0.130						
RS1746048	91	0.96	72	0.96	23	0.97	23	0.97	T/C	40/437/1434	0.23/0.23	0.328						
RS2505083	13	0.99	9	1.00	5	0.99	5	0.99	C/T	326/958/690	0.49/0.48	0.852						
RS2895811	9	1.00	7	1.00	2	1.00	2	1.00	C/T	333/952/691	0.48/0.48	0.889						
RS3825807	11	1.00	10	0.99	1	1.00	1	1.00	C/T	354/942/677	0.48/0.49	0.405						
RS4380028	10	1.00	9	1.00	1	1.00	1	1.00	A/G	317/910/747	0.46/0.48	0.156						
RS4773144	8	1.00	7	1.00	1	1.00	1	1.00	G/A	392/992/592	0.50/0.49	0.555						
RS6725887	127	0.95	100	0.95	34	0.95	34	0.95	G/A	34/400/1449	0.21/0.22	0.290						
RS6922269	64	0.97	40	0.98	28	0.96	28	0.96	A/G	140/733/1070	0.38/0.39	0.347						
RS9818870	45	0.98	40	0.98	6	0.99	6	0.99	T/C	49/536/1358	0.28/0.27	0.740						
RS9982601	71	0.97	54	0.97	17	0.97	17	0.97	T/C	28/460/1441	0.24/0.23	0.238						
RS10757278	103	0.96	82	0.96	22	0.97	22	0.97	G/A	450/938/513	0.49/0.50	0.614						
RS10953541	10	1.00	9	1.00	1	1.00	1	1.00	T/C	123/718/1133	0.36/0.37	0.503						
RS11206510	38	0.99	27	0.99	12	0.98	12	0.98	C/T	77/572/1307	0.29/0.30	0.155						
RS11556924	10	1.00	9	1.00	1	1.00	1	1.00	T/C	313/935/726	0.47/0.48	0.672						
RS12190287	10	1.00	8	1.00	2	1.00	2	1.00	G/C	272/881/822	0.45/0.46	0.144						
RS12413409	11	1.00	10	0.99	1	1.00	1	1.00	A/G	17/297/1659	0.15/0.15	0.377						
RS12526453	63	0.98	41	0.98	27	0.96	27	0.96	G/C	237/841/864	0.43/0.45	0.142						

SNP	CAREMA N=2548			Subcohort N=1983			Cases N=667			HWE in subcohort		
	N miss	Success rate	N miss	Success rate	N miss	Success rate	N miss	Success rate	A1/A2	GENO	Obs/Exp	P-value
RS12936587	10	1.00	8	1.00	2	1.00			A/G	372/981/622	0.50/0.49	0.681
RS17114036	24	0.99	16	0.99	9	0.99			G/A	21/303/1643	0.15/0.16	0.119
RS17465637	35	0.99	29	0.99	8	0.99			A/C	157/753/1044	0.39/0.40	0.210
RS17609940	13	0.99	9	1.00	5	0.99			C/G	61/613/1300	0.31/0.30	0.298
<i>SNPs selected for their association with blood pressure</i>												
RS381815	16	0.99	13	0.99	3	1.00			T/C	144/741/1085	0.38/0.39	0.267
RS1004467	15	0.99	13	0.99	3	1.00			C/T	17/319/1634	0.16/0.16	0.680
RS1378942	8	1.00	8	1.00	0	1.00			G/T	210/854/911	0.43/0.44	0.643
RS1530440	10	1.00	9	1.00	1	1.00			T/C	67/581/1326	0.29/0.30	0.705
RS2681472	7	1.00	7	1.00	0	1.00			C/T	51/516/1409	0.26/0.26	0.670
RS3184504	7	1.00	7	1.00	0	1.00			C/T	466/998/512	0.51/0.50	0.653
RS9815354	12	1.00	11	0.99	1	1.00			A/G	58/563/1351	0.29/0.29	1.000
RS11014166	14	0.99	12	0.99	3	1.00			T/A	228/872/871	0.44/0.45	0.687
RS11191548	9	1.00	8	1.00	1	1.00			C/T	17/295/1663	0.15/0.15	0.304
RS12946454	7	1.00	7	1.00	0	1.00			T/A	136/759/1081	0.38/0.39	0.861
RS16948048	7	1.00	7	1.00	0	1.00			G/A	296/922/758	0.47/0.47	0.568
RS16998073	9	1.00	8	1.00	1	1.00			T/A	173/811/991	0.41/0.41	0.704
RS17367504	10	1.00	10	0.99	0	1.00			G/A	39/503/1431	0.25/0.25	0.590
<i>SNPs selected for their association with diabetes type 2</i>												
RS5215	74	0.97	58	0.97	16	0.98			G/A	269/894/762	0.46/0.47	0.807
RS5219	154	0.94	125	0.94	36	0.95			A/G	248/848/762	0.46/0.46	0.616
RS864745	70	0.97	58	0.97	13	0.98			G/A	455/933/537	0.48/0.50	0.201
RS1111875	65	0.97	47	0.98	20	0.97			T/C	332/902/702	0.47/0.48	0.157
RS2237892	100	0.96	78	0.96	26	0.96			T/C	9/212/1684	0.11/0.11	0.415
RS2237897	85	0.97	61	0.97	30	0.96			T/C	8/141/1773	0.07/0.08	0.012
RS4402960	32	0.99	28	0.99	6	0.99			A/C	198/845/912	0.43/0.43	0.917
RS4506565	74	0.97	63	0.97	13	0.98			T/A	197/901/822	0.47/0.45	0.032
RS4607103	57	0.98	36	0.98	25	0.96			T/C	130/714/1103	0.37/0.38	0.334
RS4689388	8	1.00	7	1.00	1	1.00			G/A	325/963/688	0.49/0.48	0.710

SNP	CAREMA N=2548			Subcohort N=1983			Cases N=667			HWE in subcohort			P-value	Remarks
	N miss	Success rate	N miss	Success rate	N miss	Success rate	A1/A2	GENO	Obs/Exp					
RS4712524	94	0.96	70	0.96	26	0.96	G/A	198/804/911	0.42/0.43	0.313				
RS6931514	92	0.96	71	0.96	23	0.97	G/A	141/762/1009	0.40/0.40	0.908				
RS7578597	52	0.98	41	0.98	13	0.98	G/A	25/384/1533	0.20/0.20	0.820				
RS7961581	93	0.96	75	0.96	21	0.97	C/T	180/779/949	0.41/0.42	0.274				
RS8050136	112	0.96	96	0.95	22	0.97	A/C	280/811/796	0.43/0.46	0.002				
RS10811661	49	0.98	32	0.98	21	0.97	C/T	69/606/1276	0.31/0.31	0.826				
RS10923931	111	0.96	90	0.95	25	0.96	A/C	21/309/1563	0.16/0.17	0.217				
RS10946398	63	0.98	47	0.98	18	0.97	C/A	197/825/914	0.43/0.43	0.598				
RS12779790	119	0.95	93	0.95	31	0.95	G/A	60/555/1275	0.29/0.29	1.000				
RS13266634	54	0.98	35	0.98	23	0.97	T/C	177/845/926	0.43/0.43	0.457				
SNPs selected for their association with blood lipid levels														
RS328	94	0.96	73	0.96	23	0.97	G/C	14/397/1499	0.21/0.20	0.027				
RS693	51	0.98	32	0.98	23	0.97	T/C	457/958/536	0.49/0.50	0.468				
RS7679	124	0.95	106	0.95	23	0.97	C/T	87/564/1226	0.30/0.32	0.040				
RS12916	21	0.99	10	0.99	12	0.98	C/T	370/966/637	0.49/0.49	0.927				
RS157580	77	0.97	66	0.97	14	0.98	G/A	273/896/748	0.47/0.47	0.884				
RS173539	na	na	na	na	na	na	na	na	na	na	No PCR product			
RS174547	56	0.98	35	0.98	25	0.96	G/A	205/870/873	0.45/0.44	0.608				
RS174570	157	0.94	121	0.94	42	0.94	T/C	50/405/1407	0.22/0.23	0.003				
RS181362	25	0.99	16	0.99	10	0.99	T/C	101/639/1227	0.32/0.34	0.140				
RS439401	105	0.96	71	0.96	41	0.94	T/C	244/828/840	0.43/0.45	0.076				
RS442177	26	0.99	18	0.99	9	0.99	G/T	343/924/698	0.47/0.48	0.225				
RS471364	93	0.96	68	0.97	30	0.96	G/A	30/387/1498	0.20/0.21	0.375				
RS492602	14	0.99	12	0.99	2	1.00	G/A	389/954/628	0.48/0.49	0.437				
RS514230	25	0.99	16	0.99	10	0.99	A/T	446/989/532	0.50/0.50	0.752				
RS515135	na	na	na	na	na	na	na	Na	na	na	Abnormal clusterplot			
RS581080	14	0.99	9	1.00	6	0.99	C/G	64/614/1296	0.31/0.31	0.461				
RS599839	na	na	na	na	na	na	na	na	na	na	No PCR product			

SNP	CAREMA N=2548			Subcohort N=1983			Cases N=667			HWE in subcohort			
	N miss	Success rate	N miss	Success rate	N miss	Success rate	N miss	Success rate	A1/A2	GENO	Obs/Exp	P-value	Remarks
RS605066	100	0.96	80	0.96	23	0.97			C/T	376/849/678	0.45/0.49	<0.001	Out of HWE
RS629301	15	0.99	11	0.99	4	0.99			C/A	86/664/1222	0.34/0.33	0.787	
RS645040	26	0.99	17	0.99	10	0.99			G/T	84/656/1226	0.33/0.33	0.786	
RS646776	23	0.99	21	0.99	3	1.00			G/A	86/665/1211	0.34/0.34	0.687	
RS737337	110	0.96	85	0.96	28	0.96			C/T	16/288/1594	0.15/0.15	0.456	
RS780093	68	0.97	53	0.97	15	0.98			T/C	331/917/682	0.48/0.48	0.451	
RS838880	11	1.00	10	0.99	1	1.00			G/A	193/820/960	0.42/0.42	0.367	
RS964184	31	0.99	28	0.99	5	0.99			C/G	41/469/1445	0.24/0.24	0.708	
RS1042034	11	1.00	10	0.99	1	1.00			G/A	102/679/1192	0.34/0.35	0.697	
RS1084651 (RS1652507)	12	1.00	8	1.00	5	0.99			C/T	44/491/1440	0.25/0.25	0.787	
RS1167998	39	0.98	31	0.98	11	0.98			G/T	213/905/834	0.46/0.45	0.174	
RS1260326	65	0.97	43	0.98	28	0.96			T/C	356/931/653	0.48/0.49	0.457	
RS1367117	506	0.80	406	0.80	113	0.83			T/C	191/448/938	0.28/0.39	<0.001	Success- rate <90%
RS1495741	10	1.00	9	1.00	1	1.00			G/A	85/714/1175	0.36/0.35	0.080	
RS1501908	73	0.97	65	0.97	10	0.99			C/G	257/888/773	0.46/0.46	0.961	
RS1532085	30	0.99	25	0.99	6	0.99			A/G	281/921/756	0.47/0.47	1.000	
RS1532624	70	0.97	50	0.97	22	0.97			A/C	354/937/642	0.48/0.49	0.710	
RS1564348	50	0.98	35	0.98	16	0.98			C/T	52/470/1426	0.24/0.25	0.086	
RS1689800	17	0.99	13	0.99	5	0.99			G/A	296/887/787	0.45/0.47	0.076	
RS1800562	11	1.00	10	0.99	1	1.00			A/G	9/231/1733	0.12/0.12	0.571	
RS1800588	21	0.99	18	0.99	4	0.99			T/C	101/698/1166	0.36/0.35	0.848	
RS1800961	70	0.97	51	0.97	21	0.97			T/C	2/98/1832	0.05/0.05	0.388	
RS1883025	101	0.96	73	0.96	33	0.95			A/G	125/686/1099	0.36/0.37	0.216	
RS2000999	8	1.00	7	1.00	1	1.00			A/G	79/632/1265	0.32/0.32	1.000	
RS2068888	15	0.99	10	0.99	6	0.99			A/G	411/958/604	0.49/0.50	0.388	
RS2072183	24	0.99	16	0.99	9	0.99			G/C	101/711/1155	0.36/0.36	0.569	
RS2075650	65	0.97	49	0.98	18	0.97			G/A	28/483/1423	0.25/0.24	0.072	
RS2081687	11	1.00	10	0.99	1	1.00			T/C	252/880/841	0.45/0.46	0.374	

SNP	CAREMA N=2548			Subcohort N=1983			Cases N=667			HWE in subcohort			
	N miss	Success rate	na	N miss	Success rate	na	N miss	Success rate	na	A1/A2	GENO	Obs/Exp	P-value
RS2083637	49	0.98	na	32	0.98	na	21	0.97	na	G/A	124/779/1048	0.40/0.39	0.200
RS2228671	28	0.99	na	26	0.99	na	4	0.99	na	T/C	31/398/1528	0.20/0.21	0.383
RS2240466	51	0.98	na	33	0.98	na	22	0.97	na	A/G	21/364/1565	0.19/0.19	1.000
RS2241213 (RS2338104)	93	0.96	na	76	0.96	na	21	0.97	na	A/G	436/917/554	0.48/0.50	0.130
RS2247056	na	na	na	na	na	na	na	na	na	na	Na	na	Na
RS2255141	10	1.00	na	9	1.00	na	1	1.00	na	T/C	152/858/964	0.43/0.42	0.045
RS2271293	115	0.95	na	96	0.95	na	24	0.96	na	A/G	35/434/1418	0.23/0.23	0.766
RS2277862	8	1.00	na	7	1.00	na	1	1.00	na	T/C	30/530/1416	0.27/0.25	0.013
RS2290159	15	0.99	na	11	0.99	na	5	0.99	na	C/G	84/617/1271	0.31/0.32	0.397
RS2293889	24	0.99	na	16	0.99	na	9	0.99	na	A/C	328/940/699	0.48/0.48	0.708
RS2304130	27	0.99	na	24	0.99	na	4	0.99	na	G/A	18/306/1635	0.16/0.16	0.393
RS2412710	11	1.00	na	10	0.99	na	1	1.00	na	A/G	0/91/1882	0.05/0.05	0.624
RS2479409	16	0.99	na	11	0.99	na	7	0.99	na	G/A	273/867/832	0.44/0.46	0.056
RS2642442	187	0.93	na	150	0.92	na	42	0.94	na	C/T	537/528/768	0.29/0.49	<0.001
RS2650000	88	0.97	na	71	0.96	na	21	0.97	na	A/C	242/816/854	0.43/0.45	0.032
RS2652834	29	0.99	na	21	0.99	na	10	0.99	na	A/G	73/646/1243	0.33/0.32	0.362
RS2737229	8	1.00	na	7	1.00	na	1	1.00	na	G/T	212/861/903	0.44/0.44	0.759
RS2814944	25	0.99	na	17	0.99	na	9	0.99	na	T/C	43/504/1419	0.26/0.26	0.930
RS2814982	10	1.00	na	9	1.00	na	1	1.00	na	A/G	17/373/1584	0.19/0.18	0.394
RS2902940	24	0.99	na	16	0.99	na	9	0.99	na	G/A	197/796/974	0.40/0.42	0.069
RS2923084	24	0.99	na	16	0.99	na	9	0.99	na	G/A	75/601/1291	0.31/0.31	0.610
RS2925979	24	0.99	na	16	0.99	na	9	0.99	na	T/C	153/856/958	0.44/0.42	0.045
RS2929282	11	1.00	na	10	0.99	na	1	1.00	na	A/T	8/187/1778	0.09/0.10	0.241
RS2954029	57	0.98	na	37	0.98	na	24	0.96	na	T/A	424/1002/520	0.51/0.50	0.159
RS2967605	184	0.93	na	150	0.92	na	42	0.94	na	T/C	147/504/1182	0.28/0.34	<0.001
RS2972146	12	1.00	na	8	1.00	na	5	0.99	na	G/T	265/877/833	0.44/0.46	0.155
RS3177928	17	0.99	na	11	0.99	na	7	0.99	na	A/G	34/276/1662	0.14/0.16	<0.001
RS3757354	11	1.00	na	10	0.99	na	1	1.00	na	T/C	85/659/1229	0.33/0.33	0.839
RS3764261	26	0.99	na	17	0.99	na	10	0.99	na	A/C	193/848/925	0.43/0.43	1.000

SNP	CAREMA N=2548		Subcohort N=1983		Cases N=667		HWE in subcohort			P-value	Remarks
	N miss	Success rate	N miss	Success rate	N miss	Success rate	A1/A2	GENO	Obs/Exp		
RS3846662	364	0.86	281	0.86	97	0.85	G/A	355/842/505	0.49/0.50	0.922	Success- rate <90%
RS3905000	137	0.95	114	0.94	30	0.96	T/C	41/376/1452	0.20/0.22	0.007	
RS4129767	13	0.99	8	1.00	6	0.99	A/G	472/971/532	0.49/0.50	0.499	Out of HWE
RS4148008	24	0.99	16	0.99	9	0.99	C/G	189/810/968	0.41/0.42	0.310	
RS4420638	128	0.95	107	0.95	24	0.96	G/A	0/604/1272	0.32/0.27	<0.001	Success- rate <90%
RS4660293	57	0.98	46	0.98	12	0.98	C/T	119/729/1089	0.38/0.37	0.903	
RS4731702	59	0.98	45	0.98	15	0.98	T/C	461/929/548	0.48/0.50	0.084	Out of HWE
RS4765127	18	0.99	13	0.99	6	0.99	T/G	180/887/903	0.45/0.43	0.077	
RS4846914	119	0.95	85	0.96	42	0.94	G/A	323/928/647	0.49/0.49	0.777	Success- rate <90%
RS4939883	175	0.93	139	0.93	43	0.94	T/C	53/495/1296	0.27/0.27	0.495	
RS5756931	247	0.90	199	0.90	90	0.87	G/A	257/844/683	0.47/0.47	0.920	Success- rate <90%
RS6029526	29	0.99	18	0.99	12	0.98	A/T	435/951/579	0.48/0.50	0.238	
RS6450176	21	0.99	13	0.99	10	0.99	T/C	151/732/1087	0.37/0.39	0.081	Success- rate <90%
RS6511720	62	0.98	39	0.98	28	0.96	T/G	39/384/1521	0.20/0.21	0.017	
RS6544713	89	0.97	78	0.96	14	0.98	T/C	172/796/937	0.42/0.42	0.870	Success- rate <90%
RS6754295	83	0.97	73	0.96	12	0.98	G/T	118/698/1094	0.37/0.37	0.621	
RS6756629	22	0.99	20	0.99	3	1.00	T/C	5/242/1716	0.12/0.12	0.345	Success- rate <90%
RS6882076	12	1.00	11	0.99	1	1.00	A/G	266/911/795	0.46/0.46	0.846	
RS6987702	19	0.99	18	0.99	1	1.00	C/T	155/793/1017	0.40/0.40	1.000	Success- rate <90%
RS7134375	9	1.00	7	1.00	2	1.00	A/C	359/926/691	0.47/0.49	0.116	
RS7255436	14	0.99	11	0.99	3	1.00	C/A	490/934/548	0.47/0.50	0.022	Success- rate <90%
RS7395662	49	0.98	39	0.98	12	0.98	A/G	345/899/700	0.46/0.48	0.060	
RS7515577	8	1.00	7	1.00	1	1.00	G/T	100/702/1174	0.36/0.35	0.750	Success- rate <90%
RS7557067	75	0.97	55	0.97	23	0.97	G/A	119/706/1103	0.37/0.37	0.667	
RS7570971	9	1.00	7	1.00	2	1.00	A/C	364/903/709	0.46/0.48	0.011	Success- rate <90%
RS7819412	69	0.97	42	0.98	31	0.95	T/C	466/972/503	0.50/0.50	0.964	
RS7941030	8	1.00	7	1.00	1	1.00	C/T	292/915/769	0.46/0.47	0.474	Success- rate <90%
RS8017377	8	1.00	7	1.00	1	1.00	A/G	418/986/572	0.50/0.50	0.892	

SNP	CAREMA N=2548			Subcohort N=1983			Cases N=667			HWE in subcohort			P-value	Remarks
	N miss	Success rate		N miss	Success rate		N miss	Success rate		A1/A2	GENO	Obs/Exp		
RS9488822	12	1.00		9	1.00		3	1.00		A/T	212/881/881	0.45/0.44	0.722	
RS9686661	10	1.00		9	1.00		1	1.00		T/C	74/609/1291	0.31/0.31	0.828	
RS9987289	15	0.99		12	0.99		4	0.99		A/G	8/283/1680	0.14/0.14	0.337	
RS10195252	na	na		na	na		na	na		na	na	na	na	Abnormal cluster plot
RS10401969	22	0.99		20	0.99		3	1.00		G/A	9/257/1697	0.13/0.13	1.000	
RS10468017	61	0.98		46	0.98		17	0.97		T/C	159/746/1032	0.39/0.40	0.139	
RS10889352	31	0.99		26	0.99		7	0.99		G/A	205/901/851	0.46/0.45	0.155	
RS10903129	110	0.96		89	0.96		24	0.96		A/G	398/902/594	0.48/0.49	0.104	
RS11206510	38	0.99		27	0.99		12	0.98		C/T	77/572/1307	0.29/0.30	0.155	
RS12272004	83	0.97		52	0.97		38	0.94		A/C	9/235/1687	0.12/0.12	0.711	
RS12670798	23	0.99		20	0.99		4	0.99		C/T	108/687/1168	0.35/0.35	0.610	
RS12678919	66	0.97		51	0.97		17	0.97		G/A	15/392/1525	0.20/0.19	0.062	
RS12740374	76	0.97		59	0.97		17	0.97		A/C	85/649/1190	0.34/0.34	0.838	
RS17216525	73	0.97		64	0.97		11	0.98		T/C	14/291/1614	0.15/0.15	0.765	
<i>SNP selected for their association with anthropomorphic traits</i>														
RS6265	146	0.94		121	0.94		32	0.95		A/G	81/610/1171	0.33/0.33	0.888	
RS29939	96	0.96		71	0.96		28	0.96		A/G	186/816/910	0.43/0.43	0.873	
RS925946	12	1.00		11	0.99		1	1.00		T/G	184/743/1045	0.38/0.40	0.003	
RS987237	10	1.00		9	1.00		1	1.00		G/A	69/604/1301	0.31/0.31	1.000	
RS1121980	na	na		na	na		na	na		na	na	na	na	No PCR product
RS1260326	65	0.97		43	0.98		28	0.96		T/C	356/931/653	0.48/0.49	0.457	
RS1558902	8	1.00		8	1.00		0	1.00		A/T	313/852/810	0.43/0.47	<0.001	Out of HWE
RS2568958	72	0.97		54	0.97		20	0.97		G/A	297/922/710	0.48/0.48	0.962	
RS2605100	14	0.99		12	0.99		3	1.00		A/G	193/774/1004	0.39/0.42	0.017	
RS2844479	na	na		na	na		na	na		Na	na	na	Na	No PCR product
RS3764261	26	0.99		17	0.99		10	0.99		A/C	193/848/925	0.43/0.43	1.000	
RS6499640	93	0.96		68	0.97		29	0.96		G/A	341/956/618	0.50/0.49	0.401	
RS6548238	97	0.96		77	0.96		24	0.96		T/C	51/554/1301	0.29/0.28	0.422	
RS7498665	71	0.97		54	0.97		17	0.97		G/A	235/926/768	0.48/0.46	0.094	

SNP	CAREMA N=2548		Subcohort N=1983		Cases N=667		HWE in subcohort			
	N miss	Success rate	N miss	Success rate	N miss	Success rate	A1/A2	GENO	Obs/Exp	P-value
RS7561317	190	0.93	148	0.93	49	0.93	A/G	51/540/1244	0.29/0.29	0.466
RS7647305	121	0.95	86	0.96	41	0.94	A/G	66/606/1225	0.32/0.31	0.421
RS7826222	na	na	na	na	na	na	na	na	na	No PCR product
RS8050136	112	0.96	96	0.95	22	0.97	A/C	280/811/796	0.43/0.46	0.002
RS10146997	7	1.00	7	1.00	0	1.00	G/A	79/602/1295	0.30/0.31	0.385
RS10838738	127	0.95	99	0.95	33	0.95	C/T	180/836/868	0.44/0.43	0.313
RS10913469	187	0.93	150	0.92	45	0.93	G/A	80/539/1214	0.29/0.31	0.049
RS10938397	130	0.95	105	0.95	31	0.95	G/A	384/862/632	0.46/0.49	0.005
RS11084753	151	0.94	120	0.94	36	0.95	A/G	197/813/853	0.44/0.44	0.874
RS12970134	91	0.96	72	0.96	25	0.96	A/G	125/731/1055	0.38/0.38	0.952

HWE, Hardy-Weinberg equilibrium; **SNP**, single nucleotide polymorphism; **N miss**, number of missing genotypes; **A1/A2**, minor allele/major allele; **GENO**, genotype distribution; **Obs/Exp**, observed/expected

Table S3 | Associations between the different genetic risk scores and future CHD in subjects with complete genotype information.

Genetic Risk Scores	N SNPs	Cases (N=272)	Sub-cohort (N=878)	Coronary heart disease		
				HR (95% CI)	HR std* (95% CI)	P
Overall GRS	179	185.4 ± 8.4	182.5 ± 9.4	1.02 (1.01-1.04)	1.25 (1.06-1.46)	0.006
Adjusted for TRF*				1.01 (1.00-1.03)	1.11 (0.96-1.28)	0.152
Risk Factor GRS	153	157.0 ± 8.1	155.3 ± 8.4	1.02 (1.01-1.04)	1.22 (1.05-1.42)	0.011
Adjusted for TRF*				1.01 (0.99-1.03)	1.08 (0.94-1.24)	0.303
CHD GRS	29	31.1 ± 3.2	30.7 ± 3.5	1.05 (1.01-1.09)	1.79 (1.03-1.35)	0.020
Adjusted for TRF*				1.03 (0.97-1.09)	1.12 (0.93-1.36)	0.285
Weighted CHD GRS	29	11.4 ± 1.3	11.1 ± 1.3	1.23 (1.10-1.37)	1.32 (1.13-1.53)	<0.001
Adjusted for TRF*				1.14 (1.01-1.29)	1.19 (1.02-1.40)	0.032
LASSO CHD GRS	3	1.80 ± 0.70	1.63 ± 0.73	1.57 (1.28-1.94)	1.39 (1.20-1.62)	<0.001
Adjusted for TRF*				1.39 (1.13-1.71)	1.19 (1.02-1.40)	0.002
LASSO GRS	14	4.1 ± 1.0	3.7 ± 1.0	1.64 (1.36-1.99)	1.65 (1.36-2.00)	<0.001
Adjusted for TRF*				1.53 (1.24-1.90)	1.54 (1.24-1.91)	<0.001

* Traditional risk factors include sex, current smoking (yes/no), systolic blood pressure, total cholesterol, HDL cholesterol, self-reported diabetes (yes/no), body mass index and parental history of MI (parents with MI, one parent with MI, both parents with MI). Age was used as the time scale variable

* Increase in hazard ratio per standard deviation

HR, hazard ratio; std, standardized; SNPs, single nucleotide polymorphisms; CHD, coronary heart disease; TRF, traditional risk factors; GRS, genetic risk score; LASSO, least absolute shrinkage and selection operator

Table S4 | Weights for the SNPs selected for their association with CHD. The weights are based on the log odds ratio as published in Schunkert et al. 2011 or Peden et al 2011.

SNP	RA	Subcohort		Cases			Weighting factor [*]	Source
		N	RAF	N	RAF	OR		
RS10757278 ¹	G	1832	0.49	623	0.55	1.29	1.00	Schunkert <i>et al.</i> (2011)
RS10953541	C	1902	0.76	641	0.75	1.08	0.30	Peden <i>et al.</i> (2011)
RS11206510	T	1881	0.81	632	0.81	1.08	0.30	Schunkert <i>et al.</i> (2011)
RS1122608	G	1887	0.76	640	0.76	1.14	0.51	Schunkert <i>et al.</i> (2011)
RS11556924	C	1902	0.60	641	0.64	1.09	0.34	Schunkert <i>et al.</i> (2011)
RS12190287	C	1901	0.64	641	0.65	1.08	0.30	Schunkert <i>et al.</i> (2011)
RS12413409	G	1901	0.92	641	0.90	1.12	0.45	Schunkert <i>et al.</i> (2011)
RS12526453	C	1865	0.66	616	0.65	1.10	0.37	Schunkert <i>et al.</i> (2011)
RS12936587	G	1900	0.56	641	0.58	1.07	0.27	Schunkert <i>et al.</i> (2011)
RS1412444	A	1900	0.33	639	0.35	1.09	0.34	Peden <i>et al.</i> (2011)
RS17114036	A	1890	0.91	635	0.92	1.17	0.62	Schunkert <i>et al.</i> (2011)
RS1746048	C	1838	0.86	622	0.87	1.09	0.34	Schunkert <i>et al.</i> (2011)
RS17465637	C	1880	0.73	636	0.75	1.14	0.51	Schunkert <i>et al.</i> (2011)
RS17609940	G	1897	0.81	637	0.81	1.07	0.27	Schunkert <i>et al.</i> (2011)
RS216172	G	1898	0.35	637	0.35	1.07	0.27	Schunkert <i>et al.</i> (2011)
RS2505083 ²	C	1897	0.41	637	0.43	1.05	0.19	Schunkert <i>et al.</i> (2011)
RS2895811	C	1901	0.41	641	0.41	1.07	0.27	Schunkert <i>et al.</i> (2011)
RS3825807	T	1901	0.58	641	0.57	1.08	0.30	Schunkert <i>et al.</i> (2011)
RS4380028	G	1902	0.61	641	0.6	1.07	0.27	Peden <i>et al.</i> (2011)
RS46522	T	1901	0.54	642	0.55	1.06	0.23	Schunkert <i>et al.</i> (2011)
RS4773144	G	1901	0.45	642	0.46	1.07	0.27	Schunkert <i>et al.</i> (2011)
RS579459	C	1890	0.23	633	0.23	1.10	0.37	Schunkert <i>et al.</i> (2011)
RS646776	A	1887	0.79	641	0.82	1.11	0.41	Schunkert <i>et al.</i> (2011)
RS6725887	G	1811	0.13	611	0.13	1.14	0.51	Schunkert <i>et al.</i> (2011)
RS6922269	A	1866	0.26	615	0.27	1.06	0.23	Schunkert <i>et al.</i> (2011)
RS964184	C	1880	0.14	639	0.14	1.13	0.48	Schunkert <i>et al.</i> (2011)
RS974819	A	1901	0.31	642	0.31	1.07	0.27	Peden <i>et al.</i> (2011)
RS9818870	T	1869	0.17	638	0.16	1.12	0.45	Schunkert <i>et al.</i> (2011)
RS9982601	T	1859	0.13	626	0.15	1.18	0.65	Schunkert <i>et al.</i> (2011)

¹ Instead of rs10757278, rs4977574 was selected for replication in Schunkert et al.² Instead of rs2505083, rs7920682 was selected for replication in Schunkert et al.^{*} The weighting factor is determined by dividing the log odds ratio of a SNP by the log odds ratio of rs10757278 (the SNP with the highest log odds ratio).

Table S5 | The association between the 179 successfully genotyped SNPs and incident CHD in the CAREMA study.

SNP	RA	Subcohort		Cases		CHD	P	CHD adjusted*	
		N	RAF	N	RAF	HR (95% CI)		HR (95% CI)	P
SNPs selected for their association with coronary artery diseases									
RS10757278	G	1832	0.49	623	0.55	1.34 (1.17-1.53)	<0.001	1.29 (1.05-1.59)	0.015
RS10953541	C	1902	0.76	641	0.75	0.98 (0.84-1.14)	0.779	0.87 (0.77-0.99)	0.028
RS11206510	T	1881	0.81	632	0.81	0.99 (0.83-1.17)	0.861	1.02 (0.88-1.17)	0.840
RS1122608	G	1887	0.76	640	0.76	1.01 (0.87-1.18)	0.869	1.01 (0.89-1.15)	0.919
RS11556924	C	1902	0.60	641	0.64	1.16 (1.01-1.32)	0.033	1.23 (0.93-1.63)	0.155
RS12190287	C	1901	0.64	641	0.65	1.05 (0.92-1.21)	0.451	1.06 (0.94-1.20)	0.323
RS12413409	G	1901	0.92	641	0.90	0.88 (0.69-1.14)	0.334	0.86 (0.71-1.04)	0.124
RS12526453	C	1865	0.66	616	0.65	0.92 (0.80-1.06)	0.255	1.00 (0.89-1.13)	0.991
RS12936587	G	1900	0.56	641	0.58	1.08 (0.95-1.24)	0.247	1.04 (0.93-1.17)	0.505
RS1412444	A	1900	0.33	639	0.35	1.07 (0.93-1.24)	0.325	1.15 (0.77-1.73)	0.489
RS17114036	A	1890	0.91	635	0.92	1.06 (0.85-1.33)	0.598	0.99 (0.80-1.23)	0.924
RS1746048	C	1838	0.86	622	0.87	1.01 (0.82-1.23)	0.960	1.05 (0.88-1.25)	0.568
RS17465637	C	1880	0.73	636	0.75	1.14 (0.99-1.32)	0.064	1.04 (0.92-1.17)	0.535
RS17609940	G	1897	0.81	637	0.81	0.99 (0.83-1.17)	0.864	0.88 (0.63-1.24)	0.459
RS216172	G	1898	0.35	637	0.35	1.02 (0.88-1.17)	0.809	1.09 (0.93-1.28)	0.286
RS2505083	C	1897	0.41	637	0.43	1.07 (0.93-1.23)	0.322	1.00 (0.89-1.12)	0.942
RS2895811	C	1901	0.41	641	0.41	0.99 (0.87-1.14)	0.905	0.99 (0.89-1.11)	0.864
RS3825807	T	1901	0.58	641	0.57	1.00 (0.87-1.14)	0.956	1.04 (0.93-1.17)	0.447
RS4380028	G	1902	0.61	641	0.60	0.99 (0.87-1.13)	0.890	1.06 (0.95-1.18)	0.337
RS46522	T	1901	0.54	642	0.55	1.10 (0.96-1.26)	0.168	1.07 (0.94-1.21)	0.327
RS4773144	G	1901	0.45	642	0.46	0.99 (0.87-1.13)	0.846	0.99 (0.88-1.11)	0.818
RS579459	C	1890	0.23	633	0.23	0.97 (0.84-1.13)	0.727	0.99 (0.87-1.12)	0.821
RS646776	A	1887	0.79	641	0.82	1.27 (1.09-1.47)	0.002	1.25 (1.07-1.45)	0.004
RS6725887	G	1811	0.13	611	0.13	1.05 (0.85-1.30)	0.637	1.08 (0.90-1.29)	0.395
RS6922269	A	1866	0.26	615	0.27	1.06 (0.91-1.24)	0.431	1.07 (0.92-1.25)	0.386
RS964184	C	1880	0.14	639	0.14	0.96 (0.8-1.170)	0.704	0.80 (0.56-1.14)	0.215
RS974819	A	1901	0.31	642	0.31	1.02 (0.88-1.18)	0.812	0.96 (0.85-1.09)	0.552
RS9818870	T	1869	0.17	638	0.16	0.94 (0.79-1.11)	0.452	0.92 (0.79-1.07)	0.285
RS9982601	T	1859	0.13	626	0.15	1.19 (0.96-1.49)	0.114	1.16 (0.98-1.37)	0.092
SNPs selected for their association with blood pressure									
RS1004467	T	1894	0.91	639	0.90	0.89 (0.70-1.14)	0.347	0.83 (0.69-0.99)	0.038
RS11014166	A	1895	0.66	639	0.68	1.05 (0.91-1.20)	0.526	1.14 (1.00-1.30)	0.051
RS11191548	T	1899	0.92	641	0.91	0.92 (0.72-1.18)	0.523	0.90 (0.75-1.09)	0.282
RS12946454	T	1900	0.26	642	0.24	0.91 (0.79-1.06)	0.218	0.83 (0.68-1.03)	0.084
RS1378942	G	1899	0.32	642	0.34	1.11 (0.96-1.28)	0.163	1.00 (0.88-1.12)	0.937
RS1530440	C	1898	0.18	641	0.18	1.05 (0.88-1.25)	0.622	0.89 (0.77-1.03)	0.112
RS16948048	G	1900	0.38	642	0.39	1.06 (0.93-1.22)	0.390	1.09 (0.97-1.23)	0.132
RS16998073	T	1899	0.29	641	0.31	1.06 (0.91-1.23)	0.460	1.08 (0.95-1.22)	0.250
RS17367504	A	1897	0.15	642	0.15	1.03 (0.85-1.26)	0.735	1.15 (0.98-1.36)	0.082
RS2681472	T	1900	0.84	642	0.84	0.93 (0.77-1.13)	0.464	0.85 (0.73-0.99)	0.038
RS3184504	T	1900	0.51	642	0.50	0.98 (0.86-1.12)	0.731	0.92 (0.81-1.06)	0.241
RS381815	T	1894	0.26	639	0.28	1.16 (0.99-1.36)	0.075	1.22 (1.08-1.38)	0.002

SNP	Subcohort			Cases		CHD		CHD adjusted*	
	RA	N	RAF	N	RAF	HR (95% CI)	P	HR (95% CI)	P
RS9815354	A	1896	0.17	641	0.17	0.99 (0.83-1.17)	0.865	1.06 (0.92-1.22)	0.437
<i>SNPs selected for their association with diabetes type II</i>									
RS10811661	T	1874	0.81	622	0.80	0.94 (0.79-1.12)	0.491	0.91 (0.76-1.08)	0.284
RS10923931	A	1821	0.09	620	0.10	1.16 (0.90-1.50)	0.253	1.14 (0.81-1.59)	0.455
RS10946398	C	1861	0.31	624	0.31	1.01 (0.88-1.17)	0.867	1.06 (0.94-1.20)	0.316
RS11111875	C	1861	0.60	622	0.61	1.03 (0.89-1.18)	0.726	1.11 (0.81-1.53)	0.512
RS12779790	G	1818	0.18	613	0.19	1.04 (0.87-1.25)	0.655	1.08 (0.92-1.25)	0.350
RS13266634	C	1871	0.69	620	0.69	1.02 (0.88-1.18)	0.838	0.99 (0.87-1.12)	0.881
RS2237892	C	1831	0.94	619	0.94	0.94 (0.70-1.27)	0.703	0.80 (0.59-1.09)	0.153
RS2237897	C	1848	0.96	613	0.96	0.98 (0.69-1.39)	0.913	0.89 (0.54-1.45)	0.635
RS4402960	A	1880	0.32	638	0.34	1.12 (0.97-1.30)	0.124	1.11 (0.98-1.25)	0.109
RS4506565	T	1847	0.34	631	0.35	1.01 (0.88-1.17)	0.853	0.94 (0.72-1.22)	0.637
RS4607103	C	1870	0.75	618	0.77	1.12 (0.97-1.30)	0.126	1.09 (0.96-1.24)	0.181
RS4689388	A	1900	0.59	641	0.60	1.07 (0.94-1.23)	0.311	1.15 (0.89-1.49)	0.291
RS4712524	G	1838	0.31	617	0.32	1.04 (0.89-1.20)	0.643	1.09 (0.97-1.24)	0.152
RS5215	G	1856	0.37	627	0.36	0.93 (0.81-1.06)	0.270	1.03 (0.92-1.16)	0.639
RS5219	A	1789	0.36	608	0.36	0.94 (0.82-1.09)	0.417	1.06 (0.95-1.20)	0.305
RS6931514	G	1843	0.27	621	0.28	1.02 (0.88-1.19)	0.801	1.09 (0.96-1.23)	0.174
RS7578597	A	1869	0.89	632	0.89	1.03 (0.83-1.27)	0.806	1.05 (0.87-1.26)	0.608
RS7961581	C	1836	0.30	623	0.31	1.08 (0.93-1.26)	0.295	1.04 (0.91-1.19)	0.541
RS8050136	A	1815	0.36	622	0.40	1.19 (1.04-1.38)	0.014	1.20 (1.07-1.34)	0.002
RS864745	T	1857	0.52	631	0.50	0.91 (0.79-1.04)	0.159	0.90 (0.80-1.02)	0.090
<i>SNPs selected for their association with blood lipid levels</i>									
RS10401969	A	1888	0.93	641	0.92	0.85 (0.64-1.14)	0.279	0.76 (0.44-1.33)	0.338
RS1042034	A	1901	0.78	641	0.75	0.88 (0.74-1.04)	0.123	0.83 (0.73-0.94)	0.005
RS10468017	C	1863	0.72	625	0.70	0.95 (0.81-1.10)	0.460	1.00 (0.88-1.13)	0.965
RS1084651	T	1898	0.85	637	0.85	0.88 (0.72-1.08)	0.234	0.92 (0.77-1.08)	0.302
RS10889352	A	1882	0.66	637	0.68	1.13 (0.98-1.30)	0.082	1.07 (0.95-1.21)	0.249
RS10903129	G	1820	0.55	620	0.53	0.96 (0.84-1.10)	0.594	1.05 (0.93-1.17)	0.437
RS11206510	T	1881	0.81	632	0.81	0.99 (0.83-1.17)	0.861	1.02 (0.88-1.17)	0.840
RS1167998	T	1877	0.66	633	0.68	1.11 (0.97-1.28)	0.138	1.07 (0.95-1.21)	0.256
RS12272004	A	1855	0.07	606	0.06	0.85 (0.66-1.09)	0.193	0.77 (0.60-0.99)	0.043
RS1260326	T	1863	0.42	615	0.44	1.07 (0.93-1.23)	0.343	0.91 (0.72-1.15)	0.426
RS12670798	C	1888	0.23	640	0.24	1.01 (0.86-1.17)	0.949	0.91 (0.80-1.04)	0.173
RS12678919	A	1857	0.89	625	0.90	1.19 (0.96-1.46)	0.107	1.02 (0.84-1.24)	0.831
RS12740374	C	1854	0.79	626	0.82	1.26 (1.08-1.47)	0.003	1.24 (1.07-1.43)	0.004
RS12916	C	1896	0.43	632	0.41	0.96 (0.84-1.10)	0.562	0.96 (0.85-1.09)	0.528
RS1495741	G	1902	0.23	641	0.24	1.05 (0.89-1.24)	0.534	1.09 (0.95-1.24)	0.233
RS1501908	G	1845	0.63	634	0.66	1.10 (0.96-1.26)	0.194	1.11 (0.99-1.25)	0.085
RS1532085	G	1883	0.62	638	0.60	0.93 (0.81-1.06)	0.276	0.95 (0.84-1.06)	0.355
RS1532624	C	1860	0.57	620	0.57	1.00 (0.88-1.15)	0.973	0.94 (0.83-1.05)	0.270
RS1564348	T	1873	0.85	628	0.86	1.05 (0.88-1.25)	0.624	1.13 (0.97-1.32)	0.123
RS157580	A	1844	0.62	630	0.63	1.01 (0.88-1.17)	0.842	0.99 (0.88-1.12)	0.837
RS1689800	G	1893	0.37	637	0.37	1.00 (0.87-1.15)	0.998	0.93 (0.83-1.04)	0.225

SNP	RA	Subcohort			Cases		CHD		CHD adjusted*	
		N	RAF		N	RAF	HR (95% CI)	P	HR (95% CI)	P
RS17216525	C	1846	0.92		633	0.91	1.02 (0.80-1.29)	0.901	0.97 (0.79-1.18)	0.729
RS174547	G	1871	0.33		618	0.33	1.05 (0.90-1.21)	0.555	1.04 (0.62-1.72)	0.896
RS174570	C	1793	0.86		602	0.86	0.95 (0.78-1.16)	0.591	1.02 (0.86-1.20)	0.832
RS1800562	G	1901	0.94		641	0.94	1.03 (0.79-1.34)	0.822	0.77 (0.62-0.96)	0.020
RS1800588	T	1890	0.23		640	0.22	0.97 (0.82-1.13)	0.675	1.00 (0.87-1.16)	0.961
RS1800961	T	1857	0.03		621	0.03	1.15 (0.72-1.85)	0.558	0.80 (0.30-2.14)	0.649
RS181362	T	1890	0.21		634	0.24	1.17 (0.99-1.40)	0.071	1.24 (1.09-1.41)	0.001
RS1883025	A	1836	0.25		609	0.25	1.02 (0.88-1.20)	0.766	0.91 (0.79-1.03)	0.140
RS2000999	A	1901	0.20		642	0.24	1.32 (1.08-1.60)	0.006	1.21 (1.02-1.43)	0.027
RS2068888	G	1896	0.55		637	0.54	0.96 (0.84-1.10)	0.568	0.86 (0.75-0.99)	0.034
RS2072183	G	1890	0.23		635	0.22	1.01 (0.86-1.19)	0.900	0.96 (0.84-1.11)	0.582
RS2075650	G	1859	0.14		624	0.15	1.01 (0.83-1.24)	0.893	1.00 (0.85-1.18)	1.000
RS2081687	T	1901	0.35		641	0.37	1.10 (0.96-1.27)	0.176	1.13 (1.01-1.27)	0.036
RS2083637	A	1874	0.74		622	0.75	1.01 (0.86-1.18)	0.933	0.90 (0.68-1.18)	0.433
RS2228671	C	1883	0.88		640	0.89	1.22 (1.01-1.46)	0.037	1.16 (0.97-1.38)	0.101
RS2240466	G	1873	0.90		621	0.89	0.91 (0.72-1.15)	0.412	0.95 (0.73-1.22)	0.670
RS2255141	T	1902	0.30		641	0.32	1.06 (0.91-1.23)	0.450	1.12 (0.85-1.47)	0.425
RS2271293	A	1817	0.87		620	0.87	1.02 (0.84-1.24)	0.837	0.93 (0.75-1.15)	0.503
RS2277862	C	1901	0.85		642	0.83	0.90 (0.74-1.10)	0.317	0.97 (0.83-1.13)	0.697
RS2290159	G	1895	0.80		637	0.80	0.93 (0.78-1.11)	0.440	0.88 (0.76-1.02)	0.078
RS2293889	A	1890	0.41		635	0.41	1.01 (0.89-1.16)	0.853	1.00 (0.89-1.11)	0.958
RS2304130	A	1886	0.91		640	0.91	1.01 (0.80-1.27)	0.933	0.93 (0.76-1.12)	0.426
rs2241213	G	1834	0.47		623	0.45	0.95 (0.83-1.08)	0.444	0.99 (0.86-1.13)	0.860
RS2412710	A	1901	0.02		641	0.03	1.08 (0.68-1.70)	0.758	0.82 (0.58-1.16)	0.251
RS2479409	G	1895	0.36		635	0.36	0.99 (0.86-1.13)	0.865	0.98 (0.88-1.10)	0.728
RS2650000	A	1839	0.34		623	0.34	1.01 (0.88-1.17)	0.858	1.00 (0.89-1.13)	0.981
RS2652834	A	1886	0.20		633	0.19	0.98 (0.83-1.16)	0.776	0.93 (0.80-1.08)	0.346
RS2737229	T	1901	0.67		642	0.68	1.00 (0.87-1.15)	0.979	0.92 (0.79-1.09)	0.343
RS2814944	T	1890	0.15		635	0.13	0.91 (0.76-1.09)	0.307	0.93 (0.77-1.11)	0.400
RS2814982	G	1902	0.90		641	0.90	0.95 (0.75-1.19)	0.643	0.98 (0.81-1.19)	0.870
RS2902940	A	1890	0.70		635	0.71	1.04 (0.90-1.19)	0.607	1.06 (0.61-1.85)	0.828
RS2923084	G	1890	0.19		635	0.18	0.93 (0.79-1.09)	0.351	0.97 (0.84-1.12)	0.707
RS2925979	T	1890	0.30		635	0.33	1.23 (1.05-1.44)	0.012	1.13 (0.93-1.38)	0.217
RS2929282	A	1901	0.05		641	0.05	0.99 (0.74-1.33)	0.952	1.01 (0.79-1.30)	0.927
RS2954029	A	1870	0.52		619	0.54	1.07 (0.94-1.23)	0.310	1.05 (0.94-1.18)	0.388
RS2972146	T	1898	0.64		637	0.67	1.19 (1.05-1.36)	0.009	1.06 (0.94-1.19)	0.353
RS328	C	1835	0.89		620	0.90	1.17 (0.95-1.44)	0.146	1.03 (0.84-1.26)	0.770
RS3757354	C	1898	0.79		642	0.80	1.05 (0.89-1.23)	0.551	1.00 (0.86-1.17)	0.994
RS3764261	C	1889	0.69		634	0.70	1.09 (0.95-1.25)	0.215	1.06 (0.94-1.20)	0.327
RS3905000	T	1798	0.12		614	0.10	0.87 (0.72-1.06)	0.161	0.86 (0.54-1.39)	0.544
RS4129767	G	1898	0.51		637	0.52	1.01 (0.89-1.16)	0.840	0.97 (0.86-1.09)	0.575
RS4148008	C	1890	0.30		635	0.32	1.13 (0.97-1.31)	0.113	0.97 (0.85-1.10)	0.583
rs6544713	T	1832	0.30		630	0.30	1.06 (0.91-1.23)	0.475	0.96 (0.85-1.08)	0.491
RS439401	C	1838	0.66		602	0.65	0.99 (0.86-1.13)	0.827	0.98 (0.87-1.10)	0.678

SNP	Subcohort			Cases		CHD		CHD adjusted*	
	RA	N	RAF	N	RAF	HR (95% CI)	P	HR (95% CI)	P
RS442177	T	1888	0.59	635	0.59	1.03 (0.90-1.18)	0.622	1.11 (0.99-1.24)	0.086
RS4660293	C	1861	0.25	633	0.25	0.97 (0.83-1.13)	0.686	0.82 (0.72-0.94)	0.003
RS471364	G	1841	0.12	612	0.13	1.08 (0.87-1.34)	0.464	1.08 (0.91-1.28)	0.369
RS4731702	C	1863	0.52	631	0.52	1.01 (0.88-1.15)	0.944	1.01 (0.91-1.13)	0.802
RS4765127	G	1893	0.68	637	0.70	1.04 (0.89-1.20)	0.648	1.02 (0.88-1.17)	0.832
RS4846914	G	1825	0.41	602	0.43	1.08 (0.94-1.25)	0.256	0.98 (0.86-1.11)	0.755
RS492602	G	1899	0.44	640	0.45	1.05 (0.92-1.19)	0.517	1.04 (0.93-1.16)	0.483
RS4939883	C	1777	0.84	602	0.86	1.15 (0.97-1.36)	0.106	1.20 (1.04-1.40)	0.015
RS514230	T	1890	0.52	634	0.54	1.14 (1.00-1.31)	0.051	1.19 (0.94-1.51)	0.151
RS581080	C	1897	0.19	636	0.19	1.01 (0.85-1.19)	0.938	0.97 (0.84-1.12)	0.639
RS6029526	A	1888	0.47	632	0.46	0.97 (0.85-1.11)	0.690	1.01 (0.90-1.13)	0.925
RS629301	A	1897	0.79	639	0.82	1.25 (1.07-1.45)	0.004	1.23 (1.06-1.42)	0.005
RS6450176	T	1893	0.26	633	0.26	0.98 (0.85-1.14)	0.789	0.86 (0.76-0.98)	0.019
RS645040	T	1890	0.79	634	0.79	0.98 (0.83-1.16)	0.808	0.91 (0.80-1.05)	0.195
RS646776	A	1887	0.79	641	0.82	1.27 (1.09-1.47)	0.002	1.25 (1.07-1.45)	0.004
RS6511720	G	1867	0.12	615	0.11	0.80 (0.67-0.95)	0.013	0.85 (0.71-1.01)	0.062
RS6544713	T	1832	0.30	630	0.30	1.06 (0.91-1.23)	0.475	0.96 (0.85-1.08)	0.491
RS6754295	T	1839	0.76	632	0.74	0.92 (0.78-1.08)	0.294	0.85 (0.75-0.97)	0.016
RS6756629	C	1888	0.94	641	0.93	0.97 (0.74-1.27)	0.817	0.81 (0.65-1.01)	0.057
RS6882076	G	1900	0.63	641	0.66	1.11 (0.97-1.27)	0.138	1.13 (1.00-1.27)	0.044
RS693	T	1874	0.48	620	0.48	1.01 (0.89-1.16)	0.860	0.93 (0.83-1.04)	0.191
RS6987702	C	1890	0.72	641	0.75	1.16 (1.01-1.34)	0.038	1.19 (1.04-1.34)	0.008
RS7134375	C	1901	0.59	641	0.60	1.01 (0.88-1.15)	0.912	0.88 (0.79-0.98)	0.019
RS7255436	C	1897	0.49	640	0.51	1.05 (0.93-1.20)	0.439	1.10 (0.89-1.36)	0.375
RS737337	C	1826	0.08	616	0.10	1.12 (0.87-1.45)	0.393	1.11 (0.92-1.33)	0.275
RS7395662	G	1871	0.59	632	0.58	0.89 (0.77-1.02)	0.091	0.88 (0.56-1.38)	0.574
RS7515577	T	1901	0.77	642	0.78	1.05 (0.90-1.23)	0.569	1.03 (0.90-1.18)	0.691
RS7557067	A	1853	0.76	620	0.74	0.92 (0.79-1.08)	0.329	0.87 (0.76-0.99)	0.029
RS7570971	A	1901	0.41	641	0.44	1.09 (0.95-1.24)	0.217	1.14 (1.02-1.27)	0.019
RS7679	C	1807	0.20	622	0.19	1.02 (0.85-1.21)	0.869	1.04 (0.90-1.20)	0.595
RS780093	T	1860	0.41	628	0.44	1.11 (0.96-1.27)	1.530	0.95 (0.85-1.07)	0.426
RS7819412	C	1866	0.51	612	0.49	0.93 (0.81-1.06)	0.271	0.93 (0.82-1.04)	0.194
RS7941030	C	1901	0.38	642	0.39	1.03 (0.89-1.18)	0.725	0.96 (0.86-1.07)	0.463
RS8017377	A	1901	0.46	642	0.46	0.99 (0.86-1.13)	0.855	0.92 (0.82-1.03)	0.158
RS838880	A	1901	0.69	641	0.71	1.03 (0.90-1.19)	0.643	1.01 (0.89-1.14)	0.890
RS9488822	T	1899	0.67	640	0.67	1.00 (0.87-1.15)	0.990	1.09 (0.97-1.23)	0.149
RS964184	C	1880	0.14	639	0.14	0.96 (0.80-1.17)	0.704	0.80 (0.56-1.14)	0.215
RS9686661	T	1902	0.19	641	0.20	1.08 (0.91-1.29)	0.378	1.15 (1.00-1.32)	0.050
RS9987289	A	1896	0.07	639	0.07	0.88 (0.69-1.11)	0.274	0.68 (0.53-0.88)	0.003
<i>SNPs selected for their association with anthropomorphic related traits</i>									
RS10146997	G	1900	0.19	642	0.21	1.11 (0.93-1.32)	0.262	1.07 (0.93-1.23)	0.341
RS10838738	C	1813	0.32	612	0.32	1.05 (0.91-1.22)	0.506	1.02 (0.90-1.16)	0.720
RS10913469	G	1764	0.19	599	0.21	1.19 (0.98-1.44)	0.081	1.29 (1.10-1.51)	0.001
RS10938397	G	1810	0.43	615	0.42	0.94 (0.83-1.08)	0.371	0.96 (0.86-1.07)	0.478

SNP	RA	Subcohort		Cases		CHD		CHD adjusted*	
		N	RAF	N	RAF	HR (95% CI)	P	HR (95% CI)	P
RS11084753	G	1794	0.68	609	0.67	0.94 (0.81-1.09)	0.389	0.81 (0.70-0.94)	0.004
RS1260326	T	1863	0.42	615	0.44	1.07 (0.93-1.23)	0.343	0.91 (0.72-1.15)	0.426
RS12970134	A	1839	0.26	621	0.28	1.10 (0.94-1.29)	0.239	1.01 (0.89-1.15)	0.833
RS2568958	A	1854	0.61	622	0.61	0.96 (0.84-1.11)	0.572	0.95 (0.83-1.10)	0.495
RS2605100	G	1895	0.70	639	0.71	1.04 (0.90-1.19)	0.616	1.10 (0.98-1.24)	0.115
rs29939	G	1837	0.69	615	0.68	0.91 (0.78-1.06)	0.212	0.79 (0.66-0.93)	0.006
RS3764261	C	1889	0.69	634	0.70	1.09 (0.95-1.25)	0.215	1.06 (0.94-1.20)	0.327
RS6265	G	1792	0.79	612	0.78	0.91 (0.76-1.08)	0.267	0.99 (0.86-1.12)	0.825
RS6499640	G	1841	0.57	614	0.57	1.03 (0.89-1.18)	0.728	1.04 (0.93-1.16)	0.551
RS6548238	C	1831	0.83	618	0.82	0.96 (0.80-1.16)	0.694	0.91 (0.79-1.06)	0.225
RS7498665	G	1859	0.36	626	0.38	1.06 (0.92-1.22)	0.455	1.04 (0.90-1.21)	0.562
RS7561317	G	1767	0.83	596	0.82	0.98 (0.82-1.18)	0.858	0.93 (0.80-1.09)	0.368
RS7647305	G	1824	0.81	602	0.80	1.01 (0.85-1.20)	0.902	0.97 (0.84-1.12)	0.671
RS8050136	A	1815	0.36	622	0.40	1.19 (1.04-1.38)	0.014	1.20 (1.07-1.34)	0.002
RS925946	T	1896	0.28	641	0.30	1.14 (0.98-1.32)	0.091	1.29 (1.10-1.51)	0.002
RS987237	G	1899	0.19	641	0.18	0.95 (0.81-1.12)	0.561	0.96 (0.83-1.12)	0.613

SNP, single nucleotide polymorphism; BP, base pair; RA, risk allele; RAF, risk allele frequency; CHD, coronary heart disease; HR, hazard ratio; CI, confidence interval; P, P-value; MI, myocardial infarction; CAD, coronary artery disease

* Adjusted for sex, current smoking, systolic blood pressure, total cholesterol, HDL cholesterol, body mass index, diabetes and family history of MI, age was used as the time scale variable

Table S6a | Reclassification of subjects when a Genetic Risk Score (GRS) composed of all 179 SNPs was used in addition to traditional risk factors (age, sex, current smoking, total cholesterol, HDL cholesterol, BMI, parental history of MI and self-reported diabetes).

Model without Overall GRS	Model with Overall GRS				Total
	<5%	5-<10%	10-<20%	≥20%	
<i>Incident cases</i>					
<5%	191	4	0	0	195
5-<10%	1	126	2	0	129
10-<20%	0	4	76	0	80
≥20%	0	0	0	30	30
Total	192	134	78	30	434
<i>Non-cases</i>					
<5%	1755	7	0	0	1762
5-<10%	11	157	4	0	172
10-<20%	0	3	61	0	64
≥20%	0	0	0	14	14
Total	1766	167	65	14	2012

Table S6b | Reclassification of subjects when a Genetic Risk Score (GRS) composed of the 153 CHD risk factor SNPs was used in addition to traditional risk factors (age, sex, current smoking, total cholesterol, HDL cholesterol, BMI, parental history of MI and self-reported diabetes).

Model without Risk Factor GRS	Model with Risk Factor GRS				Total
	<5%	5-<10%	10-<20%	≥20%	
<i>Incident cases</i>					
<5%	193	2	0	0	195
5-<10%	1	126	2	0	129
10-<20%	0	1	79	0	80
≥20%	0	0	0	30	30
Total	194	129	81	30	434
<i>Non-cases</i>					
<5%	1759	3	0	0	1762
5-<10%	2	169	1	0	172
10-<20%	0	1	63	0	64
≥20%	0	0	0	14	14
Total	1761	173	64	14	2012

Table S6c | Reclassification of subjects when a Genetic Risk Score (GRS) composed of the 29 SNPs previously associated with CHD was used in addition to traditional risk factors (age, sex, current smoking, total cholesterol, HDL cholesterol, BMI, parental history of MI and self-reported diabetes).

Model without CHD GRS	Model with CHD GRS				Total
	<5%	5-<10%	10-<20%	≥20%	
<i>Incident cases</i>					
<5%	190	5	0	0	195
5-<10%	4	116	9	0	129
10-<20%	0	4	73	3	80
≥20%	0	0	4	26	30
Total	194	125	86	29	434
<i>Non-cases</i>					
<5%	1751	11	0	0	1762
5-<10%	18	150	4	0	172
10-<20%	0	5	56	3	64
≥20%	0	0	1	13	14
Total	1769	166	61	16	2012

Table S6d | Reclassification of subjects when a Genetic Risk Score (GRS) composed of the 29 weighted CHD risk SNPs was used in addition to traditional risk factors (age, sex, current smoking, total cholesterol, HDL cholesterol, BMI, parental history of MI and self-reported diabetes).

Model without weighted CHD GRS	Model with weighted CHD GRS				Total
	<5%	5-<10%	10-<20%	≥20%	
<i>Incident cases</i>					
<5%	183	12	0	0	195
5-<10%	3	114	12	0	129
10-<20%	0	9	68	3	80
≥20%	0	0	4	26	30
Total	186	135	84	29	434
<i>Non-cases</i>					
<5%	1748	14	0	0	1762
5-<10%	22	144	6	0	172
10-<20%	0	8	51	5	64
≥20%	0	0	1	13	14
Total	1770	166	58	18	2012

Table S6e | Reclassification of subjects when a weighted Genetic Risk Score (GRS) composed of 3 SNPs selected using LASSO regression on the 29 SNPs previously associated to CHD was used in addition to traditional risk factors (age, sex, current smoking, total cholesterol, HDL cholesterol, BMI, parental history of MI and self-reported diabetes).

Model without weighted CHD GRS	Model with weighted CHD GRS				Total
	<5%	5-<10%	10-<20%	≥20%	
<i>Incident cases</i>					
<5%	180	15	0	0	195
5-<10%	11	104	14	0	129
10-<20%	0	10	65	5	80
≥20%	0	0	5	25	30
Total	191	129	84	30	434
<i>Non-cases</i>					
<5%	1743	19	0	0	1762
5-<10%	29	134	9	0	172
10-<20%	0	10	52	2	64
≥20%	0	0	2	12	14
Total	1772	163	63	14	2012

Table S6f | Reclassification of subjects when a weighted Genetic Risk Score (GRS) composed of 14 SNPs selected using LASSO regression on all 179 SNPs was used in addition to traditional risk factors (age, sex, current smoking, total cholesterol, HDL cholesterol, BMI, parental history of MI and self-reported diabetes).

Model without LASSO GRS	Model with LASSO GRS				Total
	<5%	5-<10%	10-<20%	≥20%	
<i>Incident cases</i>					
<5%	173	22	0	0	195
5-<10%	15	83	31	0	129
10-<20%	1	17	53	9	80
≥20%	0	0	4	26	30
Total	189	122	88	35	434
<i>Non-cases</i>					
<5%	1730	32	0	0	1762
5-<10%	47	113	12	0	172
10-<20%	0	17	42	5	64
≥20%	0	0	4	10	14
Total	1777	162	58	15	2012

Table S8 | Information about the subsets of SNPs, their coefficients and the appropriate weighting factors. The subsets of SNPs were selected using 10-fold cross-validated LASSO regression on only the 29 SNPs previously associated with CHD and all 179 SNPs.

SNP	Locus	Nearby genes	Coefficient	weighting factor*	Previously associated with:
<i>LASSO regression on the 29 CHD SNPs</i>					
RS10757278	9p21.3	CDKN2A, CDKN2B	0.160	1.000	CHD
RS11556924	7q32.2	ZC3HC1	0.038	0.238	CAD
RS646776	1p13.3	CELSR2, PSRC1, SORT1	0.041	0.256	MI
<i>LASSO regression on all 179 SNPs</i>					
rs10757278	9p21.3	CDKN2A, CDKN2B	0.159	1.000	CHD
rs2925979	16q23.2	CMIP	0.066	0.414	HDL
rs6882076	5q33.3	TIMD4	0.006	0.039	TC
rs2954029	8q24.13	TRIB1	0.001	0.007	TG
rs6987702	8q24.13	TRIB1	0.011	0.067	TC
rs10889352	1p31.1	DOCK7	0.023	0.142	TC
rs2972146	2q36.3	IRS1	0.050	0.316	HDL-C
rs11556924	7q32.2	ZC3HC1	0.031	0.198	CAD
rs514230	1q42.3	IRF2BP2	0.021	0.135	TC
rs8050136	16q12.2	FTO	0.063	0.394	DM2, DMI
rs181362	22q11.21	UBE2L3	0.015	0.096	HDL
rs646776	1p13.3	CELSR2, PSRC1, SORT1	0.065	0.406	MI, LDL-C
rs925946	11p14.1	BDNF	0.022	0.136	BMI
rs2000999	16q22.2	HPR	0.128	0.804	TC

* The weighting factor is set to 1 for the SNP with the highest coefficient

Supplement IV

Table S1 | Association between baseline δ -5 desaturase activity and incident coronary heart disease according to rs174547 genotypes

rs174547 (# cases)	Quintile of δ -5 desaturase activity ¹					P value for trend ²
	First	Second	Third	Fourth	Fifth	
AA (n = 205)	1 ³	0.61 (0.27-1.41)	0.35 (0.15-0.79)	0.36 (0.16-0.80)	0.25 (0.11-0.54)	<0.0001
	1 ⁴	0.60 (0.25-1.43)	0.35 (0.15-0.83)	0.48 (0.21-1.11)	0.35 (0.15-0.81)	0.022
	1 ⁵	0.63 (0.26-1.53)	0.40 (0.17-0.97)	0.55 (0.24-1.27)	0.44 (0.19-1.04)	0.087
AG/GG (n = 276)	1 ³	0.69 (0.46-1.03)	0.62 (0.38-1.00)	0.54 (0.31-0.93)	1.15 (0.57-2.33)	0.027 ⁶
	1 ⁴	0.62 (0.39-0.98)	0.71 (0.42-1.21)	0.72 (0.40-1.30)	1.64 (0.76-3.53)	0.463
	1 ⁵	0.63 (0.40-1.02)	0.75 (0.44-1.27)	0.78 (0.43-1.41)	1.86 (0.84-4.12)	0.649

¹ δ -5 desaturase activity was assessed by the ratio of C20:4n-6 to C20:3n-6 in plasma cholesteryl esters.

² From models with desaturase activity included as a continuous variable.

³ Model was adjusted for age, sex, systolic blood pressure, hypertensive medication use, current smoking, and diabetes.

⁴ Model was adjusted for age, sex, systolic blood pressure, hypertensive medication use, current smoking, diabetes, total cholesterol, and high-density lipoprotein cholesterol.

⁵ Model was adjusted for age, sex, systolic blood pressure, hypertensive medication use, current smoking, diabetes, total cholesterol, high-density lipoprotein cholesterol, and percentages of C22:6n-3 (DHA) in plasma cholesteryl esters.

⁶ Hazard ratio per unit desaturase activity (95% confidence interval) = 0.89 (0.80-0.99).

Table S2 | Association of rs174547 with incident coronary heart disease (CHD) risk

	rs174547 genotype			rs174547 G allele ¹
	AA	AG	GG	
Incident CHD, n	234	259	56	-
Model 1 ²	1.0	1.00 (0.80-1.26)	0.96 (0.67-1.37)	0.99 (0.84-1.16)
Model 2 ³	1.0	1.15 (0.88-1.49)	0.91 (0.60-1.39)	1.02 (0.85-1.22)

¹ The model assumes that each copy of the G allele contributes equally to coronary heart disease risk.

² Model 1 was adjusted for age and sex.

³ Model 2 was adjusted for age, sex, systolic blood pressure, hypertensive medication use, current smoking, diabetes, total cholesterol, and high-density lipoprotein cholesterol.

Table S3 | Association between baseline n-6 PUFA in plasma cholesteryl esters (precursors and products of $\delta 5$ - or $\delta 6$ -desaturase) and incident coronary heart disease (CHD)

	Quintile of C18:2n-6 (linoleic acid)					<i>P</i> value for trend ¹
	First	Second	Third	Fourth	Fifth	
Incident CHD, n	127	125	102	97	86	
Model 1 ²	1	1.11 (0.80-1.54)	1.02 (0.72-1.44)	1.10 (0.78-1.56)	0.80 (0.57-1.13)	0.247
Model 2 ³	1	1.16 (0.81-1.66)	1.27 (0.88-1.83)	1.30 (0.89-1.89)	0.91 (0.62-1.32)	0.861
Model 3 ⁴	1	1.09 (0.74-1.60)	1.34 (0.91-1.96)	1.38 (0.92-2.09)	1.01 (0.68-1.48)	0.395
Model 4 ⁵	1	1.09 (0.74-1.61)	1.34 (0.91-1.96)	1.39 (0.92-2.09)	1.01 (0.69-1.49)	0.389
	Quintile of C18:3n-6					<i>P</i> value for trend ¹
	First	Second	Third	Fourth	Fifth	
Incident CHD, n	93	109	100	107	128	
Model 1 ²	1	1.05 (0.74-1.50)	0.89 (0.62-1.28)	0.92 (0.64-1.32)	1.00 (0.71-1.41)	0.706
Model 2 ³	1	1.12 (0.76-1.64)	0.90 (0.61-1.33)	1.02 (0.69-1.50)	0.93 (0.64-1.37)	0.867
Model 3 ⁴	1	1.13 (0.75-1.72)	0.90 (0.58-1.38)	1.01 (0.66-1.53)	1.05 (0.70-1.57)	0.605
Model 4 ⁵	1	1.14 (0.75-1.73)	0.90 (0.58-1.39)	1.02 (0.67-1.56)	1.06 (0.70-1.61)	0.536
	Quintile of C20:3n-6					<i>P</i> value for trend ¹
	First	Second	Third	Fourth	Fifth	
Incident CHD, n	88	80	107	125	137	
Model 1 ²	1	0.98 (0.68-1.43)	1.14 (0.80-1.64)	1.29 (0.91-1.82)	1.39 (0.99-1.96)	0.024
Model 2 ³	1	0.91 (0.61-1.34)	1.06 (0.73-1.55)	1.14 (0.78-1.66)	1.43 (1.00-2.05)	0.011
Model 3 ⁴	1	0.94 (0.61-1.45)	1.13 (0.75-1.69)	1.00 (0.67-1.50)	1.11 (0.75-1.67)	0.494
Model 4 ⁵	1	0.95 (0.62-1.47)	1.15 (0.76-1.73)	1.03 (0.67-1.57)	1.14 (0.75-1.74)	0.420
	Quintile of C20:4n-6 (arachidonic acid)					<i>P</i> value for trend ¹
	First	Second	Third	Fourth	Fifth	
Incident CHD, n	127	107	120	79	104	
Model 1 ²	1	0.89 (0.63-1.24)	0.96 (0.70-1.34)	0.65 (0.45-0.92)	0.76 (0.54-1.06)	0.031
Model 2 ³	1	0.86 (0.60-1.22)	0.92 (0.65-1.31)	0.64 (0.43-0.94)	0.77 (0.53-1.10)	0.088
Model 3 ⁴	1	1.06 (0.72-1.56)	1.09 (0.74-1.60)	0.79 (0.51-1.21)	0.86 (0.58-1.28)	0.163
Model 4 ⁵	1	1.04 (0.70-1.55)	1.07 (0.73-1.58)	0.76 (0.48-1.20)	0.83 (0.53-1.28)	0.150

¹ From models with fatty acids included as continuous variables.² Model 1 is adjusted for age and sex.³ Model 2 is adjusted for age, sex, systolic blood pressure, hypertensive medication use, current smoking, and diabetes.⁴ Model 3 is adjusted for all covariates in model 2, total cholesterol, and high-density lipoprotein cholesterol.⁵ Model 4 is adjusted for all covariates in model 3 and baseline n-3 PUFA in plasma cholesteryl esters.

Table S4 | Association of baseline C20:5n-3 (EPA) and C22:6n-3 (DHA) in plasma cholesteryl esters with incident coronary heart disease (CHD)

	Quintile of C20:5n-3 (EPA)					P value for trend ²
	First (0.00) ¹	Second (0.28)	Third (0.38)	Fourth (0.52)	Fifth (0.83)	
Incident CHD, n	96	100	104	114	123	
Model 1 ³	1	0.81 (0.57-1.16)	0.88 (0.62-1.26)	0.82 (0.57-1.16)	0.82 (0.58-1.17)	0.348
Model 2 ⁴	1	0.86 (0.59-1.24)	0.91 (0.62-1.33)	0.88 (0.61-1.27)	0.76 (0.52-1.11)	0.243
Model 3 ⁵	1	0.90 (0.60-1.34)	0.90 (0.59-1.39)	0.79 (0.52-1.20)	0.89 (0.58-1.35)	0.733
Model 4 ⁶	1	0.90 (0.60-1.35)	0.90 (0.59-1.39)	0.80 (0.52-1.22)	0.89 (0.58-1.35)	0.724

	Quintile of C22:6n-3 (DHA)					P value for trend ²
	First (0.00) ¹	Second (0.20)	Third (0.27)	Fourth (0.34)	Fifth (0.46)	
Incident CHD, n	89	139	93	112	104	
Model 1 ³	1	1.45 (1.03-2.05)	0.87 (0.60-1.25)	1.07 (0.74-1.53)	0.93 (0.65-1.34)	0.286
Model 2 ⁴	1	1.49 (1.05-2.13)	0.84 (0.58-1.24)	1.13 (0.77-1.67)	0.86 (0.58-1.29)	0.268
Model 3 ⁵	1	0.96 (0.65-1.43)	0.61 (0.41-0.92)	0.89 (0.59-1.34)	0.65 (0.42-0.99)	0.049
Model 4 ⁶	1	0.95 (0.64-1.40)	0.57 (0.38-0.87)	0.82 (0.53-1.26)	0.59 (0.37-0.93)	0.027

¹ Median percentages of EPA and DHA in each quintile are listed between brackets.² From models with fatty acids included as continuous variables.³ Model 1 is adjusted for age and sex.⁴ Model 2 is adjusted for age, sex, systolic blood pressure, hypertensive medication use, current smoking, and diabetes.⁵ Model 3 is adjusted for all covariates in model 2, total cholesterol, and high-density lipoprotein cholesterol.⁶ Model 4 is adjusted for all covariates in model 3 and baseline n-6 PUFA in plasma cholesteryl esters.

Supplementary Note S1

Analysis of intermediate factors of coronary heart disease (CHD)

To explore whether δ -5 desaturase activity might exert its protective effect against CHD through known intermediate risk (or protective) factors of CHD (total and HDL cholesterol levels, systolic blood pressure, diabetes mellitus, EPA and DHA), we regressed these factors against δ -5 desaturase activity with adjustment for age and sex. It seems some of these intermediate factors (HDL cholesterol levels, EPA and DHA) might be involved in the protective effects of δ -5 desaturase activity against CHD risk (See Table S5).

Table S5 | Association between intermediate risk (protective) factors for coronary heart disease and δ -5 desaturase activity at baseline in a random subcohort (n = 1263).¹

Covariates	<i>B</i>	<i>P</i>
Total cholesterol (mmol/L) ²	-0.021 ± 0.011	0.076
HDL cholesterol (mmol/L) ²	0.020 ± 0.003	< 0.0001
Systolic blood pressure (mmHg) ²	-0.29 ± 0.17	0.078
EPA (%) ²	0.044 ± 0.004	< 0.0001
DHA (%) ²	0.023 ± 0.002	< 0.0001
Diabetes mellitus (Yes/no) ³	-0.020 ± 0.12	0.86

¹ δ -5 desaturase activity was assessed by the ratio of C20:4n-6 to C20:3n-6 in plasma cholesteryl esters.

² *B* and *P* are β regression coefficients $\pm$ SE and corresponding *p* values obtained by using a regression model adjusted for age and sex.

³ *B* and *P* are β regression coefficient $\pm$ SE and corresponding *p* value obtained by using a logistic model adjusted for age and sex.

Dankwoord

Het is zeven jaar geleden dat ik naar Leiden kwam. Nu heb ik mijn proefschrift afgerond en is het tijd voor een nieuwe stap in mijn carrière. Eén van de dingen die ik geleerd heb tijdens die zeven jaren is dat je wetenschap samen doet en niet alleen. Daarom wil ik graag diegenen bedanken die het voor mij mogelijk hebben gemaakt dit proefschrift te schrijven.

Om te beginnen bij mijn promotor Professor Eline Slagboom. Hartelijk bedankt voor de tijd die jij in mij hebt gestoken om dit proefschrift tot een goed einde te brengen. Daarnaast wil ik mijn co-promotor Bas Heijmans bedanken voor zijn inzet en de discussies die hebben geleid tot beter proefschrift.

Zonder statistiek geen proefschrift. Voor de statistische analyses en het schrijven van R-scriptjes was Stefan een grote steun voor mij. Ook wil ik Jolanda bedanken voor haar uitleg over het case-cohort design en de CAREMA studie.

Zonder data geen proefschrift. Dennis en Eka bedankt voor jullie steun en hulp op het lab. Aswin, Sibel, Axel en André ik heb jullie hulp met betrekking tot het uitvoeren van de $^1\text{H-NMR}$ experimenten en het interpreteren van de spectra zeer gewaardeerd.

Claudia, onze samenwerking heeft geleid tot een beter artikel en daar ben ik heel blij mee. Yingchang thank you for your collaboration on this NHS project which resulted in this thesis. Ook wil ik mijn collega's bedanken, niet alleen voor de wetenschappelijke discussies, maar ook voor de gezelligheid bij ons op de afdeling. Ik denk dat de stelling "Werken en feesten maakt schoone geesten" zeker van toepassing is op onze afdeling.

Ana ik vond het erg leuk om jouw paranimf te zijn tijdens je promotie. Hierdoor weet ik al een beetje hoe het voelt om als promotie kandidaat aan tafel te zitten.

Naast werken moet er ook tijd zijn voor ontspanning. Die ontspanning heb ik gevonden in de vele leuke uitjes die ik samen met mijn vrienden heb ondernomen.

Acrobatiek doe je samen, niet alleen. De gezelligheid van mijn mede-acrobaten en hun belangstelling voor mijn proefschrift heb ik zeer gewaardeerd.

Ook wil ik mijn ouders en broer bedanken voor hun steun en interesse tijdens mijn promotietraject.

Annemieke, zonder jou had ik Bob nooit leren kennen en dit brengt mij bij mijn allerlaatste bedankje. Dankjewel lieve Bob voor jouw steun tijdens het schrijven van dit proefschrift.

Curriculum Vitae

Anika Vaarhorst was born on 3th of March 1981 in Enschede, the Netherlands. She attended secondary school in Holten at the Waardenborch and passed her exams in 1999. In that same year she started her study Health Sciences at Maastricht University, the Netherlands. Her first internship was at the department Farmacology and Toxicology at Maastricht University. Here she studied the effect of poly(ADP-ribose)polymerase-1 inhibition on NAD⁺ depletion in human lung cells. A part of these results were published in *Biochem Biophys Res Commun.* 2005;338(4):1805-10. Her main internship was conducted at the department of Health Risk Analysis and Toxicology at Maastricht University where she studied the role of high glucose levels on the production of reactive oxygen species using an in vitro model. After graduating in 2004, she started her master Epidemiology in 2005 at Maastricht University. During that study she conducted her internship at the department of Epidemiology at Maastricht University where she studied the role of pet exposure in the development of atopic diseases in children between 0 and 2 years old. In 2006 she received her degree. In 2007 she started her PhD at the department of Molecular Epidemiology at the Leiden University Centre under the supervision of Professor P. Slagboom and Dr. B.T. Heijmans. Her main aim was to find novel metabolomics and genetic risk markers for incident coronary heart disease. The results of this research are outlined in this thesis.

List of publications

Vaarhorst AAM*, Verhoeven A*, Weller CM*, Böhringer S, Göraler S, Meissner A, Deelder AM, Henneman P, Gorgels APM, van den Brandt PA, Schouten LJ, van Greevenbroek MM, Merry AHH, Verschuren WMM, van den Maagdenberg AMJM, Willems van Dijk K, Isaacs A, Boomsma D, Oostra BA, van Duijn CM, Jukema JW, Boer JMA, Feskens E, Heijmans BT, Slagboom PE. A metabolomics profile is associated with the risk of incident coronary heart disease. Accepted for publication in Am Heart J.

Draisma HH, Beekman M, Pool R, van Ommen GJ, Adamski J, Prehn C, **Vaarhorst AA**, de Craen AJ, Willemsen G, Slagboom PE, Boomsma DI. Familial resemblance for serum metabolite concentrations. Twin Res Hum Genet. 2013;16(5):948-61

Lu Y, **Vaarhorst A**, Merry AH, Dollé ME, Hovenier R, Imholz S, Schouten LJ, Heijmans BT, Müller M, Slagboom PE, van den Brandt PA, Gorgels AP, Boer JM, Feskens EJ. Markers of endogenous desaturase activity and risk of coronary heart disease in the CAREMA cohort study. PLoS One. 2012;7(7):e41681

Vaarhorst AA, Lu Y, Heijmans BT, Dollé ME, Böhringer S, Putter H, Imholz S, Merry AH, van Greevenbroek MM, Jukema JW, Gorgels AP, van den Brandt PA, Müller M, Schouten LJ, Feskens EJ, Boer JM, Slagboom PE. Literature-based genetic risk scores for coronary heart disease: the Cardiovascular Registry Maastricht (CAREMA) prospective cohort study. Circ Cardiovasc Genet. 2012;5(2):202-9

Slagboom PE, Beekman M, Passtoors WM, Deelen J, **Vaarhorst AA**, Boer JM, van den Akker EB, van Heemst D, de Craen AJ, Maier AB, Rozing M, Mooijaart SP, Heijmans BT, Westendorp RG. Genomics of human longevity. *Philos Trans R Soc Lond B Biol Sci*. 2011;366(1561):35-42

Vaarhorst AA, Beekman M, Suchiman EH, van Heemst D, Houwing-Duistermaat JJ, Westendorp RG, Slagboom PE, Heijmans BT; Leiden Longevity Study (LLS) Group. Lipid metabolism in long-lived families: the Leiden Longevity Study. *Age (Dordr)*. 2011;33(2):219-27

Lu Y, Feskens EJ, Boer JM, Imholz S, Verschuren WM, Wijmenga C, **Vaarhorst A**, Slagboom E, Müller M, Dollé ME. Exploring genetic determinants of plasma total cholesterol levels and their predictive value in a longitudinal study. *Atherosclerosis*. 2010;213(1):200-5

Moonen HJ, Geraets L, **Vaarhorst A**, Bast A, Wouters EF, Hageman GJ. Theophylline prevents NAD⁺ depletion via PARP-1 inhibition in human pulmonary epithelial cells. *Biochem Biophys Res Commun*. 2005;338(4):1805-10

*These authors contributed equally to this work