FRONT PAGE

PROJECT FINAL REPORT

Grant Agreement number: 305422

Project acronym: COMBI-BIO

Project title: Development of COMBInatorial BIOmarkers for subclinical atherosclerosis

Funding Scheme: SME-targeted Collaborative Project (small-scale focused research project)

Period covered: from 01/10/2012 to 30/09/2014

Name of the scientific representative of the project's co-ordinator[1], Title and Organisation:

Professor Paul Elliott, Imperial College London

Tel: 0044 (0) 20 7594 3328

Fax:

E-mail: p.elliott@imperial.ac.uk

Project website: http://www.combi-bio.eu

[1] Usually the contact person of the coordinator as specified in Art. 8.1. of the Grant Agreement.

4.1 Final publishable summary report

Executive summary

The overall aim of this project was to seek biomarkers of sub-clinical atherosclerosis using novel ways of analysing human blood serum, specifically by assessing the small molecule metabolite composition. The project brought together world-leading expertise in cardiology, biostatistics, bioinformatics, metabolic phenotyping and cardiovascular disease (CVD) epidemiology.

A total of 8,000 serum samples was amassed from three existing epidemiological cohorts (LOLIPOP study in UK, Rotterdam study in the Netherlands and MESA study in USA) along with a wealth of participant data such as age, gender, ethnicity, smoking, blood pressure, and various other risk factors for CVD. Crucially, each participant had been assessed using two measures of pre-symptomatic arterial disease – coronary artery calcification (CAC) and intima-media thickness (IMT).

The metabolic phenotyping was achieved using a combination of proton nuclear magnetic resonance (NMR) spectroscopy (measuring three complementary data sets) and ultra-performance-liquid-chromatography linked to mass spectrometry (measured under four different conditions to target different combinations of metabolites). The resulting data represent one of the largest coherent sets of metabolic phenotyping data ever accumulated. All of the analytical goals were achieved within the tight 24 month schedule by a combination of parallel instrumentation and dedication of the involved staff. Initially, it was planned to analyse 4,000 samples in a discovery mode and use the metabolic phenotyping of a further 4,000 samples to validate the prior results. However, by combining all 8,000 metabolic phenotypes by NMR into one large data set, the improvement in statistical power was such that this approach was agreed and adopted.

An array of metabolic biomarkers were associated significantly with CAC and IMT. Highest priority was given to biomarkers common to all three cohorts, using a statistical model that took into account known CVD risk factors (including, age, gender, blood cholesterol values, smoking, blood pressure, statin use and blood pressure medication) such that any extra variation in CAC or IMT ascribed to the metabolic component could be viewed as explanation beyond the classical Framingham heart disease risk factors. A statistical model that allowed for age, gender, ethnicity, and cohort +/- phase of the study was also calculated as this would be analogous to helping explain the variation in CAC or IMT described by the Framingham risk factors. This showed many associated metabolic biomarkers that are being investigated for biochemical pathways to derive information on the underlying mechanisms of atherosclerosis. To this end, correlations between the metabolic biomarkers and genome-wide data have been used to help identify such pathways and to aid the characterisation of metabolic features not easily ascribed to particular metabolites.

A wide range of dissemination activities has been planned. These include publications in high impact journals in the cardiovascular and epidemiological areas, and papers on the new methodologies that have been necessarily developed in order to be able study such large sample cohorts, plus presentations at international conferences, and outreach activities to a lay audience. In addition, patent lawyers have been engaged to investigate the feasibility of filing patents on the new intellectual property that has been invented.

We have achieved all of the milestones and deliverables that were set out for the project, and it has led to discovery of a range of biomarkers for sub-clinical atherosclerosis and links to established risk factors for heart disease. The results obtained, knowledge generated and methodologies developed have potential to aid the European research effort in this fast moving area for better understanding of the biochemical mechanisms underlying the development of atherosclerotic plaques, and hence possible new avenues of therapy, for the medical benefit of humankind generally.

Summary description of project context and objectives

Background and Context of the COMBI-BIO project

Cardiovascular disease is the leading cause of mortality worldwide, and is a late manifestation of the pathophysiological processes leading to development of atherosclerosis, the deposition of fatty and fibrous material on the artery walls. Atherosclerosis starts early in life and remains asymptomatic for many years, often decades. When the disease becomes symptomatic, atherosclerotic disease is usually advanced; death from Coronary Heart Disease (CHD) is often sudden before medical care is available. However, CHD is largely preventable. There are two complementary approaches to tackle the CHD epidemic. First is the public health approach that aims to reduce the overall population risk of CHD by lifestyle changes (e.g., diet, exercise, smoking cessation); the second is targeted at high-risk individuals who can benefit from personalised therapeutic and lifestyle interventions to reduce their CHD risk.

Currently there are no readily available and reliable early markers of atherosclerosis for use in routine clinical practice. Two promising candidate markers of subclinical atherosclerosis, coronary artery calcium (CAC) detected by Computerised Tomography and carotid intima-media thickness (IMT) detected on ultrasound, both have limitations that have so far prevented them from being adopted into routine clinical care. While CAC has been shown, in particular, to improve prediction of future risk of CHD, there are limitations, including cost, radiation dose and limited availability in clinical settings. Carotid IMT is also predictive of cardiovascular disease, but is time-consuming, requires specific training and expertise, and is not widely used in routine clinical practice.

Current risk stratification algorithms are most applicable later on in the disease process, and less so for younger individuals, who nonetheless are still at risk as the atherosclerotic process starts in adolescence or young adulthood.

The concept is that ***early*** identification of people at highest risk of atherosclerosis will give opportunity for ***early intervention*** (lifestyle/pharmacologic), with potential to halt or even reverse the disease process. The General Aim of the COMBI-BIO project is to use a systems biology approach, specifically metabolic profiling and computational medicine, to develop novel combinatorial biomarkers and risk scores for subclinical atherosclerosis, as a means of ***early detection*** and stratified patient management for people at risk of CHD and cardiovascular disease.

Systems biology involves the analysis of relationships among the elements in a biological system, viewed as an integrated and interacting network of genes, proteins and biochemical reactions, and its response to genetic or environmental perturbations. We have been proponents of a “top-down” systems biology approach, using minimally invasive methods, to capture the properties of systemic homeostasis and its dysregulation through application of multivariate metabolic profiling technologies via metabolomics – defined as ‘‘the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification’’. COMBI-BIO adopted such a “top down” systems approach to the analysis and interpretation of data and samples from three leading international epidemiological resources with measurements of subclinical atherosclerosis (CAC and IMT), an important precursor in the pathophysiological pathway leading to clinically manifest CHD and cardiovascular disease.

In proof-of-concept studies, we pioneered the concept of the Metabolome-wide Association Study (MWAS) to discover novel biomarkers of cardiovascular disease risk using data and biological

samples from large-scale epidemiological resources [1]. The MWAS approach capitalises on the advances in metabolomic technologies such as proton Nuclear Magnetic Resonance (^{1}H NMR) and mass spectrometry (MS) coupled with advances in computational and statistical approaches to reveal latent data dependencies and spectral inter-correlations. It involves the untargeted (discovery) analysis of biological samples using high-throughput metabolic profiling to detect novel biomarkers in relation to disease risk [1]; this was the approach we adopted in COMBI-BIO.

With ^{1}H NMR, there is underrepresentation of metabolites at low concentrations and these can be identified and quantified using more sensitive MS methods; these give good coverage of metabolites in terms of chemical class and concentration range and provide a complementary approach to ^{1}H NMR. Metabolic profiling by MS can use either a targeted or untargeted strategy. In COMBI-BIO we used untargeted screening by both ^{1}H NMR and ultra performance liquid chromatography MS (UPLC-MS) to maximize biomarker discovery combined with innovative approaches to describe, model, and visualise the data. This involved the integration of NMR and MS 'omics' data (as well as genome wide (GWAS) data) to investigate biochemical pathways and regulatory networks, and thus gain biological insights into the pathophysiological processes underlying associations of metabolic profiles with atherosclerosis.

Aims and Objectives

Our ***General Aim*** was to use a systems biology approach via metabolic profiling to develop novel combinatorial biomarkers and risk scores for subclinical atherosclerosis – as a means of early detection and stratified patient management for people at risk of CHD and cardiovascular disease. Since atherosclerotic disease remains subclinical for many years, efforts to predict development and progression are urgently needed to guide early prevention and provide insights into disease aetiology and pathogenesis.

Our ***Specific Aims*** were:

1) To use systems biology approaches, based on metabolic profiling and computational medicine, to discover, test and validate novel biomarkers for subclinical atherosclerosis;
2) To use cross-platform (NMR, MS) and multi-omics (genome-wide, metabolome-wide) analyses to investigate underlying biochemical connectivities and pathways, and hence to advance understanding of aetiopathogenesis of atherosclerosis development and progression;
3) To develop prognostic combinatorial biomarkers and risk scores to improve early prediction and patient stratification/management for subclinical atherosclerosis.

Cohorts in COMBI-BIO and their biological and epidemiological resources

The consortium investigated biomarkers related to subclinical atherosclerosis (CAC and IMT) using data and samples from three well-established ethnically diverse epidemiological cohorts in both Europe and the USA (see Table 1). These were the LOLIPOP Study (U.K.), the Rotterdam Study (the Netherlands) and the MESA Study (U.S.A.). Extensive data and stored serum samples from 8,000 individuals with CAC and IMT measurements were made available for the research.

The following table gives an overview of the samples and epidemiological resources provided by each study:

Cohort description	LOLIPOP	ROTTERDAM	MESA
Recruitment	Recruited 2002 to 2008, from 58 General Practices in W London	Recruited 1990 and 2000, living in a suburb of Rotterdam	Recruited 2000 to 2002, living near 6 US field centers
Age	ages 35-74	ages 55+	ages 45-84
Population	N= 17,606 Indian Asian origin: 7,766 European origin, other groups 3,911	N= 7983 (1990), N=3011 (2000)	N= 6,814 multi-ethnic origin: white (38%), US African (28%), Hispanic (22%), US Chinese (12%)
Measurements	CAC and IMT measures, covariates, physical and biological measurements	CAC and IMT measures, covariates, physical and biological measurements	CAC and IMT measures, covariates, physical and biological measurements
Samples provided	2,000	2,000	4,000

Table 1 Cohort descriptions

The COMBI-BIO study design and Work Packages

The Work Packages (WPs) are summarized in the figure below:

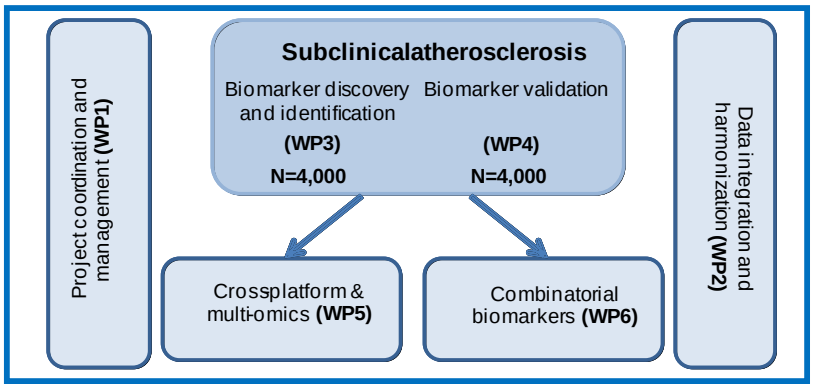

Figure 1 COMBI-BIO study design and Work Packages

- **WP1** created a robust and flexible management structure within COMBI-BIO to support the execution and coordination of the project as a whole.
- **WP2** brought together the data and stored serum samples from the three cohorts (8,000 individuals) and harmonized across cohorts to ensure comparability of data and analyses.
- **WP3** comprised analysis by 600 MHz ^{1}H NMR spectroscopy and UPLC-MS of serum samples from 4,000 individuals with CAC and IMT measurements, complemented by a suite of statistical/chemometric techniques, for biomarker discovery.
- **WP4** comprised ^{1}H NMR and UPLC- MS analysis of serum from a further 4,000 individuals with CAC and IMT measurements, analysis of metabolomic features vs CAC and IMT, and structural identification of associated metabolites.
- **WP5** involved the construction of multiple association networks within and between analytical platforms (NMR/MS), generating networks of associated metabolites; and use of multi-omics computational approaches to investigate GWAS-metabolome associations.

- **WP 6** used robust methodology to investigate predictive ability of a combinatorial set of available traditional measures together with novel biomarkers (WPs 3 and 4) with respect to subclinical atherosclerosis.

The COMBI-BIO Consortium

The consortium is multidisciplinary, and brought together leading groups in systems medicine, biomarker discovery, epidemiology, cardiology, critical evaluation and development of CHD/cardiovascular disease risk functions, and computational medicine. The Project Lead was Prof. Paul Elliott (Imperial College) who is highly experienced in leading international projects and consortia.

The research is SME (small and medium sized enterprises)-driven. The SME Metabometrix Ltd. is a small Imperial College spin-out company formed over ten years ago, and a bespoke supplier of metabolic profiling services (metabolomics) to industry and academia. It has independent analytical (NMR and MS) and computational facilities, which were made available to the project.
The table below gives an overview of the project partners and their contribution to COMBI-BIO.

Project Partner	Specialist Expertise	Lead Investigator	Cohorts
Imperial College of Science, Technology and Medicine	Project management and coordination, biostatistics, epidemiology	Prof Paul Elliott	West London Lifesciences Population Study (LOLIPOP)
Metabometrix Ltd.	Systems biology, systems medicine and metabolic profiling, biomarker discovery and validation, computational medicine	Prof John Lindon	
University of Ioannina	Development, statistical evaluation and validation of the combinatorial biomarkers and risk scores	Prof John Ioannidis	
Erasmus Universiteit Rotterdam	Epidemiology, risk factor modelling	Prof Albert Hofmann	Rotterdam Study
Helmholtz Zentrum, Munich	Analysis of multi-omics data (genome-wide, metabolome-wide), pathway modelling	Dr Christian Gieger	
Northwestern University, Chicago	Evaluation of novel risk markers of cardiovascular risk	Prof Philip Greenland	Multi Ethnic Study of Atherosclerosis (MESA)

Table 2 COMBI-BIO project partners

References

1. Holmes E et al, *Nature* 2008;453:396-400.

Description of the main S&T results/foregrounds

The COMBI-BIO project was planned to ensure delivery of the project objectives within tight timelines. An overview of the project tasks colour coded by their Work Packages is depicted below.

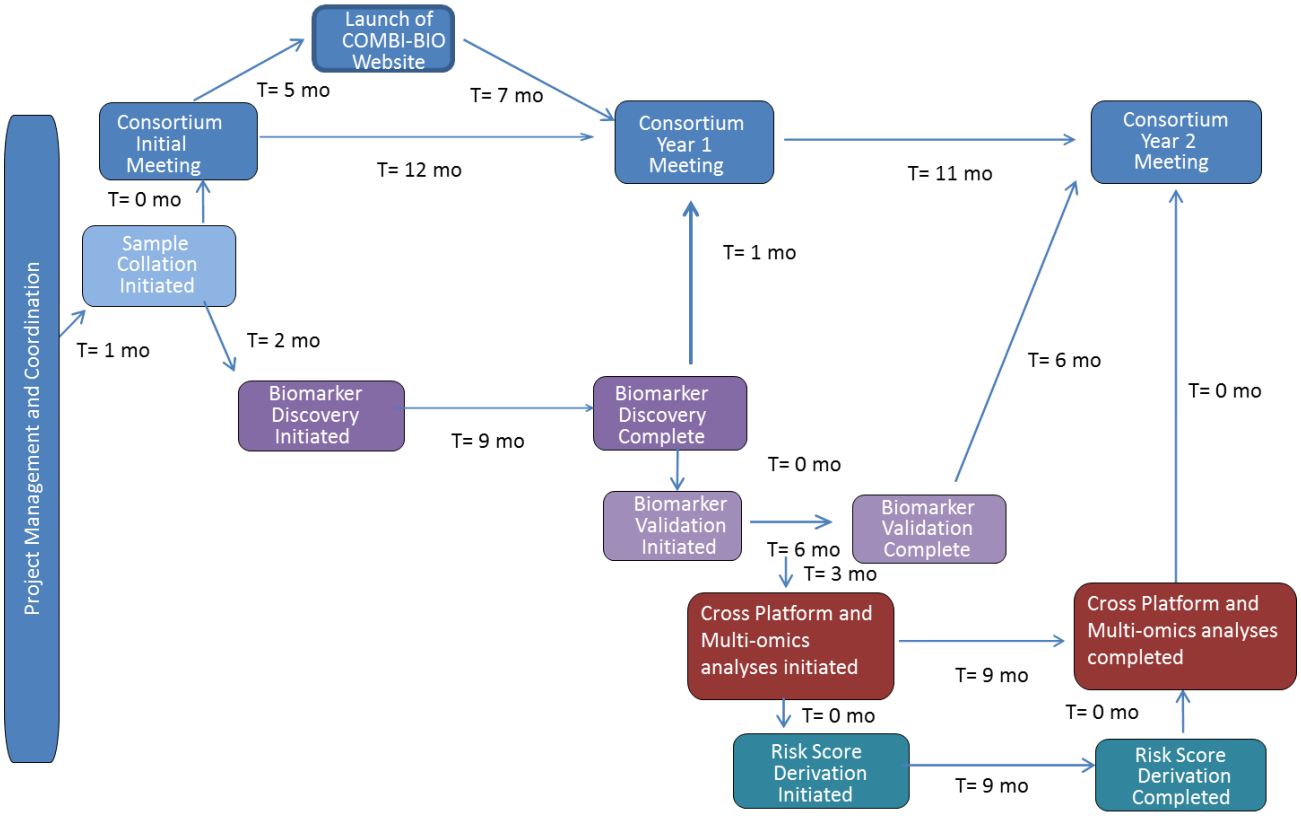

Figure 2 The COMBI-BIO Pert diagram

The project was delivered through six distinct, but integrated Work Packages as described in the previous section. WP1 provided the management of the project. The work in the remaining WPs is described below.

WP2 – DATA INTEGRATION AND HARMONISATION

The main objective of WP2 was to integrate the wealth of data generated within the project, and provide the other WPs with harmonised data for subsequent analysis and for biological interpretation. As such WP2 worked in close collaboration with all WPs.

Receipt and storage of serum samples from participating cohorts and generated metabolomic data

Procedures for shipment of samples from the cohorts to Metabometrix Ltd were agreed and samples were despatched frozen. Standard Operating Procedures (SOPs) for sample shipment, sample preparation and storage were made publicly available via the COMBI-BIO website.

Samples from the three cohorts were stored in lockable alarmed freezers, monitored 24/7. All samples were logged and given unique identifiers. The information on the newly received samples was entered into the Metabometrix Ltd freezer inventory. Samples were randomised and assigned a new number (Metabometrix ID) prior to metabolic analysis to prevent any systematic bias in the

analyses. Samples were aliquoted for NMR and MS and study-specific quality control samples were generated. Samples were stored at -80°C pending NMR/MS data acquisition.

All raw and processed spectroscopic data were stored securely; the raw analytical data (data directly off the instrument) produced in WP3 and WP4 were stored in two separate locations, fully backed-up, and made available to the project consortium members. Processed data were uploaded to an active database, accessible by all partners.

The raw data generated by the NMR and UPLC-MS assays were exported to a MATLAB platform with standardised IDs comprising the following information: cohort, assay, phase, rack number and unique sample ID. Pre-processing workflows for both NMR and UPLC-MS datasets were implemented and employed on the raw data. The pre-processed data sets were exported in ASCII format for further statistical analysis.

Receipt and storage of epidemiological data from participating cohorts and data harmonisation
The epidemiological data were requested from all three cohorts using a specially designed template. All cohorts provided the data in this format. The distributions of all variables were checked and there was no need for further data cleaning. Datasets with similar naming of the variables were created to facilitate the epidemiological analyses.

Intima-media thickness (IMT) is a measurement of the thickness of the innermost two layers of the arterial wall. It is measured by ultrasound to detect the presence of atherosclerotic plaques in carotid arteries. After a teleconference among the three cohorts, we decided to include measures of *common carotid IMT* which was available for all participants.

Coronary Artery Calcium (CAC) is a non-invasive measure of subclinical coronary atherosclerosis. It is measured by CT scan to detect the presence of calcium in plaques on the walls of the heart. The statistical distribution of this measurement was highly skewed in our populations (many participants with zero or small amounts of calcium but few participants with very high amounts). Therefore, we applied some scaling correction to the data (i.e. log transformation) and used binary and categorical variables for CAC in our statistical models.

- CAC was recoded using three transformations: logarithmic (log(CAC+1)), binary (CAC_binomial), or coded in three clinically relevant classes of CAC (0; -100, -1,000; CAC_clinical).
- IMT was analysed using either the raw values or log-transformed values.

Harmonisation of variables concerning established risk factors for cardiovascular disease (CVD) was done after receipt of the epidemiological data. The scaling of different biochemical measures was harmonised across cohorts to make the measurements as comparable as possible for all participants.

A list of eligible variables for use in the risk prediction analyses was created and availability of variables was checked in each cohort. The list included basic demographic information, Framingham risk score (FRS) variables and a list of emerging cardiovascular biomarkers based on recent recommendations. The list also included information on the outcomes of interest (CAC and IMT), comorbidities and family history of disease. Table 3 below summarises the data obtained from each cohort.

	LOLIPOP	Rotterdam	MESA
Demographic/ socioeconomic			
Age	X	X	X
Sex	X	X	X
Socioeconomic status		X	X
Education		X	X
Ethnicity	X	X	X
Traditional CVD risk factors			
Smoking (current, past, never)	X	X	X
Total cholesterol	X	X	X
LDL cholesterol	X	X	X
HDL cholesterol	X	X	X
Triglycerides	X	X	X
Systolic blood pressure	X	X	X
Diastolic blood pressure	X	X	X
Fasting glucose	X	X	X
Weight	X	X	X
Height	X	X	X
Waist circumference	X	X	X
Hip circumference	X	X	X
Emerging biomarkers			
CRP		X	X
BNP		X	X
Fibrinogen			x (antigen)
D-dimer			X
ICAM-1		X	
VCAM-1		X	X
PAI-1		X	
Lipoprotein a		X	
Albumin	X		x (urinary)
Apolipoprotein AI		X	x
Apolipoprotein B			x
Uric acid	X		
Homocysteine			X
Chlamydia pneumoniae			X
Cystatin C			x
Comorbidities			
Self reported/ doctor diagnosis of diabetes	X	X	X
Previous suffered MI or stroke	X	X	
Self reported/ doctor diagnosis of peripheral arterial disease		x (defined based on ABI)	x (defined based on ABI)

Disease prevalence and history of disease			
Parent or mother with heart disease before 65 years	x	X	X
Family history of CVD	X	X	X
Family history of diabetes	X	X	X
Medications			
Currently on medication for BP	X	X	X
Currently on medication for diabetes	X	X	X
Currently on medication for lipids	X	X	X
Outcomes			
CAC	X	X	X
IMT right	X	X	X
IMT left	X	X	X
Carotid plaque	X	X	X
Carotid plaque density	X	X	X
Carotid plaque maximum lesion width	X	X	X

Table 3 Epidemiological data obtained from each cohort

X: data available. Blank cells indicate data not available. CAC Coronary artery calcium; IMT intima-media thickness; ABI Ankle-Brachial Index

WP3 – BIOMARKER DISCOVERY AND IDENTIFICATION

This WP was led by the SME Metabometrix Ltd. We used ^{1}H NMR spectroscopy and MS platforms, operating in untargeted mode for biomarker discovery.

The main tasks in this WP were:

- To carry out metabolic profiling for novel metabolic biomarker detection
  - ^{1}H NMR untargeted analysis of serum samples from 4,000 individuals with measurements of atherosclerosis (CAC and IMT)
  - UPLC-MS untargeted analysis of the same serum samples.
- Multivariate chemometric interrogation of the data to characterize discriminatory novel metabolic biomarkers.
- Structural identification of unknown discriminatory metabolites.

It was agreed to analyse a further 4,000 samples by NMR (to increase the statistical power) within the existing budget as part of the validation phase (WP4). This major addition to the project deliverables was agreed by the project General Assembly.

NMR spectroscopic analyses of blood serum samples from the three population cohorts

Serum samples were received from the three cohorts, logged and stored frozen as outlined in WP2. The specific NMR data sets comprised –

- a standard 1-D spectrum (so-called NOESY) showing resonances from all proton-containing molecules in the sample, including broad, largely undefined bands from serum proteins, sharper and well-defined bands from serum lipoproteins (with some classification into their main groups), and sharp peaks from a range of small molecule metabolites such as amino acids, simple carbohydrates, organic acids, organic bases and a number of osmolytes;
- Carr-Purcell-Meilboom-Gill (CPMG) spectrum that attenuates the peaks from the macromolecules and allows better definition of the small molecules;
- a 2-D (J-resolved) spectrum that separates the various NMR bands and their complex splittings into orthogonal directions while editing out the macromolecule peaks.

The analyses above resulted in one of the largest comprehensively profiled datasets using NMR spectroscopy ever compiled. The data produced by the NMR instruments required extensive pre-processing prior to further statistical analysis. This pre-processing stage included development and optimisation of new techniques for cross-cohort alignment and 'binning' of the NMR spectral intensity to reduce dimensionality, based on the correlation structure between spectral signals. The SOP for the NMR analysis was made publicly available via the COMBI-BIO website.

NMR data handling, processing and harmonisation

For both CPMG and NOESY NMR data, in-house written MATLAB (Mathworks Inc., USA) functions were utilised for phasing and baseline correction of each sample spectrum including the calibration of chemical shifts using the glucose doublet at δ 5.23. The trimethylsilyl propionate (TSP) signal at δ 0.00 was not used because the TSP peak was influenced by sample pH.

Six NMR data tables were generated representing spectra from the three cohorts analysed in two phases. These data tables were then concatenated in order to have one large data table consisting of ~8,000 samples (including the additional 4,000 samples measured by NMR in WP4) and 34,001 variables (δ 0.500–9.000). Prior to spectral peak alignment, the region δ 4.400–5.100 corresponding to H_2O resonances was removed and the data table was divided into six slices on the ppm axis because of computational issues arising from the high computer memory demand. The ends of the slices were selected from the noisy regions without metabolite signals. Spectral peak alignment on each slice was performed by the Recursive Segment-wise Peak Alignment (RSPA) algorithm and thereafter the slices were concatenated back to the full ppm range. In Figure 3, spectral peak alignment of the ~8,000 samples is shown. Aligning the peaks using the RSPA algorithm resulted in better peak position distributions for specific peaks. In order to evaluate spectral peak alignment, alignment quality measures (aq_{bin}) were calculated for bin sizes of δ = 0.02 and 0.08 ppm for unaligned (0.3299 and 0.4460 respectively for CPMG, 0.5437 and 0.7447 for NOESY) and aligned (0.3988 and 0.5242 for CPMG, 0.5915 and 0.7816 for NOESY) data. Higher values of $aq_{0.02}$ and $aq_{0.08}$ in the aligned compared to the unaligned data indicated successful spectral peak alignment. Subsequently, selected regions such as δ 1.180–1.240, δ 2.244–2.261, δ 3.375–3.400 and δ 3.660–3.710, where peaks for suspected contaminants occur, were removed from the spectra.

The remaining spectral regions were normalised by probabilistic quotient normalisation using the median spectrum as the reference. The normalised high resolution spectra contained 30,590 data points for both CPMG and NOESY datasets. To decrease the number of variables and account for minor misalignments, 'binning' was applied by grouping correlated spectral peaks. The Statistical Recoupling of Variables (SRV) algorithm was used which generates bins by searching correlated structures throughout the full spectrum. After binning, the number of variables for both CPMG and NOESY datasets decreased to ~500.

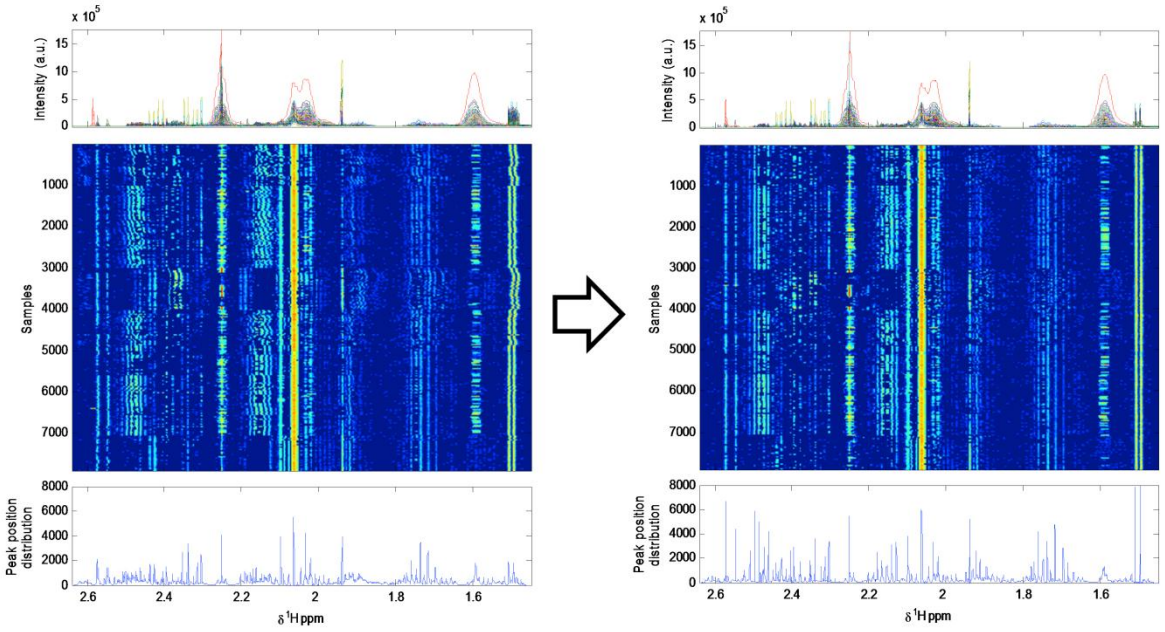

Figure 3 Illustration of the spectral peak alignment procedure. Note that spectra have not been normalised so systematic differences in intensity between the cohorts were corrected later in the pipeline.

The variation between data tables due to differences between the cohorts and phases was removed by mean-centring each variable in each data table. This approach was applied to both high resolution and binned spectra.

Analysis of NMR data with respect to CAC and IMT measurements

We used univariate linear models coupled with a multiple testing correction strategy controlling the family wise error rate (FWER) to investigate associations with the sub-clinical atherosclerosis measures (CAC and IMT). We developed a strategy for declaring statistical significance of the observed associations which took account of the multivariate nature of the data and the correlation structure within metabolomic datasets. Specifically we derived a per-test significance level (α') that accounted for the correlation across the metabolomic covariates, using an in-house permutation procedure. The Effective Number of Tests (ENT) was estimated, defined as the number of independent tests that would be required to obtain the same significance level by using Bonferroni corrections. The ENT implicitly measures the level of dependency within the data.

The permutation procedure used to calculate the per-test significance level was based on 10,000 resamples. The permutation procedure used to calculate the per-test significance level was based on 10,000 resamples. For each resample the P-value distribution over the (~30K) tests performed gives the null P-value distribution and the minimal P-value defines the significance level above which at least one false positive conclusion would be expected. The distribution of that statistic was estimated over the 10,000 permutations to infer the per-test significance threshold controlling the FWER at the desired level. The estimated α' depended on the type of NMR analysis (NOESY or CPMG), the model covariates (Model 1 or Model 2, see Table 4 below) and outcome (CAC or IMT), Table 5.

NMR Models*	Adjustments
Model 1	Adjusted for age, gender, cohort, phase and ethnicity
Model 2	Adjusted as in Model 1 plus low density lipoprotein (LDL) and high density lipoprotein (HDL) cholesterol, systolic blood pressure, lipid and blood pressure lowering treatment, smoking status and diabetes

Table 4 Model covariates used in the statistical analyses
*For LC-MS models: no adjustment for phase since the two phases were not combined; adjustment for cohort only for the features matched across cohorts

Spectrum	N data points	Outcome	α'	ENT
NOESY	30,590	Log_{10}(IMT)	1.86×10^{-05}	2,684
NOESY	30,590	Log(CAC+1)	1.81×10^{-05}	2,766
CPMG	30,590	Log_{10} (IMT)	3.36×10^{-06}	14,861
CPMG	30,590	Log(CAC+1)	3.80×10^{-06}	13,156

Table 5 Metabolome Wide Significance Level (MWSL) derived from a permutation-based approach for CAC and IMT, NOESY and CPMG. Results for Model 2 only are presented. See Table 4 for definition of Model 2.

Typical NMR spectra are shown in Figure 4 below for all three NMR data acquisitions. The top left plot is the NOESY data showing peaks from all hydrogen-containing molecules in the sample, the top right is the CPMG spectrum in which peaks from macromolecules are attenuated, emphasising the small molecule metabolite peaks. At the bottom is the 2-dimensional J-resolved spectrum that is used to aid metabolite identification by splitting chemical shifts (in ppm) from J-coupling multiplets (in Hz) for better resolution.

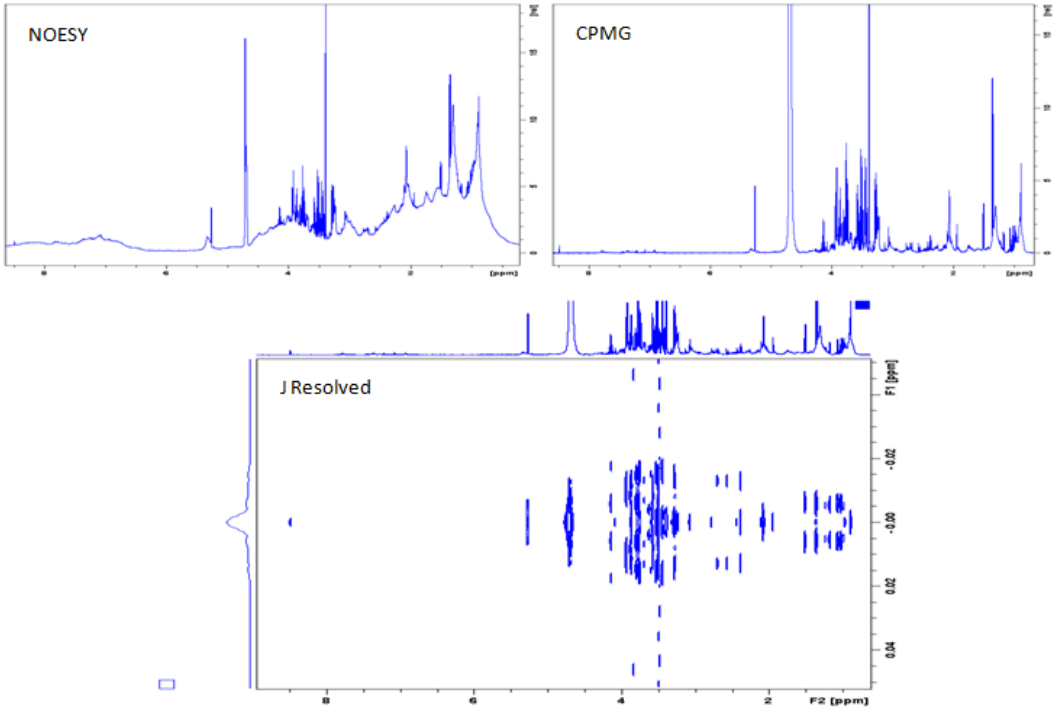

Figure 4 Typical NOESY, CPMG and J-resolved NMR spectra of serum from a COMBI-BIO sample

A small number of data sets were outliers, due, for example, to instrument malfunction during measurement. These were all investigated and flagged as such in the database. Multivariate data analysis also showed minor differences in metabolic profiles between the samples from three cohorts. Nevertheless, data sets from all three cohorts were combined for subsequent analysis (for detais see the report on WP3).

To quantify the potential per-cohort heterogeneity we ran a series of analyses in each cohort separately (adjusting for age, gender, and ethnicity). Our results indicated that the number of findings was greater and more stable in the largest population (MESA). We iteratively sub-sampled in MESA the same number of participants as in the two other cohorts (N≈1,000) and ran an MWAS on each sub sample. Our results showed that the greater number of significant findings in MESA could mainly be attributed to its larger sample size and hence greater statistical power (Figure 5). This suggested that pooling data from the three cohorts (with the appropriate additional adjustment by study and ethnicity) would increase statistical power and therefore facilitate the discovery of novel signals.

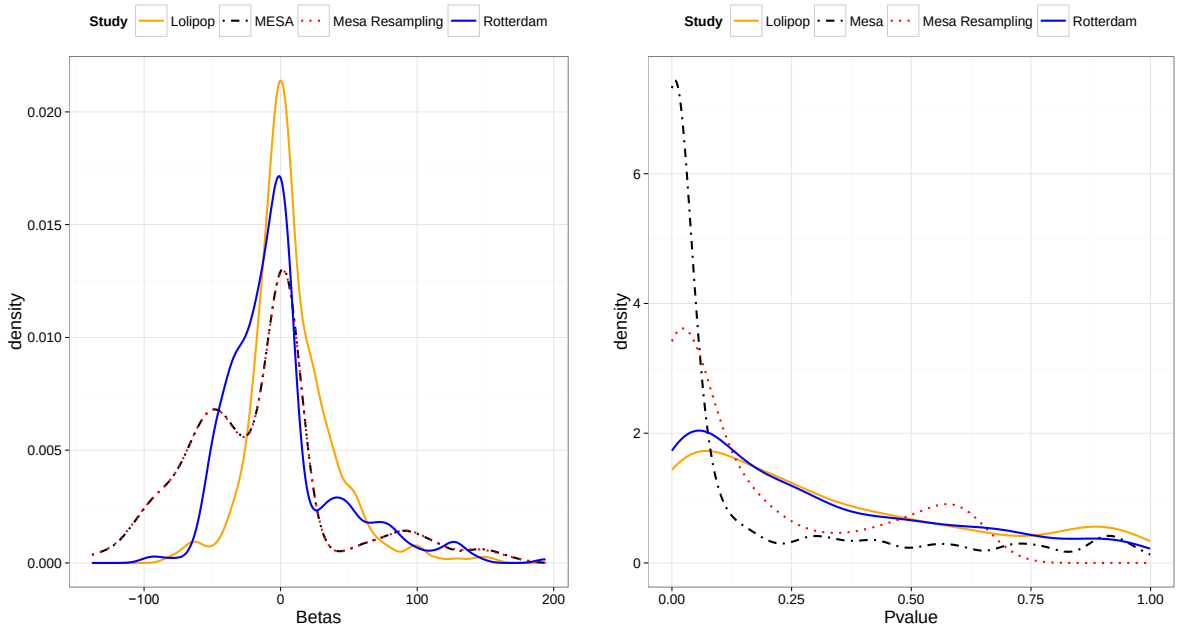

Figure 5 Density estimates of the effect size estimates (left panel) and corresponding strength of the association (right panel) for the MWAS of log(CAC+1). Results are presented for NOESY, adjusted age, gender, ethnicity, phase, cohort (Model 1) in LOLIPOP (orange), MESA (black) and Rotterdam (blue). The red lines represent the median value across resampling of size N=1,000 from MESA done 2,500 times.

To maximise discovery, we decided to pool the NMR data across cohorts, using raw peak intensity, and log(CAC+1) and $\log_{10}$(IMT) as main outcomes. We defined two main models as shown in Table 4. As expected, additional adjustment for the CVD risk factors in Model 2 resulted in fewer significant signals. The signals that were no longer significant when moving from Model 1 to Model 2 identified numerous spectral features associated with the CVD risk factors, which might give insights into possible mechanisms and pathways.

Results from the reconciled model based on phase 1 samples (N=4,000) are summarised in Table 6.

	Log(CAC+1)		Log_{10}(IMT)	
	Model 1	Model 2	Model 1	Model 2
NOESY	4,386	81	5,767	13
CPMG	916	3	1,424	0

Table 6 Summary of the number of significant spectral associations identified for log(CAC+1) and log_{10}(IMT) using phase 1 data for Models 1 and 2 (see Table 4 for definition). Results are presented for both outcomes, and two NMR spectral types.

We then ran the same NMR assays on all 8,000 biosamples included in phase 1 (4,000 samples) and phase 2 (4,000 samples) in COMBI-BIO. Several strategies were considered to integrate data from the two phases without inflating the number of false positive findings, which could arise from technical variation across phases. This included a unified pre-processing procedure based on all the spectra and the implementation of a series of stratified analyses by phase, cohort, and ethnic group (Figure 6).

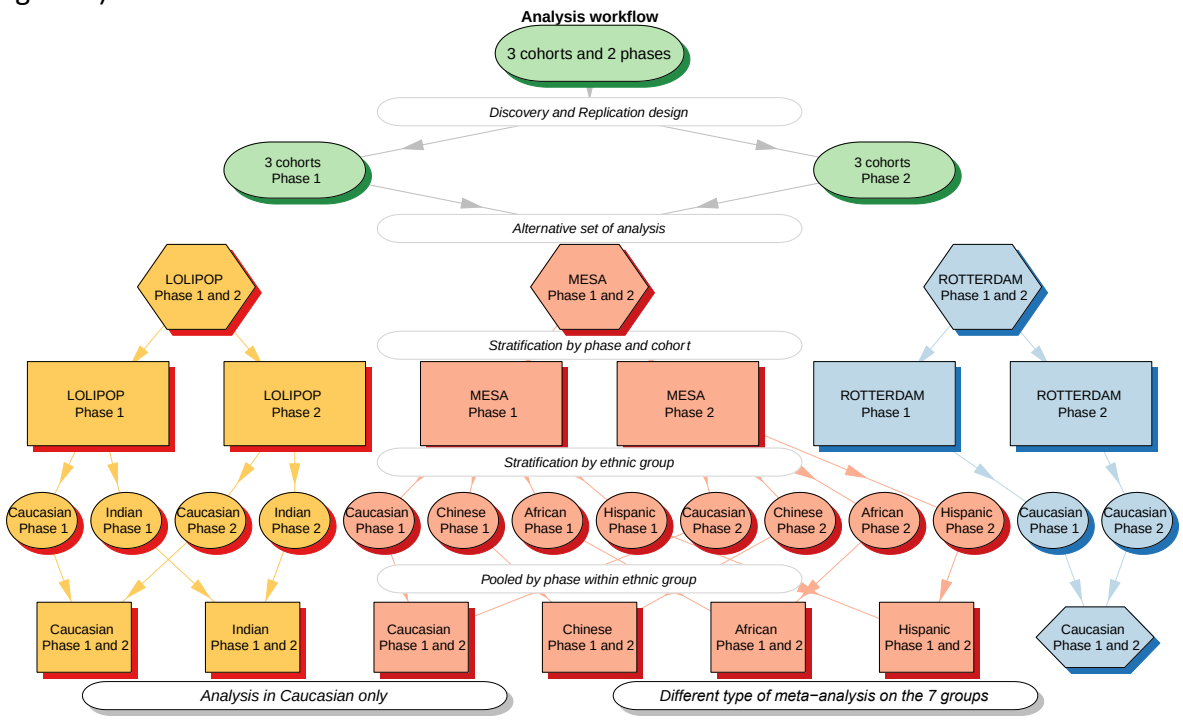

Figure 6 Schematic representation of the NMR stratification procedure.

Our analyses showed that the factor primarily affecting the sensitivity of our strategy was sample size. The number of significant findings vs CAC and IMT is shown in WP4**.**

UPLC-MS sample preparation and analyses

Each sample was prepared according to standard protocols optimised for serum. In brief, both hydrophilic and lipophilic substances were extracted separately and subjected to UPLC separation using a UPLC column optimised for lipidomics and a HILIC column for polar molecules; MS analysis was conducted in both positive and negative ion mode, so each sample provided 4 data sets.

In-house, on-line and commercial databases were used to identify novel metabolic biomarkers, as well as reference to standard libraries (e.g., the Human Metabolome Project). For unidentified metabolites, we used a range of techniques for identification of putative biomarkers found to be

significantly associated with CAC and/or IMT. These included tandem mass spectrometry (MS/MS) and simultaneous broad spectrum fragmentation (MSE) using UPLC-Quadrupole Time of Flight (QToF)-MS; hyphenated directly-coupled LC-cryo-NMR-MS was also employed as necessary, using solid phase extraction pre-concentration and fractionation, together with a suite of novel statistical spectroscopy methods. As necessary, other NMR approaches (2-D, ^{1}H-^{13}C spectra, cryoprobe-UPLC-NMR-MS, etc) were used for metabolite identification.

The data required extensive pre-processing in order to make them amenable to statistical analysis for biomarker discovery. Figure 7 below shows a typical UPLC-MS data generation workflow. SOPs for MS analysis, namely MS HILIC and MS LIPID profiling were made publicly available via the COMBI-BIO website.

Pre-processing of the LC-MS data. An LC-MS data set is three-dimensional with retention time, mass-to-charge ratio and intensity as the dimensions. First, the data set was reduced by the XCMS pre-processing software to two dimensions where the data table consisted of the samples for each row and the variables (retention time / mass-to-charge ratio pairs) in each column. This part included data trimming, peak detection, chromatogram deconvolution and retention time alignment.

The XCMS data table was trimmed to retain only the informative parts of the chromatogram by removal of features from chromatographic regions of column washing and equilibration. As the generated data table included many uninformative variables, the data were subject to several filtering steps. The first aimed at removing the variables unobserved in any of the quality control samples (iQC: internal quality control, eQC: external quality control, LTR: long term reference) assuming that 'real' peaks (i.e. non-artefacts) appeared in the quality control samples. The second was a dilution filter where a standard dilution series of the iQC samples was used to check the linearity of the feature intensity measurements. If a feature did not show a linear trend according to dilution, it was removed. In this step, robust linear regression was used to remove the influence of outlying samples on the dilution filter with the R^2 threshold chosen empirically.

In large-scale MS measurements, instability or drift in the response of the instrument may occur leading to systematic changes in the intensities of the features over the course of an analytical run. In order to correct these drifts, the change in intensity of each feature in consecutive quality control samples (iQCs in this case) were analysed. The feature lists of quality control samples were used to generate a locally weighted polynomial regression model for each feature between intensity and run order, which was used to correct the intensity of each feature (applied to each rack separately).

The above drift correction also effectively removed batch/rack differences. Thus, the data for each rack were merged and repeatability filtering was applied to the whole dataset. The coefficient of variation (CV) was calculated based on the internal quality control samples. The features that varied least within this set of injections were assumed to be the most reliable and reproducible measures. Therefore, features that had a CV of more than a threshold (50%) were removed.

The final pre-processing step was necessary because UPLC-MS derived features show heteroscedasticity as the standard deviation of the measured intensity increases with the increase in the mean intensity. In order to remove heteroscedasticity from the data, the natural logarithm of the intensity of each feature was calculated and used for further analysis.

The efficacy of pre-processing was monitored by Principal Components Analysis (PCA) as well as the CV distribution of all features in both biological and QC samples. Results were visualised by PCA

scores plots. The pre-processed MS data comprised different numbers of features for each cohort and each of the four assays, typically data set sizes being between 100MB to ca 1.5 GB.

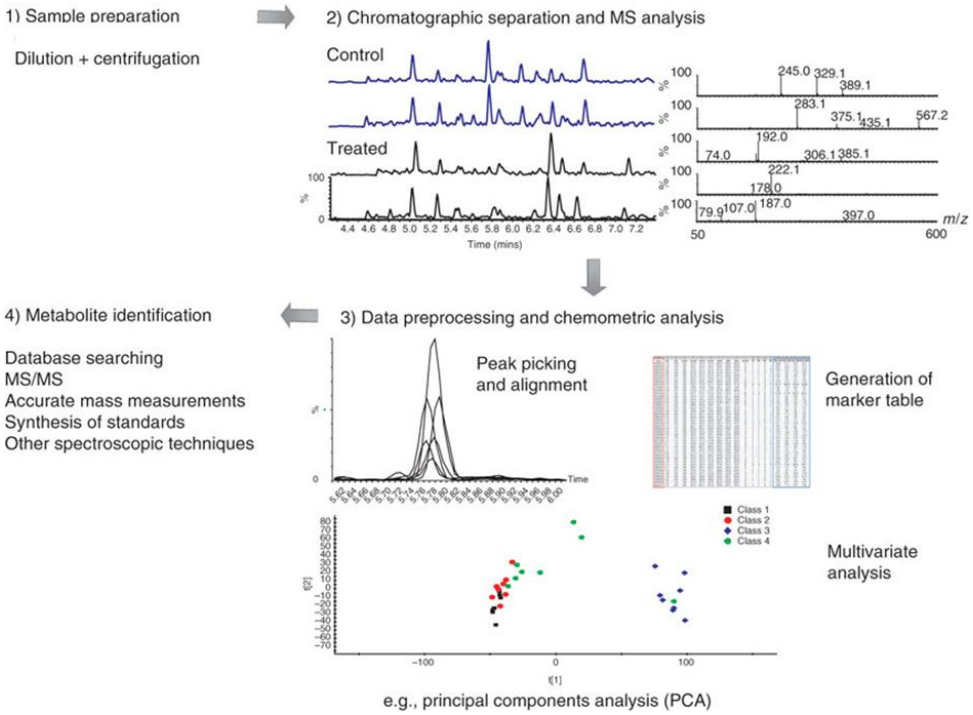

Figure 7 Schematic for Mass Spectrometry analysis of biological samples (adapted from Want E et al 2010)

Figure 8 shows a representative PCA score plot. iQCs were scattered among the biological samples before pre-processing whereas the iQCs are clustered together on the right hand side after all the pre-processing steps, demonstrating the effectiveness of the pre-processing.

To combine the LC-MS datasets from different cohorts, a peak matching algorithm was developed and applied to the three cohorts and seven datasets were generated: one consisting of features common to all three cohorts, three datasets consisting of features common to pairs of cohorts, and three datasets consisting of unmatched features.

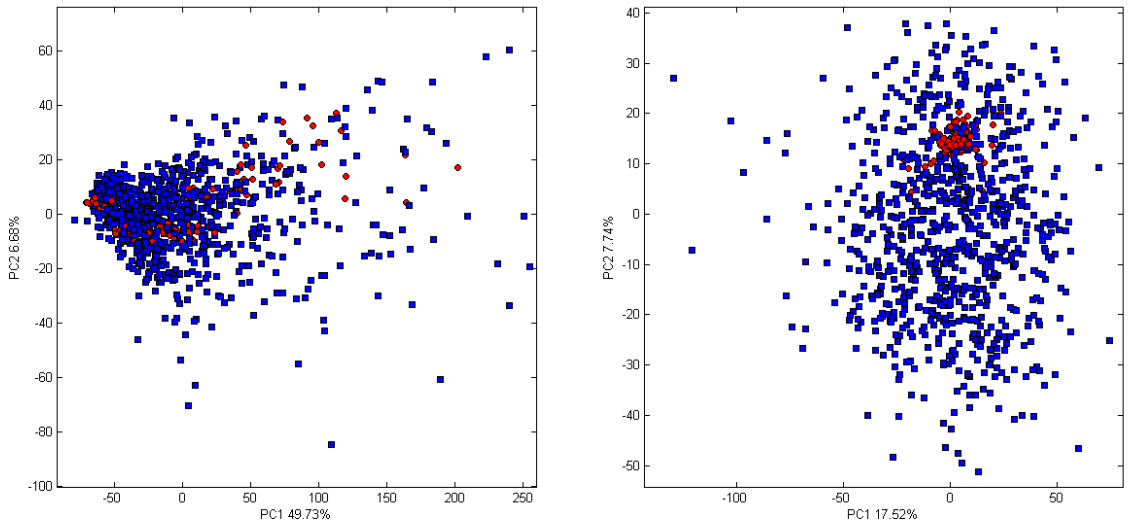

Figure 8 Comparison of PCA score plots of iQC (red circles) and biological (blue squares) samples before (left) and after (right) pre-processing workflow

Initial identification of discriminatory spectral features across groups (e.g., CAC absent vs score >300, upper and lower tertiles of IMT), was carried out using a variety of multivariate statistical approaches including orthogonal partial least squares (O-PLS) regression. Multivariate analysis of the MS data sets allowed identification of anomalous, outlying samples, evaluation of analytical reproducibility by ensuring close similarity of all internal QC samples, and identification of spectral regions linked to CAC and IMT.

Structural identification of unknown metabolites

NMR. Selected samples were analysed using a range of 2-D NMR experiments such as TOtal Correlation SpectroscopY (TOCSY) or Heteronuclear Single Quantum Coherence (HSQC) to aid molecular identification. Cross-correlation of spectral data allows further characterisation and improved molecular descriptors of metabolites identified as candidate biomarkers. We used approaches such as Statistical Total Correlation SpectroscopY (STOCSY) and STORM to provide information on molecular structure; such approaches can also give information on metabolic pathway connections (inter-molecular signal correlations). The spectral information was also compared with available literature and existing databases such as the Human Metabolome Data Base (HMDB): http://www.hmdb.ca/ and the Biological Magnetic Resonance Data Bank (BMRB): http://www.bmrb.wisc.edu. The NMR features were confirmed ultimately by purchase of authentic materials and when these were added to the samples and the spectra re-measured, exact superposition of all peaks was observed if a correct assignment had been made.

UPLC-MS. The structural assignment of metabolites remains a significant challenge. Features that ranked highly in the statistical models were entered into the biomarker ID workflow. For structural elucidation UPLC-MS^E and UPLC-MS/MS data were used. The MS^E method allows identification of the exact-mass precursor and fragment ion information while simultaneously obtaining accurate MS full scan profiles from every detectable component in the sample. MS^E data were collected on pooled samples at the end of the run for the phase 1 set of samples, whereas MS^E data acquisition was applied throughout the run of individual samples for phase 2.

According to published guidelines for metabolite annotation and identification, there are four levels of metabolite characterisation:

1. Unknown compounds—although unidentified or unclassified these metabolites can still be differentiated and quantified based upon spectral data.
2. Putatively characterized compound classes (e.g. based upon characteristic physicochemical properties of a chemical class of compounds, or by spectral similarity to known compounds of a chemical class).
3. Putatively annotated compounds (e.g. without chemical reference standards, based upon physicochemical properties and/or spectral similarity with public/commercial spectral libraries).
4. Identified compounds, requiring comparison of analytical data to a chemical reference standard.

Molecular formula and structural elucidation were therefore initially assisted by matching accurate m/z measurements to metabolites from online available databases such as the Metlin Metabolite Database (http://metlin.scripps.edu), the LIPID MAPS Lipidomics Gateway, (http://www.lipidmaps.org/) and the Human Metabolome Database HMDB (http://www.hmdb.ca).

These databases contain information regarding thousands of endogenous and drug metabolites, including MS spectra and in some cases MS/MS spectra. Where possible, comparison of MS/MS fragmentation patterns and chromatographic retention time between candidate biomarkers and reference chemicals is necessary to claim metabolite identification.

It was agreed by the whole consortium that structural identification of discriminatory features discovered in phase 1 will continue in parallel with the work carried out in phase 2. This refinement of the project protocol allowed work to proceed more effectively and did not impact adversely on the timescale for the phase 2 validation activities (see WP4 below).

WP4 – BIOMARKER VALIDATION

Phase 2 comprised analysis of 4,000 additional samples by NMR to provide a pooled NMR dataset of 8,000 samples, to maximise statistical power for discovery, and semi-quantitative MS analysis of the same 4,000 serum samples for validation of the results found in WP3.

Phase 2 NMR analysis. The SOP for phase 2 NMR was the same as used in phase 1 using 600 MHz ^{1}H NMR spectroscopy. However, the spectral calibration was performed on glucose at 5.23 ppm rather than TSP to improve the data alignment of phase1 and phase2 spectra. The two data sets (phase 1 and phase 2) were combined and aligned using novel data processing approaches.

NMR pre-processing. In phase 1, 4,000 spectra were acquired in each of two modes (NOESY and CPMG) across the three cohorts. Since there are often small shifts in peak position between samples, these spectra required alignment both within and between the cohorts. This was accomplished using a slightly modified version of conventional alignment procedures. We applied the same procedure to the full 8,000 phase 1+2 data set (Figure 9).

Table 7 shows the numbers of significant NMR features from the pooled analysis of 8,000 samples. Adjusting for Framingham Risk Score (FRS) variables (Model 2) dramatically reduced the number of associations, suggesting that most of the signals identified by Model 1 reflected metabolic changes associated with (and possibly on the causal pathway of) these factors. Sequential investigation of the FRS variables showed that the relative impact of each of these variables depended on both the type of NMR profiles and the clinical outcome (CAC or IMT). For instance, analyses of NOESY spectra against CAC-log2 showed that adjusting for diabetes and LDL-cholesterol yielded the largest drop in the number of associations.

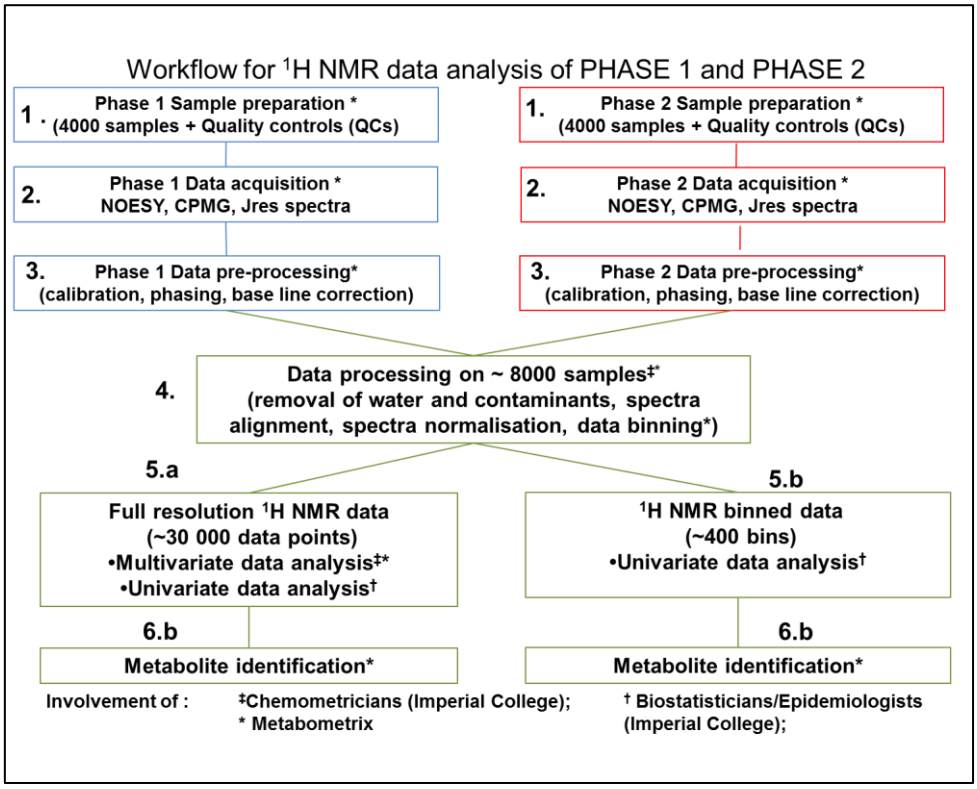

Figure 9 Workflow for NMR pre-processing of Phase 1 and 2 spectra.

		Log(CAC+1)		Log_{10}(IMT)	
		Model 1	Model 2	Model 1	Model 2
	N ppm	N significant spectral features	N significant spectral features	N significant spectral features	N significant spectral features
NOESY	30,590	11,217	546	11,717	240
CPMG	30,590	3,686	13	3,410	0

Table 7 Number of associations of NMR spectral features with log(CAC+1) and log_{10}(IMT) for Model 1 and Model 2, N=8,000. See Table 4 for definitions of Model 1 and Model 2.

Phase 2 MS analyses

To perform validation of biomarkers discovered in phase 1 by UPLC-MS profiling, a targeted validation approach to phase 2 analysis was initially planned utilising targeted chromatographic methods and multiple reaction monitoring (MRM)-guided detection using a tandem quadrupole mass spectrometer (TQ-MS). However, this approach was dependent on the successful naming and prioritization of molecular features of interest of biomolecules from phase 1. Because of the urgent need to commence phase 2 analysis as soon as possible after phase 1, it was decided that an enhanced profiling approach be applied as an alternative strategy. The revised strategy utilised MS^E detection as a means to capture information on structural specificity through correlation between simultaneously generated datasets with high and low molecular fragmentation – applied without sacrificing untargeted molecular coverage. In this manner, profiling was repeated, but with additional structural information approximating the approach through use of a TQ-MS with MRM,

allowing biomolecule characterization from phase 1 to coincide with data acquisition in phase 2. The Q-ToF MS systems used in phase 2 were of a newer generation than used in phase 1 and therefore provided additional sensitivity to the profiling assay.

New SOPs were developed for the phase 2 lipidomics and HILIC UPLC methods in both positive and negative mass spectrometry ionisation modes. For each of the 4 UPLC-MS assays, minor changes to the HILIC and lipid profiling chromatographic gradients were made in order to enhance profiling performance based on results observed in phase 1 analysis. These changes resulted in slightly longer methods which were determined to be appropriate for MS^E detection. Dynamic Range Extension was also applied to the lipid profiling assay to compensate for the increased sensitivity of the MS system and the wide dynamic range of observed lipid species. Whilst not providing a full unambiguous identification of metabolites, this approach gave a good degree of structural specificity. The resultant profiling of a total of 8,000 samples by UPLC-MS (i.e. over phase 1 and 2) represents one of the largest blood serum sample sets profiled by untargeted MS metabolic phenotyping to date.

MS pre-processing. At the start of the project, the LC-MS pre-processing protocol included peak detection, alignment, grouping, integration, normalisation and repeatability filtering. Owing to the very large, multi-cohort and multi-batch nature of the acquired data sets, a significantly augmented pipeline was developed. New additions included a linearity filter, variance stabilising transform and drift/batch correction, the latter being a critical component in such a large study. We also implemented a multi-cohort alignment/matching algorithm which was able to match signals across the separate cohorts, i.e. identify signals originating from the same metabolites in each cohort. We applied all these algorithms similarly to the phase 2 data.

MS data analysis. In the phase 2 analyses, the four UPLC-MS assays were run for approximately 4,000 samples. All data sets were pre-processed as above. From this a novel approach was devised to identify features in the MS data that correlated with CAC and IMT, either as raw data or as transformed data (i.e. log). A list of relevant features was derived for statistical models that were adjusted for age, gender, ethnicity and cohort (Model 1, see Table 4) and for Framingham criteria (Model 2). Feature lists were accumulated by considering all three cohorts together, the three combinations of two cohorts and the three cohorts separately. The highest priority was assigned to the feature list for all three cohorts together, for Model 2. These features were identified using a variety of MS-based techniques including use of the internal MS^E data, matching to standard databases and by addition of authentic standard materials.

We kept the analytical strategy consistent for both phase 1 and phase 2 as far as possible (Figure 10), though it was not possible to pool data from both phases and to match all the features across all cohorts. Metabolome-Wide Associations (MWAS) were performed using the same model parameterisation for each analytical approach (presented in columns in Figure 10) and using each sub-dataset (presented in rows). Using a response permutation procedure we calculated the per-test significance level (α') for the features common to the three cohorts and extrapolated for all other subclasses (Table 8). These estimates showed marked differences across assays and phases supporting the use of study-specific significance level estimates.

We conducted a series of MWAS on both phases. For candidate associations found in either phase (summarised in Table 9), we sought replication (using nominal 5% significance level) in the other phase.

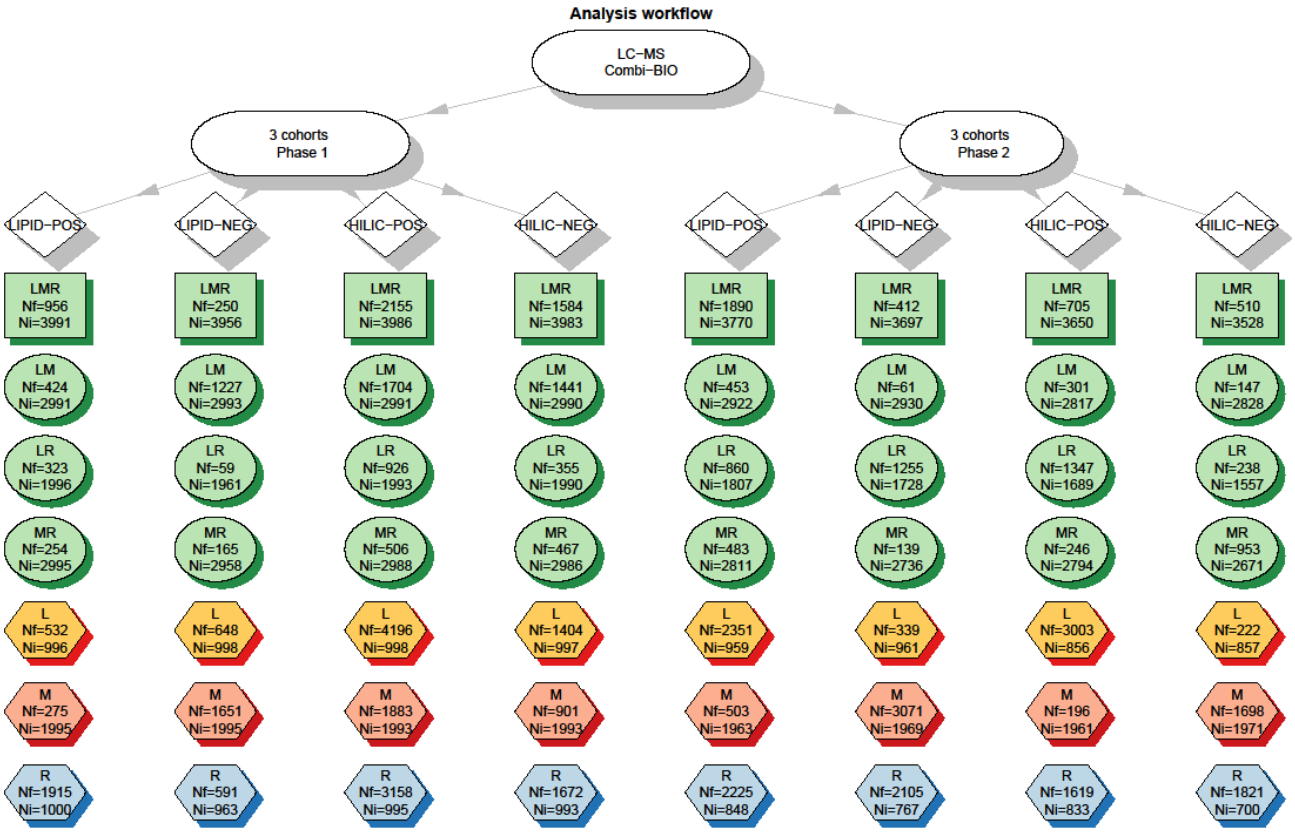

Figure 10 Schematic represent of the LC-MS analyses. For each assay, LIPID-POS, LIPID-NEG, HILIC-POS and HILIC-NEG, 7 datasets were generated after applying a between cohort matching procedure. LMR: features matched across the 3 cohorts (LOLIPOP, MESA & Rotterdam); LM: features matched across LOLIPOP and MESA; LR: features matched across LOLIPOP and ROTTERDAM; MR: features matched across MESA and Rotterdam. Ni: number of individuals; Nf: number of features

			CAC-log2		IMT-log	
	Assay	N features	α'	ENT	α'	ENT
Phase 1	Lipid +	956	1.07×10^{-04}	465	9.97×10^{-05}	501
	Lipid -	250	2.61×10^{-04}	192	2.38×10^{-04}	210
	Hilic +	2155	4.89×10^{-05}	1022	4.70×10^{-05}	1065
	Hilic -	1584	4.82×10^{-05}	1036	4.33×10^{-05}	1155
Phase 2	Lipid +	1890	5.10×10^{-05}	979	4.79×10^{-05}	1044
	Lipid -	412	1.61×10^{-04}	312	1.62×10^{-04}	310
	Hilic +	705	1.84×10^{-04}	272	1.74×10^{-04}	287
	Hilic -	510	2.01×10^{-04}	249	1.45×10^{-04}	344

Table 8 LC-MS Metabolome Wide Significance Level (MWSL) derived from a permutation-based approach for each outcome and type of data investigated. Only results from Model 2 for the LMR dataset are presented. ENT – Effective Number of Tests. See Table 4 for definition of Model 2.

			Log(CAC+1)		Log_{10}(IMT)	
	Assay	N features	N significant Model 1	N significant Model 2	N significant Model 1	N significant Model 2
Phase 1	Lipid +	956	2	0	22	1
	Lipid -	250	1	0	4	0
	Hilic +	2155	21	1	69	14
	Hilic -	1584	27	2	6	0
Phase 2	Lipid +	1890	74	0	11	2
	Lipid -	412	0	0	1	0
	Hilic +	705	5	0	0	0
	Hilic -	510	5	0	0	0

Table 9 Number of associations with CAC and IMT for Models 1 and 2. Only the features that were matched across all cohorts are presented. See Table 4 for definitions of Model 1 and Model 2.

Interaction by gender

We investigated the impact of gender on both the NMR and UPLC-MS significant features by adding an interaction term between gender and feature in Model 2. Few significant gender interaction terms were found. None were observed for CAC for either NMR ppm or the LC-MS peaks.

Enhancements to the programme

A number of enhancements to the project were agreed by the consortium to be carried out as an addition to the original project plan within the original budget. These related both to the substantive question under study by adding in a case-control component focussed on clinical cardiovascular events, and various quality control assessments to assist in understanding and interpreting the data.

1. Metabometrix investigated the spectroscopic profiles of approximately 300 participant serum samples from people who later went on to have a cardiovascular event. The data were embedded in the large data sets already acquired in phase 1 and phase 2.
2. Metabometrix undertook a series of quality control experiments to improve understanding of the data and analyses:
    a. examine a small sample set of serum samples (initially using NMR spectroscopy) that had been stored in Rotterdam at either -20°C or -80°C to investigate differences in biochemical composition which may be due to storage conditions;
    b. investigate any differences between serum and plasma samples based on 32 additional MESA samples;
    c. examine stability of metabolic profiles in stored samples over time.

1. <u>Inclusion of cardiovascular disease events</u>

Cardiovascular disease events were available in MESA and Rotterdam. Three cardiovascular disease outcomes were considered: myocardial infarction (MI), stroke and all CVD events grouping MI and stroke together. The number of cases and controls is reported in Table 10.

	MI		Stroke		All CVD	
	N cases	N controls	N cases	N controls	N cases	N controls
Rotterdam	114	1560	185	1574	267	1358
MESA	117	3827	143	3801	254	3690
Total	231	5387	328	5375	521	5048

Table 10 Numbers of incident cases and controls in Rotterdam and MESA cohorts in COMBI-BIO

For each metabolite associated significantly with CAC or IMT, we systematically screened for associations between the metabolic data and MI, stroke and all CVD events using a logistic regression model for Model 1 and Model 2 (see Table 4 for definitions).

2. Quality control studies

Sample stability according to temperature. The Rotterdam Study provided us with 30 samples from 15 individuals, where samples stored at -20°C and at -80°C were available. Both NMR and UPLC-MS were carried out for the serum samples from the same individuals stored at -20°C and -80°C. The study was done since the Rotterdam samples used in COMBI-BIO had been stored at -20°C while those for the other cohorts had been stored at -80°C (see Table 11), and it was important to know if storage temperature may have contributed to metabolic profiling differences seen in COMBI-BIO between Rotterdam and the other two cohorts.

Study	No. of samples	Metabolic Profiling Method					
		NMR (cpmg)	NMR (neosy)	LIPID UPLCMS (ESI+)	LIPID UPLCMS (ESI-)	Hilic UPLCMS (ESI+)	Hilic UPLCMS (ESI-)
ROTTERDAM Temperature	30	●	●	●	●		●
MESA Serum Vs Plasma	32	●	●	●	●	●	●
MESA Timecourse	21	●	●	●	●	●	●

Table 11 Sample assay details of QC enhancement projects in COMBI-BIO

Comparison of the metabolic profiles of serum and plasma. The MESA Study provided 32 samples to study the metabolic differences between blood serum, platelet-rich plasma, platelet-poor plasma and platelet-removed plasma from 8 individuals in each group. The study was carried out by both NMR and UPLC-MS (see Table 14). Preliminary findings from the NMR data show clustering by individual, whilst all multivariate models for MS data sets discriminated between serum and plasma. These biochemical differences were however much smaller than other sources of variation (e.g., by age, gender).

Stability over time of serum from the same individuals. The MESA Study provided us with 21 additional samples to conduct a study of the stability over time (sampling over three days: day 0, 3 and 6) of serum from the same 7 individuals, using both NMR and UPLC-MS (see Table 11). The study found that in general the metabolic profiles from individuals were stable at least over a period of six days studied. Any variation within an individual was much smaller than differences between individuals.

WP5 – CROSS-PLATFORM AND MULTI-OMICS ANALYSES

WP5 integrated metabolomic data from NMR and MS with the aim to make the most of the differential sensitivity/specificity of the two platforms. For computational reasons, in a first approach, data-driven methods were used, in order to explore the correlation structures between the set of MS-derived candidate biomarkers and the full NMR data. As detailed below, further functional investigations were carried out by integrating genome wide association study (GWAS) data with metabolomics data.

Cross-correlation of data sets based on NMR spectroscopy and mass spectrometry

Intra-platform correlations. Extensive use has been made of correlation of peak intensities found in the 8,000 NMR spectra using the STOCSY algorithm. This allowed the identification of different NMR peaks from the same metabolite and aided the metabolite identification. A similar but less formal approach was used for the MS data. In that case it was possible to identify different adducts of the metabolite that are formed in the mass spectrometer and thus again aid the identification of the molecular mass of the metabolite features, a crucial step in metabolite identification.

Cross-platform correlations. Given the huge data set sizes and the sensitivity of such analyses to small baseline and other artefacts, we cross-correlated the significant metabolic features found in the MS data with the full NMR data set. As a first priority, we explored the features identified for all three cohorts in the MS data for Model 2; these were cross-correlated with the full NMR data using Spearman's rank-order correlation (Table 12).

	Reference set			**Test set**		
	Log(CAC+1)					
	Ref.	N features	Phase	Test	N features	Phase
Intra NMR	NOESY	546	Pooled	CPMG	13	Pooled
NMR vs MS	NOESY	546	Pooled	HILIC $^{+}$	1	Phase 1
	NOESY	546	Pooled	HILIC^{-}	2	Phase 1
	CPMG	13	Pooled	HILIC $^{+}$	1	Phase 1
	CPMG	13	Pooled	HILIC^{-}	2	Phase 1
	Log_{10}(IMT)					
NMR vs MS	NOESY	546	Pooled	LIPID $^{+}$	1	Phase 1
	NOESY	546	Pooled	HILIC^{-}	14	Phase 1
	NOESY	546	Pooled	LIPID $^{+}$	2	Phase 2

Table 12 Combination of data used to estimate the correlation between significant metabolic features by NMR and MS.

Initial interpretation of the results from the correlation analysis was done by visual inspection of a heatmap; a matrix in which the correlation is indicated using colours. With very few exceptions, the significant NMR-NOESY features were negatively correlated with the significant HILIC^{-} features and positively correlated with the HILIC^{+} features.

Integration of metabolomic (NMR and MS) data with GWAS data

The list of NMR spectroscopic regions and the corresponding metabolite identities, together with the corresponding MS data and metabolite identities were distributed to the Helmholtz Center Munich for genomic-metabolomics cross-correlation. The metabolites obtained by NMR and having significant association with CAC and IMT were selected for GWAS in LOLIPOP and Rotterdam studies (MESA GWAS-metabolomic data are not yet available). This was done for the subsets of metabolites associated with CPMG-CAC (number of significantly associated metabolites N=13), NOESY-CAC (N=546) and NOESY-IMT (N=240). We then performed meta-analysis using inverse-variance weights for each subset; we found 66 significant SNP-metabolite associations for CPMG-CAC, 56716 for NOESY-CAC and 17 for NOESY-IMT, which were further classified into 29, 704 and 10 according to the metabolite class. From these results we used the metabolite-SNP association database created by KORA and TwinsUK (http://metabolomics.helmholtz-muenchen.de/gwas/) to infer metabolite names. This data resource contains the results of genome-wide association scans with high-throughput metabolic profiling comprising 7,824 adult individuals from 2 large European population studies [1]. Fifteen of the SNP-metabolite features were also found to have association with specific metabolites in the database.

We have shown previously that strong associations in GWAS with metabolic traits can point to interesting associations with clinical endpoints that otherwise would not be considered as relevant.
A large number of significant SNP-metabolite associations found in meta-analysis were not associated with specific metabolites in the metabolite-SNP association database of KORA. We regrouped these SNPs according to the genes found in the vicinity of the genomic position of the SNP and were able to find 215 genes associated with metabolites. These may give a hint as to the identity of the metabolite as well as the pathway involved.

References

1. Shin SY et al. *Nat Genet* 2014;46(6):543-50.

WP6 – PREDICTIVE COMBINATORIAL BIOMARKERS

WP6 brings together results from previous WPs on novel metabolic biomarkers/validation (WPs 3, 4) to develop novel risk prediction scores for subclinical atherosclerosis. First we obtained data on subclinical atherosclerosis (CAC and IMT), traditional and emerging cardiovascular risk factors, lifestyle and other potential explanatory variables from the participating cohorts. Specifically, the cohorts have data on smoking, blood pressure, lipid markers (total cholesterol, LDL cholesterol, HDL cholesterol) as well as a range of emerging biomarkers including apolipoprotein A and B, thrombotic markers such as d-dimer, fibrinogen, C-Reactive Protein (CRP) and other inflammatory markers (Table 3). Data on genetic variants from GWAS were also available in LOLIPOP and Rotterdam as described in WP5.

Data assembled and harmonized in WP2 were received from the LOLIPOP, Rotterdam and MESA studies. All data were anonymized and securely stored on a central server with daily back up. Table 12 summarises descriptive characteristics for the risk factors that were selected for inclusion in the prediction risk models, based on the main cardiovascular risk prediction scores such as the Framingham Risk Score and the Pooled Cohorts Equations, and in the main risk scores that have been developed for CAC prediction.

	LOLIPOP N=2,442	Rotterdam N=1,843	MESA N=4,050	ALL N=8,335
Age	54.78 (10)	70.82 (5.62)	62.85 (10.26)	62.25 (10.97)
Females (%)	836 (34.2%)	979 (53.12%)	2052 (50.66%)	3867 (46.39%)
Systolic blood pressure, mm Hg	131.50 (18.96)	143.23 (21.07)	126.99 (21.3)	131.91 (21.54)
Total cholesterol, mg/dL	210.73 (41.46)	225.06 (36.94)	194.5 (36.04)	206.02 (39.85)
HDL cholesterol, mg/dL	52.17 (13.32)	53.76 (15.01)	50.69 (14.75)	51.80 (14.45)
Diabetes (%)	295/2442 (12.08%)	262/1838 (14.25%)	577/4043 (14.27%)	1203/8323 (14.45%)
Current smoker	282/2442 (11.5%)	324/1836 (17.64%)	495/4034 (12.27%)	1101/8312 (13.24%)
CAC present	1276/2442 (52.2%)	1408/1843 (90.08%)	2105/4050 (51.97%)	5055/8335 (60.64%)
IMT, mm	0.67 (0.14)	1.07 (0.19)	0.75 (0.22)	0.80 (0.24)
Ethnicity	43.1% Caucasian 56.9% South Asian	100% Caucasian	38.66% Caucasian 13.3% Asian 24.81% African 23.2% Spanish	53.53% Caucasian 23.10% Asian 12.05% African

Table 13 Descriptive statistics (mean (sd) or percent) for traditional risk factors, CAC and IMT in each cohort

Predictive scores were generated for CAC and IMT in separate models. CAC was included as a dichotomous variable. Specifically we examined CAC>10 vs CAC ≤10 and CAC>0 vs CAC=0. IMT was used as a continuous variable since clinically valid thresholds are currently lacking. We examined the utility of emerging biomarkers such as CRP as additional risk predictors in the newly developed composite risk scores and evaluated whether the metabolomic features further added to the predictive ability of the models.

Discrimination and calibration were assessed for all models. Discrimination metrics included regression coefficients and confidence intervals in multivariate analyses, proportion of variance explained, improvement in the area under the curve (AUC) and comparison of risk distribution curves. Calibration was evaluated by estimation of the calibration slope and tested with the Hosmer-Lemeshow test and Harrell's E statistic. k-fold cross-validation including all data was performed to assess the generalizability of the findings in independent populations and if risk modelling produced unbiased estimates of effects.

For NMR we used the binned data, in order to reduce the number of variables included in the risk prediction models, based on the correlations between spectral variables. From each bin we picked the feature that explained the largest percentage of variance of the outcome, limiting the number of features included in the model to 37 for CAC and NOESY, 17 for IMT and NOESY, 3 for CAC and CPMG. This conservative approach prevented overfitting the model and minimised the false positive signals included in the risk prediction modelling.

For the evaluation of the LC-MS findings, the features that were significant under Model 2 in Phase 1 data and were matched across all cohorts were used for the development of a prediction risk score, based on data available in a sub-sample of 3,827 participants.

The prioritized metabolic features from the NMR and the LC-MS were then added to the model including the traditional risk factors described above. We used a backward stepwise selection procedure and calculated the AUC using the significant metabolic features and we also performed an additional analysis by forcing all the features into the model. We also assessed the models including the metabolic features derived both from NMR and LC-MS for the subsample where all data were available.

Cross-validation was done to account for over-fitting. The AUC calculated using a 10-fold approach yielded similar results indicating that our models provided robust estimates. Calibration was assessed by comparing the observed and predicted risk in different risk deciles; calibration was excellent for all models examined.

We also examined the utility of recently discovered genetic variants affecting coronary heart disease. One hundred and fifty three SNPs that were genome-wide or suggestive in association with coronary heart disease [1] were used to calculate a genetic risk score in the Rotterdam cohort and in participants of European descent in LOLIPOP.

References

1. The CARDIoGRAMplusC4D consortium *Nat Genet*; 2013;45:25-33

Potential impact (including the socio-economic impact and the wider societal implications) and the main dissemination activities and exploitation of results

Background

COMBI-BIO brings excellence in epidemiology, discovery of novel metabolic biomarkers and biomarker quantification, chemometrics, development and critical evaluation of risk scores, and cost effectiveness analyses. The COMBI-BIO consortium includes some of the largest and best characterised collections in the world of individuals with measurements of subclinical atherosclerosis (CAC and IMT), comprising the LOLIPOP Study (UK), Rotterdam Study (The Netherlands) and the Multi-Ethnic Study of Atherosclerosis (MESA) in the USA. The consortium thus brings together international leaders and resources across the multiple research disciplines necessary to make significant and rapid advance in this vital area of clinical research. This has allowed for the first time a consolidated effort to construct risk prediction scores for subclinical atherosclerosis based on novel metabolic markers using a "top down" systems biology approach. Development of a risk prediction score for sub-clinical atherosclerosis is a key first step in the early identification and management of susceptible individuals at high risk of future CHD events.

The scale and scope of the problem to be addressed required cooperation of various multi-disciplinary areas of expertise, knowledge and research capacities on a European/international scale. Moreover, carrying out the work at a European/international level added value in terms of reliability of data from large cohorts around the world at differing levels of risk of atherosclerosis, CHD and cardiovascular disease. Simply stated, the needed expertise and data are not available anywhere in the world in one place – thus cooperation was essential to make progress in this important area of clinical research. Bringing this consortium together at this time gave European researchers and SMEs a world-lead and competitive advantage, with potential benefits to both individual healthcare and society as a whole.

This collaborative project addressed one of today's major challenges in chronic disease prevention. The prognostic literature on predictive biomarkers even for CHD/cardiovascular disease is currently fragmented with many small studies examining one or two candidate biomarkers with poor standardisation and reproducibility between studies; the issue of subclinical atherosclerosis, allowing earlier prognostic information, interventions and treatment, had to our knowledge not been addressed before. Such fragmentation leaves little hope for making major advance in the area. By contrast, COMBI-BIO brought together data from leading, well characterized epidemiological cohorts in a needed large-scale study, which allowed the assessment of novel metabolic biomarkers using a "top down" systems biology, metabolic profiling approach based on both NMR and MS technologies. To our knowledge no other study has generated the breadth and depth of metabolic profiling on this scale. And no other study has brought together the epidemiological cohort resources with which to address the question of risk markers for sub-clinical atherosclerosis across a wide span of ages, different ethnic groups and countries covering individuals at widely differing risks of atherosclerotic disease.

One major aim of this FP7 programme was the reinforcement and development of European small and medium enterprises with the goal of strengthening the European scientific and economic outputs. An established university spin-out, the SME Metabometrix was core to this SME-led

COMBI-BIO project. In particular their expertise in the field of high-throughput platform-based metabolic phenotyping and computational medicine was key to the biomarker discovery, validation, computational and modelling aspects of the programme. Metabometrix, with its core expertise, is also well placed to capitalize on the biomarker and systems biology discoveries. The COMBI-BIO project is highly innovative, research-geared and research led. The SME Metabometrix took a leading role in the project, including leadership on 3 of the 6 WPs. Thus research and innovation were central to the SME's involvement in the project, strengthening its level of expertise and positioning the company to continue to be at the forefront of developments in this exciting and fast-moving arena.

Highly skilled personnel were employed by Metabometrix and the other European partners for delivery of the COMBI-BIO project. Personnel received specialist training and had the opportunity to contribute to a project at the edge of scientific knowledge, gaining further insights into their particular specialist areas and benefitting from international collaboration, and top-level international expertise, a combination highly valuable for their future careers. Thus the project contributed to strengthening R&D capacity in this highly specialized work, strengthening European capacity in the fields of computational biology, epidemiology, chemistry, bioinformatics, chemometrics and biostatistics, all requiring high computational ability, and all of which are skills-shortage areas in great demand both by industry and academia. These are precisely the skills bases that will deliver to society and European industry future knowledge and wealth generation.

Scientific and technological knowledge of the SME was reinforced by innovative solutions developed within the project in the area of systems biology for medical and clinical applications. Through the identification of novel metabolic and emerging biomarkers, it might be possible to target people for therapeutic and preventative interventions as part of a personalised and predictive medicine approach. Demonstrating the medical and clinical utility of systems biology approaches as well as the usefulness of their results for exploitation, Metabometrix is, as noted, currently exploring possibilities for patents arising from discoveries made within the COMBI-BIO project.

Socio-economic impact and societal implications

Introduction

Comprehensive multi–omics knowledge of large epidemiological cohorts, with careful systems biology functional characterisation of the identified biochemical variation, is a key product emerging from COMBI-BIO. The goal (dependent on current discussions regarding patentability) is to help improve the early detection of subclinical atherosclerosis, leading ultimately to earlier interventions, prevention and treatment, with impact on prognosis, treatment and clinical management. The new metabolic biomarkers that have been discovered in COMBI-BIO have relevance in terms of systems biology of atherosclerotic disease development. These novel findings can potentially lead to new mechanistic insights into atherosclerosis and hence the possibility of finding new targets for treatment early in the disease process, before clinical disease is manifest. The novel molecular signatures uncovered by COMBI-BIO may thus provide new mechanistically based information relevant to disease aetiology and prognosis.

The socio-economic impact of such developments in applying new omic technologies to a major clinical problem, such as atherosclerosis and coronary heart disease, is potentially great. Coronary heart disease and cardiovascular disease are the leading causes of death in Europe and throughout

the world, leading to a huge societal burden of premature morbidity and mortality, loss of economically active individuals (men and women) from the workforce, and increased health costs. The hope is that the application of high-throughput metabolic profiling, through innovations in metabolic phenotyping, accompanied by computational biology approaches to gain meaning and knowledge from the data, as has been developed and pioneered in COMBI-BIO, will lead to a new era of stratified and precision medicine. This in turn has prospects for huge benefit to individual patients and to society as a whole.

We have demonstrated in COMBI-BIO that such an approach can yield important dividends and discover completely new and un-thought of biochemical associations indicative of the disease process. In turn, these methodologies and findings will help equip the health sector R&D effort in Europe with the tools to deliver new solutions to tackle the heart disease epidemic, with potential to bring novel products and treatments to the market place.

This strategy will have a particular impact on those individuals for whom early diagnosis of subclinical atherosclerosis will lead to interventions to reduce risk of CHD/cardiovascular disease, which would not otherwise have been made. Such advances will provide a rational basis for the development of personalized and systems medicine approaches.

Impact on risk prediction

One of the most dissatisfying aspects in the care of patients with sub-clinical atherosclerosis is the often inability to predict individual disease risk. Previous attempts to develop genetic or biochemical biomarkers to establish individual disease risk profiles have yielded relatively few clinically useful tools. These efforts were hampered by small cohort sizes, imprecise disease definitions, the lack of technological prerequisites enabling the application of high-throughput screening paradigms, and inappropriate statistical approaches to biomarker detection. COMBI-BIO was set up to help overcome these challenges through an unprecedented European and international collaborative effort utilising large, well phenotyped patient cohorts, multi-molecular -omics profiling based on cutting-edge biotechnology platforms and integrative bioinformatics strategies.

Impact on patients

The impact on patients of the development and application of new -omics technologies for early disease screening and patient stratification is potentially profound. Identification of at risk individuals, and new potential disease targets, will guide the development of effective and safe therapies, to be started early in the disease process before major arterial, cardiac or neurological damage has occurred. This has the potential for a life-long benefit, allowing patients who might otherwise have suffered a major clinical CHD or cardiovascular disease event to live a normal life.

Impact on cardiovascular disease prevention

The new knowledge on metabolic correlates of subclinical atherosclerosis, has the potential to lead to new insights into the prevention and management of cardiovascular disease. In particular, the new knowledge generated, when combined with person- specific information on diet, lifestyles etc., has potential to help the clinician make more accurate prognosis and instigate intervention/treatment tailored to the needs of the patient much earlier in the disease process than hitherto. At the same time, this information will help to raise awareness of the disease, facilitate the communication of knowledge on risk estimates and may also help improve adherence

to lifestyle modifications and proposed therapeutic interventions in at risk populations. This individualised approach to tackling the CHD/cardiovascular disease epidemic is complementary to parallel public health strategies (what Rose termed “sick individuals and sick populations” [1]), and could result in significant health gains and reductions in health-care costs.

Impact beyond atherosclerosis

The development and application of high-throughput metabolic profiling and associated computational biology approaches, demonstrated in COMNBI-BIO in thousands of individuals, has great potential for the wider scientific approach to disease management and prevention. Although the project focussed on subclinical atherosclerosis, it provides a paradigm for discovery of mechanistic information that could lead to improvements in the management and treatment of other chronic diseases including cancer, and enhance scientific efforts toward personalised medicine. The evaluation of concordant biomarkers which are predictive for subclinical atherosclerosis will help throw light on disease mechanisms which can be further investigated through basic and population science by European groups. Finally, the project effort has contributed to building European leadership in the application of advanced ‘-omics’ to important clinical problems, and helped create a strong network of leading European research centres working on these cutting-edge problems.

Economic impact

The close collaboration of academic and SME partners in the consortium has provided an excellent example of public-private partnership necessary to maintain the European R&D effort at the forefront of knowledge and wealth generation. Such partnerships will be essential in the future to generate commercially exploitable diagnostic tools (such as assays based on metabolite profiles), proprietary technologies (e.g., assay kits, novel diagnosis algorithms, etc.), and ultimately new therapeutic agents (small-molecule drugs, biologics) based on the systems biology knowledge and understanding generated.

The market size for such diagnostic products is huge given the universal occurrence of atherosclerosis and the predominance of chronic diseases in an ageing society. Some of the innovative technological developments that have already been generated in COMBI-BIO may become the basis for the development of proprietary diagnostic tools for other related diseases and generic research technologies with a potentially wide range of applications. In this way, the economic impact of the exploitable research output will be amplified.

In summary, the commercial exploitation of technological advances achieved in the COMBI-BIO project has potential to contribute to the competitiveness of European biotechnology and pharmaceutical industries. At the same time ensuing advances from use of this technology may result in a more efficient utilisation of the health care resources available in the European community, with consequent improvements in health and welfare more generally.

Impact on data management & methodology

Due to the unique and complex nature of the dataset generated in COMBI-BIO, some methodological issues and problems were addressed. We developed an analytical pipeline which optimised statistical power and provided robust candidate biomarkers. Our approach relied on integrating data from different (and heterogeneous) populations (adjusting for study-specific characteristics) and assessing their validity by repeatedly splitting the study population for internal

validation to provide an estimate of the stability of our findings (i.e. their sensitivity to outlying observations). Our approach was compared to alternative strategies, such as the use of meta-analysis, and showed better statistical performance. In addition, we developed several novel visualisation tools to exploit the rich set of results arising from such a complex set of data.
The validity of our approach and visualisation tools as well as their portability was further explored by using them on an independent dataset (from INTERMAP study). Applicability to lower-resolution data was also explored using binned data from COMBI-BIO. Reassuringly, results showed marked consistency with those obtained from the un-binned data. In order to ensure a wide dissemination of our methodological developments publications describing our approach are currently in preparation.

Dissemination activities

Introduction

COMBI-BIO results have relevance that goes far beyond that of the partners directly involved in the project. COMBI-BIO is generating a number of reports and publications describing results generated and methods and rationale underlying these. These reports will represent end results of our project and will be widely disseminated through:

- COMBI-BIO website
- Publications
- Scientific conferences
- Other possible means, including personal influence, e.g., through physician societies, links to health departments, etc.

COMBI-BIO web site

The Project Management Team has created a dedicated COMBI-BIO web site. The web site includes a web page for the public where relevant deliverables for public use and consultation are posted. The site also has a section dedicated to researchers where research protocols and information have been posted for the scientific community. The web site also has an internal link where consortium participants are able to log in and exchange information and results between each other prior to the publication of results. The public relations departments of the partner institutions have been briefed to popularise and disseminate the COMBI-BIO project and the research results.

Publications in scientific journals

Many COMBI-BIO deliverables have resulted in internal reports and these are being adapted to form high impact peer-reviewed publications. Publications will be submitted by members of the consortium to peer-reviewed journals as these are read across Europe by all relevant target groups.

Presentations at European and other workshops/conferences

The project findings will be presented in meetings to scientific conferences of learned societies such as the European Society of Cardiology, the American Heart Association and the European Atherosclerosis Society. We are also planning to present our findings at local/national scientific conferences and regulatory forums.

Other means for creating awareness about improved prevention

To create awareness about improved/ new risk prediction models, COMBI-BIO will continue to give presentations to meetings of colleague scientists, hospitals, general practitioners, health

informatics managers and government agencies throughout Europe. Articles will be placed in relevant journals. Relevant material includes a recently generated video, and an article published about COMBI-BIO [2] together with an interview with Prof Paul Elliott about the project; these have been made available to European and American Cardiology Societies, as well as to press offices of our partner institutions. We are also aware of the opportunity for increasing awareness and disseminating standards by use of information technology, in particular the Internet. Members of our consortium are actively engaged in the dissemination of knowledge that informs development of national and international guidelines in cardiovascular health, and these channels will be adopted in parallel. All results achieved by COMBI-BIO will be made publicly available for the benefit of research and the public. In particular, metabonomic data will be deposited in appropriate public databases.

As a consequence of the above activities, the accumulated knowledge from COMBI-BIO is being disseminated throughout Europe and beyond. We are currently in the process of exploring patent options for the novel findings emerging from the project, which necessarily (and temporarily) has constrained further dissemination efforts while we seek to protect the IP arising from the study. Nonetheless, the inclusion of leading experts in the consortium from Europe and the U.S.A. with wide connections to policy makers and professional bodies (such as the European Society of Cardiology and the American Heart Association) is ensuring that the methodology and findings are being discussed at the highest levels. For example, we had the opportunity to present our methodologies to a closed meeting of the US National Heart, Lung, and Blood Institute on the application of metabolomics to prospective cohort data for the improvement of cardiovascular disease stratification and prediction. This was an excellent demonstration as to how a European initiative is leading the world in this cutting-edge area of research.

We have noted the need for special attention to be given to dissemination of COMBI-BIO methods and results to countries and centres not directly involved in this consortium including Eastern European countries and countries outside Europe.

Dissemination to commercial organisations

The SME working with the COMBI-BIO partners, Metabometrix Ltd., has access to a large audience of colleagues in both clinical medicine and public health, opinion formers, industry and patients. Therefore our collaboration provides a platform for a further uptake of research results that would have been difficult to achieve without the international collaborative approach adopted in COMBI-BIO. As a consequence, Metabometrix has mandated patent attorneys to evaluate their options of patenting scientific and technological outcomes of the project. Further dissemination into commercial activities, including into clinical practice, is dependent on the outcome of the patent discussions and the extent that the discovered biomarkers can be demonstrated to add clinical benefit and be commercialised.

Summary

The COMBI-BIO Coordinator has been responsible for ensuring that the activities of the COMBI-BIO consortium project and the role of the European commission, are publicised both through the routes described above, and also through press releases, and the sharing of appropriate information with stakeholder groups such as health authorities, governments, charities, medical societies and patient groups. Overall, dissemination of knowledge has targeted scientists from

academic institutions, private companies and other stakeholders, and in the clinical arena, both primary care and hospital practice.

Exploitation of project results and knowledge management

Introduction

COMBI-BIO is investigating the feasibility of securing patents on the methodologies and results achieved during the project. The project has also developed new IT tools to support the project research. The project will enhance the visibility and promotion of European Biotechnology and IT companies as qualified providers of outsourcing services in '-omics'; producing new knowledge that provides competitive advantages in the development of further research in '-omics' (specifically metabolomics).

An IP agreement including an exploitation plan is being developed by Metabometrix on behalf of the project partners. The underlying principles will be that partners will allow free academic access to existing IP and to IP generated during the project and that ownership of project technology will depend on whether it results from a sole or joint contribution. In the former case, the IP is the property and responsibility of the participant who will also notify the other participants of plans to protect or exploit IP generated from this project. In the latter case, the co-owners will jointly decide on a case-by-case basis how to protect their joint invention.

Metabometrix Ltd is involved in several national and international research activities (including with the pharmaceutical and diagnostics industries) and has previously filed both methodological and disease application patent applications that reflect its broad innovation activities. It has taken on the role of advising the project group of the feasibility of patent protection and has produced proposals as to how other intellectual property might be exploited.

According to the EC rules, valuable foreground should be protected. Protection is not mandatory, but any decision not to protect foreground will be made in consultation with the other project participants, which may wish to take ownership. If valuable foreground is left unprotected, it is understood that the European Commission may take ownership.

Various routes to commercial realisation of any such IP will be explored using the expertise of the interested project partners, especially the SME Metabometrix and Imperial College. A summary of the possible routes to commercialization, together with a suggested course of action has been circulated to the project partners. These routes include licensing out of new methodologies to scientific instrument manufacturers, licensing of new data handling ad processing algorithms for incorporation into chemometrics software by specialised suppliers of such products, the in-house development of new biomarker tests of coronary artery calcification (CAC) and intima-media thickness (IMT) as evidence of sub-clinical atherosclerosis, and the creation of alliances with industrial partners such as contract research organisations (CROs) who would offer such tests, nutrition companies promoting healthy food products, companies supplying diagnostics test and pharmaceutical companies, the most appropriate of which also have diagnostic kit products so as to provide theranostic capabilities. Any patents would need to be consistent with the policy of the funding bodies that support the individual cohort studies, cohort Steering Committee policies, as well as the informed consents that govern the commercial use of the data.

Since Metabometrix Ltd is not a software or diagnostics company, and has no product support staff or ability to offer on-going maintenance contracts or other after-sales support, the in-house development route is not preferred.

IP categories

The various types of IP generated in the COMBI-BIO project are summarised below, together with proposals for their development and exploitation.

Methodology: New methods and pipeline for preparing large sets of NMR spectra for statistical analysis would be suitable for licensing to an NMR instrument manufacturer. New methods and pipeline for preparing large sets of UPLC-MS spectra for analysis would also be suitable for licensing to a MS instrument manufacturer. New methods and pipeline for identification of significant regions of metabolic phenotyping data for biological endpoints across cohorts based on statistical approaches would be suitable for licensing to a chemometrics software manufacturer.

Databases: The main databases comprise NMR spectra (3 types) on 8,000 serum samples and UPLC-MS data (4 types) on 8,000 serum samples, and various licensing options are possible.

Biomarkers: The main IP is the panel of NMR- and MS-derived metabolic biomarkers associated with sub-clinical atherosclerosis based on CAC and IMT end-points. This could be patentable and advice has been sought from a specialist legal and patent company. It is possible that filings in the USA and European jurisdictions will ensue. After any filing occurs, it has been suggested that Metabometrix, on behalf of the consortium, begins negotiations with major diagnostics companies or CROs for licensing the IP. Identification of genetic linkage to significant metabolic biomarkers for sub-clinical atherosclerosis is novel and could produce new diagnostic methods or provide biochemical mechanistic information such as possible protein drug targets. Again, discussions are underway involving Metabometrix and their patent legal experts on feasibility and a plan of action.

Conclusion statement

A major assumption of the project was that through a dual-pronged approach involving both discovery and validation we would find novel metabolic biomarkers that are predictive of subclinical atherosclerosis. By including the validation step, as well as rigorous P-value thresholds for the discovery phase, we minimised the possibility of false positive associations and maximised the chances of a successful outcome of the project. By including both NMR and MS approaches, we mitigated against the possibility of the failure of one or other method/technology to deliver truly discriminatory and predictive biomarkers. Although the area of biomarker discovery has traditionally been challenging in the past, we have shown that our systematic approach, use of unique population collections with extensively curated epidemiological datasets, large sample size, state-of-the-art technology, and an international consortium involving leading research active SME and academic partners, has been able to deliver a substantial advance in metabolic phenotyping delivery and subsequent analysis over the previously fragmented efforts. In addition, our research has potential for new insights into disease pathways and mechanisms underlying atherosclerosis from consideration of the biomarkers associated with subclinical atherosclerosis and their metabolic connectivities.

References

1. Rose G. *Int. J. Epidemiol.,* 1985, **14**, 32-38; 2. COMBI-BIO. *International innovation* 2014; 162:62-64; 69

Address of the project public website, relevant contact details, project logo, video, list of all beneficiaries with the corresponding contact names

Project website address: http://www.combi-bio.eu

Beneficiary	Contact
Imperial College of Science, Technology and Medicine	Prof. Paul Elliott
Metabometrix Ltd.	Prof. John Lindon
University of Ioannina	Prof. John Ioannidis
Erasmus Universiteit Rotterdam	Prof. Albert Hofman
Helmholtz Zentrum, Munich	Dr. Christian Gieger
Northwestern University	Prof. Philip Greenland