Work Placement Project Portfolio

2010

Jason Wu
Preface

Introduction

This work place portfolio (WPP) was conducted between September 2009 and May 2010. The project was a prospective cohort analysis within the Cardiovascular Health Study (CHS). The CHS is an on-going study, which was established in order to examine risk factors associated with cardiovascular disease in community dwelling, older adults. The aim of our analysis was to investigate the association between several plasma phospholipids fatty acid biomarkers in the de novo lipogenesis pathway, with incidence of coronary heart disease.

Student’s role

This study involved collaborators both within my work place as well as with investigators in external institutions. As the primary investigator of the project, I was responsible for designing the project. To accomplish this, I carried out extensive review of the literature and obtained vital input from study collaborators. I submitted a detailed study protocol which was approved by the CHS steering committee. This allowed access to relevant data and I implemented all study analysis. Finally, I was the primary author for drafting a manuscript for this project which was submitted to a peer-reviewed journal. Through out this process I had regular meeting and sought advice from Dr Dariush Mozaffarian (primary supervisor at my work place) and also received valuable statistical support from Dr Fumiaki Imamura (co-investigator) and Mr Kevin McGeechan (WPP advisor).

Reflections on learning

The WPP gave me a first hand experience of being involved in a large prospective cohort study. I had the opportunity to extensively utilize statistical principles and methodologies throughout the WPP. While I believe the BCA course work was an excellent preparation for many of these challenges, the WPP experience enabled me to consolidate key concepts and develop many new skills. I gained an appreciation for the importance of careful preparation and planning required, prior to carrying out the statistical
analysis. The CHS has accumulated an immensely rich amount of data, and it was vital to figure out *a priori*, questions of central importance, without getting side tracked and dredge data.

I was already able to implement several of the key analytical approaches used in the study (linear regression, Cox regression), but developed new skills such as calculation and use of sampling weights. The Stata statistical package was used for all primary analysis, and I improved my programming ability (such as more efficient use of macros), by searching for Stata help files; and obtaining help from experienced users of this program. Finally, I was surprised that the time spent on computation was actually relatively minor, compared to designing and interpretation of results.

**Team work and communication skills**

Team work was instrumental in the completion of this WPP and I also had the opportunity to improve my communication skills through each stage of the WPP. I wrote an initial clear and concise project proposal to obtain approval to carry out the analyses, including a detailed section on statistical analysis methodology. My research group had weekly group meetings, where I regularly presented my latest results in tabular or graphical formats, with key statistical summaries. I discussed my interpretation of the findings with my colleagues, who provided constructive criticism and feedback. I believe these discussions enhanced overall understanding of epidemiological analyses and statistical knowledge of all those involved. Finally, my written communication skills are developed further by writing a manuscript for submission to a peer reviewed journal for the WPP.

**Ethical considerations**

The CHS involves four study centers and each center’s institutional review committee approved the study and all participants gave written informed consent. We gratefully acknowledge all volunteers’ participation in the CHS.
Project report

Project title

Biomarkers of endogenous fatty acids and risk of coronary heart disease: the Cardiovascular Health Study.

Location and dates

Harvard School of Public Health, Harvard University, USA.


Context

I am currently working as a post-doctoral fellow at the Harvard School of Public Health, under the supervision of Dr Dariush Mozaffarian. The research for our group focuses on the impact of nutrition and physical activity on cardiovascular diseases. One of the ongoing studies which we have access to, is the Cardiovascular Health Study. The CHS is a study based in the USA, which examines a range of anthropometric, lifestyle and biological risk factors for cardiovascular diseases in older adults. For the current analysis, we examined the association between plasma phospholipids fatty acids (as biomarkers for the fatty acids) in 1992 (baseline for the analysis), with incident coronary heart disease.

Student contribution

I was responsible for designing the study, carrying out all statistical analysis and writing the manuscript for publication. Throughout each stage of the project, I discussed my ideas and work progress with my co-investigators (principally Dr Mozaffarian and Dr Fumiaki Imamura) and WPP supervisor (Mr Kevin McGeechan).

Statistical issues

The primary analysis involved using Cox regression model. A range of modeling issues were considered such as covariate selection, dealing with missing covariate data, testing for linear effect of
primary exposures, testing the proportional hazards assumption and correction for regression
dilution bias. Linear regression was also used to investigate factors which could be associated with
phospholipids fatty acid biomarkers.

Acknowledgements

I would like to thank the CHS steering committee for allowing me access to the data. I am also
most grateful for all the guidance and feedback I received from Dr Mozaffarian, Dr Imamura, Mr
McGeechan and other study collaborators.

Student declaration

I declare this project is my own work, with direction and assistance provided by Dr Mozaffarian
and Mr McGeechan, and I have not previously submitted it for academic credit.

Jason Wu
Date 17.06.10

Supervisor declaration

I declare that Jason worked independently on this project. I have been very impressed by Jason’s
ability to apply the statistical methods that he learned in the Master of Biostatistics and his understanding
and application of new methods. Jason required minimal input from me on the choice of statistical
methods and the interpretation of results.

Kevin McGeechan
Date 17.06.10
Biomarkers of endogenous fatty acids and risk of coronary heart disease: the Cardiovascular Health Study

Introduction

De novo lipogenesis (DNL) is an endogenous pathway for lipid synthesis, whereby carbohydrates are converted to lipids (1). In humans, DNL occurs mainly in the liver, and there is a contribution from adipose tissue which is generally found to be low (2, 3). A normal function of DNL is the conversion of excess carbohydrate energy into fatty acids for storage as triglycerides (TG), given human’s relatively limited capacity for storing energy as glucose or glycogen (4). However, recent studies suggest that altered DNL may also affect risk of coronary heart disease (CHD). For example, hepatic DNL increases circulating fasting and postprandial TG levels (5, 6). Postprandial TG has been associated with other pro-atherogenic mediators (7), and is an independent risk factor for CHD and total mortality (8). Elevated hepatic DNL also appears to increase TG deposition in the liver (9), potentially contributing to development of non-alcoholic fatty liver disease (NAFLD) (10), which is associated with increased CHD risk (11, 12). Conversely, in animal models, whole body knockout or inhibition of stearoyl-CoA desaturase, a key enzyme in DNL, leads to increased aortic atherogenesis (13, 14). Thus, evidence for potential cardiovascular effects of DNL is limited and conflicting.

Another possible mechanism through which DNL may affect CHD risk is through generation of fatty acids. Fatty acids play crucial roles in the regulation of metabolic homeostasis and signaling pathways, and affect the development of CHD (15). The main fatty acids of DNL are myristic acid (14:0), palmitic acid (16:0); which can be processed by delta-9 desaturation and/or elongation to palmitoleic acid (16:1n-7), vaccenic acid (18:1n-7), stearic acid (18:0), and oleic acid (18:1n-9). Finally, oleic acid can be further beta-oxidized to 7-hexadecenoic acid (16:1 n-9) (Figure 1). As these fatty acids can be synthesized via DNL, they will be referred to as endogenous fatty acids in this paper, although they are also derived from dietary sources.
Circulating fatty acid levels are objective biomarkers to assess relationship of individual fatty acids with disease (16). Circulating levels of endogenous fatty acids are directly influenced by de novo synthesis. Low-fat, high-carbohydrate diets increased plasma phospholipids levels of 14:0, 16:0, 16:1 n-7, 16:1 n-9, 18:1 n-7, and 18:1 n-9 (17, 18), consistent with the endogenous synthesis of these fatty acids, and the stimulatory effect of low-fat, high-carbohydrate diets on hepatic DNL both in isocaloric and excess calorie states (19). Alcohol is another dietary factor which stimulates hepatic DNL (20), and increased alcohol intake has also been associated with higher cholesterol ester and phospholipids levels of 14:0, 16:0, 16:1 n-7 and 18:1 n-9 (21).

Thus, DNL influence levels of endogenous fatty acids, which may affect CHD risk via their effects on cellular function and metabolism. Only a limited number of studies have investigated the relationships of endogenous fatty acid biomarkers with CHD risk (22-24). These studies mostly assessed major endogenous fatty acids such as 16:0 and 16:1 n-7, and found null or small positive associations with total CHD risk. The role of relatively minor endogenous fatty acids such as 16:1 n-9 and 18:1 n-7 was investigated in a recent retrospective case-control study (25). Red blood cell membrane levels of 16:0 and 16:1 n-9, were positively associated with risk of sudden cardiac arrest (SCA), independent of both traditional and established fatty acid risk factors (25). None of the other endogenous fatty acids were independently associated with SCA. These studies suggest both major and minor endogenous fatty acids could affect CHD risk, which needs to be additionally investigated in relation to specific CHD outcomes. Previous studies also focused on middle aged adults, whereas relationships between these fatty acids with CHD in other age groups have not been studied. To address these important issues, we investigated the relationship between endogenous fatty acids, using plasma phospholipids biomarkers, with incidence of CHD in the Cardiovascular Health Study (CHS), a population based cohort of older adults. Furthermore, we investigated the association between potential inducers of DNL and these biomarkers.
Subjects and Methods

Study Design and Participants

CHS is an NHLBI funded prospective cohort study designed to investigate risk factors for cardiovascular disease in community-based older adults (26, 27). Participants were randomly selected from Medicare eligibility lists from 4 US communities: Sacramento County, CA; Washington County, MD; Forsyth County, NC; and Allegheny County, PA. 5201 men and women were initially recruited in 1989-1990, with an additional 687 black participants recruited in 1992-1993. Individuals were eligible if they were \( \geq 65 \) years old, not institutionalized, expected to remain in current community for >3 years and not under active hospice or cancer treatment. Each center’s institutional review committee approved the study and all participants gave written informed consent. Plasma samples were obtained and stored in 1992 to 1993. For the current study, we measured plasma phospholipids fatty acids using stored plasma samples in 3419 participants. This included 2906 new measurements, and 513 measurements from a previous case-control study of myocardial infarction (MI) nested in the CHS cohort (28). Analyses accounted for this sampling within the cohort using inverse-probability-of-sampling weights (see appendix D). We excluded 673 participants with prevalent CHD (history of MI, angina, or coronary re-vascularization at baseline), leaving 2746 participants for the current analysis.

Blood Sample Collection and Fatty Acid Analysis

Individual plasma phospholipids fatty acids were measured as percent of total fatty acids by the Fred Hutchinson Biomarkers Laboratory (25). Blood was sampled after a 12-hour fast and stored at \(-70^\circ C\) before shipping on dry ice for long term storage at \(-80^\circ C\). Prior studies confirmed that under these conditions phospholipids fatty acids are stable and do not degrade due to lipolysis or oxidation after 10 years of storage (28, 29). Total lipids were extracted from plasma according to Folch (30), and phospholipids separated from neutral lipid using one dimensional thin layer chromatography. We followed Lepage’s method to prepare transesterified fatty acid methyl esters (FAMEs) (31), which were
analyzed using gas chromatography (Agilent 5890 Gas Chromatograph FID detector; Supelco fused silica capillary column SP-2560 (100m x 0.25mm, 0.2μm, from Sigma Aldrich); initial 160°C for 16min, ramp 3°C / min to 240°C, hold for 15min). Identification, precision, and accuracy were evaluated using model mixtures of known FAMEs and an established in-house control pool, with identification confirmed by GC-MS at the USDA lipid laboratory in Peoria, IL. Pooled quality control samples were run concurrently with study samples to confirm low batch-to-batch variation, with CVs between 3% (16:0) to 23% (16:1 n-9) for endogenous fatty acids.

Among the fatty acids known to increase following a low-fat, high-carbohydrate diet (14:0, 16:0, 16:1 n-7, 16:1 n-9, 18:1 n-7, and 18:1 n-9)(17), 18:1 n-9 is influenced by dietary vegetable oil intake (32), and 14:0 is also a biomarker of dietary dairy intake (33). Thus, phospholipids levels of 16:0, 16:1 n-7, 16:1 n-9 and 18:1 n-7 may more strongly reflect endogenous synthesis.. As the primary focus for this study was fatty acids which may most strongly reflect endogenous synthesis, we therefore a priori chose primary exposures to be phospholipids levels of 16:0, 16:1 n-7, 16:1 n-9 and 18:1 n-7.

Assessment of coronary heart disease

The CHS protocol for ascertainment of incident CHD events has been described (34). Our primary outcome was total incident CHD, including fatal or non-fatal MI and CHD death. We also separately evaluated incident nonfatal MI, incident fatal CHD and SCA. All CHD events were adjudicated by a centralized Morbidity and Mortality Committee. MI was ascertained on the basis of cardiac enzymes, chest pain, and serial electrocardiogram changes. CHD deaths were defined as fatal MI or suspected fatal CHD events where the participant also had chest pain within 72 hours of death or had history of chronic CHD. SCA was defined as sudden pulseless condition from a cardiac origin in a previously stable individual occurring out of the hospital or in the emergency room, based on review of fatal CHD cases by a cardiologist (29). These cases could not have a life threatening, non-cardiac comorbidity or be under hospice or nursing home care. A second cardiologist conducted a blind review of
a random sample of 70 potential SCD cases and found 88% inter-reviewer agreement and a \( \kappa \) value of 0.74.

Assessment of Other Covariates

All covariates were assessed in 1992 (time of blood collection) except for usual dietary habits (see below). Participants completed standardized questionnaires on health status, medical history, cardiovascular risk factors, and physical activity and alcohol intake using validated questionnaires (26). They also underwent a clinical examination and laboratory evaluation. Laboratory assays and quality control have been reported (35). Usual dietary intake was assessed using semi-quantitative food frequency questionnaires (FFQs) in 1989-90 (36) as well as in 1996 (37). Detailed description of the reproducibility and validity of these FFQs have been published elsewhere (38, 39). As fatty acid measurements used blood samples from 1992, we averaged responses on the two FFQs to minimize misclassification due to errors in measurement and changes in lifestyle. In the African American sub-cohort, the only data available were from the second FFQ.

Statistical Analysis

All exposure, covariate and outcome data were obtained from the CHS central data base. Prior to carrying out the regression analysis I generated several new variables. These included: sampling weights; average dietary macronutrient intake (based on two FFQ); imputation of missing covariate data; generation of standardized fatty acid and categorizing participants according to their fatty acid levels and generated total follow up time for each participant. See Appendix A for detailed Stata codes and explanations.

For analyses of risk, phospholipids fatty acids biomarkers were evaluated as indicator variables (quartile categories) and continuously (per inter-decile range, comparing the median of the first and fifth quintiles, i.e. the 10th to 90th percentile). The inter-decile range is calculated by calculating the inter-decile range (90th percentile subtract 10th percentile value for each exposure variable), then dividing each
subject’s value of the exposure variable by this range. The calculated variable is then used in
the Cox regression models. We calculated this to make the analysis comparable to previous work done in
the group. The unadjusted inter-relationship between the different endogenous fatty acids were evaluated
using Spearman correlations. The relationship of each endogenous fatty acid with incident CHD was
evaluated using multivariable-adjusted Cox proportional hazards (See Appendix B for Stata codes used
for Cox regression modeling and Appendix D for checking proportional hazards assumption). Previous
studies in this cohort (Dariush Mozaffarian, manuscript in preparation) found risks associated with fatty
acid biomarkers are most stable over the first 10 years of follow up, possibly due to mis-classification of
exposure beyond 10 years. Therefore, time at risk was until first diagnosis of CHD, or censored at non-
CHD death or follow up limited to 10 years.

To minimize potential confounding, covariates were selected based on biologic interest, being
well-established risk factors for CHD risk in older adults, or associations with exposures and outcomes in
the current cohort. Based on these considerations and the goal of parsimony in covariate selection, 3 final
multivariate models were fitted: (1) adjusted for cardiovascular (age, gender, race, education, smoking
status, diabetes, hypertension, prevalent stroke or transient ischemic attack) and lifestyle risk factors
(body mass index, leisure time physical activity, alcohol use, total fat intake and total caloric intake); (2)
further adjusted for long chain n-3 fatty acids (eicosapentaenoic acid, EPA; and docosahexaenoic acid,
DHA) and trans fatty acids (total Trans 18:1 and total Trans 18:2); and (3) further adjusted for potential
confounders or mediators (systolic blood pressure, fasting HDL-C, LDL-C, triglycerides, C-reactive
protein and fibrinogen). Potential effect modifications were investigated for age and gender by assessing
the significance of multiplicative interaction terms using Wald tests. Missing continuous covariates were
imputed (all ≤6.6% missing) by single imputation using baseline data, including age, gender, race,
smoking status, alcohol use, education, physical activity, body mass index, prevalent diabetes and stroke
as predictors (see appendix D). All categorical covariates had <0.1% missing and results were similar
excluding participants with missing values. Regression dilution bias were adjusted according to published
methods (40, 41), using repeat fatty acids measurements (3 years apart) available among a subset of subjects (n = 75), (see Appendix D). The Spearman correlations for repeated fatty acid measurements were: 14:0 (r = 0.49); 16:0 (r = 0.73); 16:1 n-7 (r = 0.72); 16:1 n-9 (0.39); 18:0 (r = 0.67); 18:1 n-7 (0.59); 18:1 n-9 (0.59).

Associations between factors which may stimulate DNL, chosen a priori based on previous studies (19) and including carbohydrate intake, alcohol intake, positive energy balance (using BMI as surrogate measure), and fasting insulin, and endogenous fatty acids levels were evaluated using linear regression with each biomarker as the dependent outcome. Carbohydrate and protein intake were assessed in these linear models using multivariate nutrient density method, representing substitution for an equal energy from total fat (42), see Appendix D. See Appendix C for Stata codes used for linear regression modeling. All p-values are two-tailed (α=0.05). Analyses were performed using Stata 10.1 (Stata Corp, College Station, Tex).
Results

At baseline, average participant age was 73 years (range, 65-97 years); 39% were male and 90% were white. Cardiovascular and other lifestyle risk factors are shown in Table 1. Participants had diverse socioeconomic backgrounds, were on average overweight, and generally obtained a higher proportion of daily energy from carbohydrate (mean ± SD, 54 ± 7 %) than fat (mean ± SD, 31 ± 5 %).

Levels of endogenous fatty acids represented between 0.09 to 25.4% of total fatty acids (Table 2). Highest levels were seen for 16:0 (mean ± SD, 25.4 ± 1.6 %) and lowest for 16:1 n-9 (mean ± SD, 0.09 ± 0.03 %) Most of the fatty acids were positively inter-correlated (r between 0.23 to 0.60). The main exception was 18:0, which was inversely correlated with all other fatty acids (r between -0.16 to -0.44). The strongest correlation was observed between 14:0 and 16:1 n-7 (r = 0.60); 16:0 and 16:1 n-7 (r=0.57) and 16:1 n-7 and 18:1 n-9 (r = 0.56). Summary statistics of the each fatty acids are given in Table 2 and the histograms attached after the Table.

During 10 years (21852 person-years) of follow-up, 485 incident cases of CHD occurred (22.2 per 1000 person-years). This included 289 non-fatal MI and 255 fatal CHD cases. Among fatal CHD cases, 76 adjudicated SCD events were identified. In unadjusted and multivariate adjusted Cox regression analyses, the only primary fatty acid of interest significantly associated with total CHD was plasma phospholipids 16:1 n-9 (Table 3). However, for 16:0 and 161: n-9, analyses of Schoenfeld’s residuals demonstrated that the assumption of proportional hazards was not met (P<0.001). Based on visual inspection of Kaplan-Meier survival curves, the risk associated with the fatty acids was stable over the first 5 years of follow up, with subsequent crossing over of the survival curves (Online Supplemental Figure 1). Therefore, separate analysis from baseline to 5 years, and 5 to 10 years was carried out. In both of these periods the proportional hazards assumption held for 16:0 (P=0.1 and 0.74 for testing the null hypothesis of proportional hazards) and 16:1 n-9 (P=0.63 and 0.14 for testing the null hypothesis of proportional hazards). 16:0 remain unassociated with CHD risk in both follow up periods (not shown). On
the other hand, 16:1 n-9 was associated with substantially increased risk in the first 5 years of follow up (Table 4). The hazard ratio in increasing quartiles of 16:1 n-9 were 1.0 (reference), 0.52 (95% CI 0.16-1.72), 2.12 (95% CI 0.69-6.54), 9.59 (95% CI 3.63-25.3). 16:1 n-9 was associated with increased risk of CHD death (HR for inter-decile difference, 2.26; 95% CI, 1.51-3.37), non-fatal MI (HR for inter-decile difference, 2.59; 95% CI, 2.05-3.28) and SCA (HR for inter-decile difference, 2.0; 95% CI, 1.20-3.33). In contrast to the first 5 years, 16:1 n-9 was associated with significantly reduced risk of total CHD in the latter 5 years of follow up. The hazard ratio in increasing quartiles of 16:1 n-9 were 1.0 (reference), 0.33 (95% CI 0.12-0.91), 0.21 (95% CI 0.07-0.62), 0.27 (95% CI 0.08-0.89). The interaction between baseline level of 16:1 n-9 and follow up time was significant ($P$ for interaction = 0.05). In the latter 5 years, higher 16:1 n-9 was associated with reduced risk of non-fatal MI (HR for inter-decile range, 0.09; 95% CI, 0.02-0.38) but not CHD death (HR for inter-decile range, 0.64; 95% CI, 0.19-2.21). The association between 16:1 n-9 and sudden cardiac death was precluded from analyses in this period due to the very small number of sudden cardiac death cases (n=27). The association with total CHD risk of 16:1 n-9 was not modified by gender or age in both follow up periods ($P$ value for interaction all >0.05).

In the first 5 years of follow up, further adjustment for additional plasma phospholipids fatty acids (EPA, DHA, total Trans 18:1 and total Trans 18:2) did not attenuate the association between 16:1 n-9 with total CHD (HR, 2.35; 95% CI, 1.92-2.88), CHD death (HR, 2.06; 95% CI, 1.50-2.83), and non-fatal MI (HR, 2.39; 95% CI, 1.86-3.06). Cross sectional analyses of association between phospholipids 16:1 n-9 with potential confounders at baseline (data not shown), found it was positively associated with plasma triglycerides, and inversely associated with LDL cholesterol and inflammatory markers (CRP and fibrinogen). Further adjustment for potential mediators (systolic blood pressure, fasting HDL-C, LDL-C, triglycerides, CRP and fibrinogen) did not alter observed associations for total CHD (HR, 2.44; 95% CI, 1.99-3.0), CHD death (HR, 2.18; 95% CI, 1.61-2.97), and non-fatal MI (HR, 2.48; 95% CI, 1.93-3.18). These adjustments also did not alter the observed associations in the latter 5 years of follow up (not shown).
In exploratory analyses investigating the other endogenous fatty acids, 14:0 was also associated with increased risk of CHD in the first 5 years of follow up (multivariate adjusted HR for inter-decile range = 4.20, 95% CI, 2.26-7.81). Inclusion of 16:1 n-9 and 14:0 simultaneously in the multivariate model significantly attenuated the association for 14:0 (HR for inter-decile range = 1.53, 95% CI, 0.71-3.28) but not for 16:1 n-9 (HR for inter-decile range = 2.21, 95% CI, 1.76-2.77).

_A priori_ selected potential stimulators of the DNL pathway were associated with plasma phospholipids endogenous fatty acids (Table 5). Consumption of carbohydrate or protein in place of fat and also alcohol use were each major independent predictors of these fatty acids. The associations were linear (P for trend, all<0.01). In contrast, BMI and plasma insulin were not strong predictors.
Discussions

In this large prospective cohort study among older US adults, higher plasma phospholipids levels of 16:1 n-9 were associated with risk of incident CHD. This increased risk was confined to the first 5 years of follow up and consistent for total CHD, fatal CHD, and non-fatal MI. In contrast, after the first 5 years of follow up, 16:1 n-9 was inversely associated with CHD risk. Other endogenous fatty acids were not independently associated with CHD risk.

The observed association with SCA is consistent with a prior case-control study (25). However, the mechanism by which 16: n-9 might increase risk of CHD is unknown. While adjustment for lipid and inflammatory risk factors did not attenuate the association between 16:1 n-9 and CHD, there are other potential confounders we could not examine in the current study. For example, elevated DNL has been associated with CHD risk factors such as increased post-prandial TG concentrations, and TG deposition in the liver (5, 9). As DNL could also partly contribute to plasma 16:1 n-9 level, future studies need to investigate the association between 16:1 n-9 and these metabolic risk factors. In addition, as fatty acids could direct modulate heart function such as myocyte metabolism and ion channel activity (15), the effects of 16:1 n-9 on these physiological parameters warrant further investigation.

The observed inverse association between phospholipids 16:1 n-9 with CHD risk between 5 to 10 years of follow up needs to be interpreted with caution. The relatively low correlation observed for repeated measurements of 16:1 n-9, suggest it has high biological variability. Using participant’s baseline measurements will therefore result in increasing mis-classification of 16:1 n-9 level over time. However, the mis-classification would be expected to be random with respect to CHD outcome and therefore attenuate the results towards the null over time, which would not explain the inverse association with CHD risk observed in the latter half of follow up. We can not exclude the possibility that this was due to chance, or real long term biological effects of 16:1 n-9. Serial measurements of 16:1 n-9 levels in later
time points are needed to ascertain possible changes in 16:1 n-9, and the validity of the observed inverse association with lower CHD risk.

We did not find any significant independent association between the other primary fatty acid exposures of interest and CHD risk. 18:1 n-7 is a relatively under-studied fatty acid. To the best of our knowledge, this is the first prospective study to assess its association with total incident CHD. In our prior case-control study, it was modestly associated with SCA risk, but this was attenuated after adjustment for 16:1 n-9 (25). Overall, these findings do not support a significant relationship between 18:1 n-7 and the risk of CHD in older adults. There has been growing interest in the role of 16:1 n-7 in cardio-metabolic diseases. The enzyme which catalyzes the in vivo synthesis of 16:1 n-7, stearoyl-CoA desaturase, has been found to affect susceptibility to obesity and diabetes in animal studies (43), but its exact role in CHD remains uncertain (13, 14). Our study is consistent with prior investigations using plasma 16:1 n-7 biomarker which have generally found no association with CHD risk (22, 23). Future studies need to investigate if 16:1 n-7 could be associated with other metabolic disturbances such as insulin resistance and diabetes. Finally, prior epidemiological evidence investigating the association between 16:0 and CHD is inconsistent (22-24). In the Atherosclerosis Risk in Communities Study, plasma phospholipids and cholesterol ester 16:0 in men and women was not associated with incident CHD (23). On the other hand, a small case-control study and a population based cohort in Sweden found that, elevated 16:0 in men was associated with CHD risk (22, 24). Our study differed in several ways from these studies; including population characteristics (older age and inclusion of both genders in our study), fraction of plasma lipid used to analyze biomarkers (phospholipids versus cholesterol esters) and study design (including more detailed adjustment for covariates in our analyses). Additional studies are required to elucidate the potential relationship between 16:0 and CHD risk.

Phospholipids concentrations of endogenous fatty acids reflect interaction between dietary intake, endogenous synthesis and metabolism. The relative quantitative contribution of these different processes to the levels of these fatty acids is not well established. In the current study, the positive inter-correlation
between the endogenous fatty acids is consistent with a potential common source or determinants. 16:0, 16:1 n-7, 16:1 n-9 and 18:1 n-7 were all positively associated with replacement of fat with carbohydrate or protein intake. The association was linear across the range of fat intake (13-50% of total energy) in this cohort of older adults. These results are consistent with the known stimulatory effect of increased carbohydrate intake on DNL (1). The potential contribution of de novo synthesis towards endogenous fatty acid levels is also strengthened by the association with alcohol intake, shown in clinical studies to induce hepatic DNL (19). While humans are thought to have a low capacity for DNL (19), these results support the importance of DNL in at least partly determining plasma endogenous fatty acid levels. This has public health relevance as the prevalent dietary guidelines have resulted in recent shift in overall energy intake from fat to carbohydrate (44). Furthermore, the increased carbohydrate intake often occurs in the form of added sugars, which may have a more pronounced inductive effect on DNL than starch (1). Further studies are required to ascertain dietary and physiological factors which drive DNL and how these relate to biomarkers of endogenous fatty acids. Additional studies are also needed to investigate the time course of such effects, and particularly effects of changes in fatty acid levels over time on the risk of CHD.

The current study has several strengths. The cohort design reduces selection bias, and exclusion of persons with known CHD at baseline reduces potential bias from reverse causation (presence of disease influence fatty acid biomarkers). Adjustment was made for major lifestyle, dietary and phospholipids fatty acids (n-3 and Trans –fatty acids) risk factors associated with CHD, which reduces the influence of confounding. Low loss to follow-up, comprehensive review and central adjudication of outcomes decrease the potential for missed or misclassified outcomes. The population based recruitment enhances generalizability.

Several limitations are also noteworthy. In observational studies, residual confounding as a result of unmeasured or imprecisely measured confounders can not be ruled out. Single baseline measurement of endogenous fatty acids will misclassify long term exposure to these fatty acids, especially those with
high biological variability such as 16:1 n-9. We found an unanticipated change in relationship of 16:1 n-9 with CHD over time, that could either relate to such misclassification or to chance. On the other hand, 16:1 n-9 was also significantly associated with increased CHD risk in the main analysis.

In conclusion, higher phospholipids 16:1 n-9 was associated with an increased short term risk of CHD among older men and women, whereas other endogenous fatty acids were not independently associated with CHD risk. Further work is needed to confirm these findings and identify dietary sources and metabolic factors which raise the level of endogenous fatty acids, in particular 16:1 n-9 given its potential association with CHD risk. Experimental studies are also needed to investigate potential mechanisms via which 16:1 n-9 could affect CHD development.
**Figure legend**

**Figure 1** The major saturated and monounsaturated fatty acids in the DNL pathway, including myristic acid (14:0), palmitic acid (16:0), palmitoleic acid (16:1n-7), 7-hexadecenoic acid (16:1 n-9), stearic acid (18:0), vaccenic acid (18:1n-7), and oleic acid (18:1n-9).

**Online Supplemental Figure 1** Survival free of CHD in 2746 men and women, according to quartiles of baseline plasma phospholipids 16:0 and 16:1 n-9 levels.
Figure 1

*De novo* fatty acid synthesis

\[ \downarrow \]

14:0

\[ \downarrow \]

16:0

Elongation \(\rightarrow\) Desaturation

18:0 \(\rightarrow\) 16:1n-7

Desaturation \(\downarrow\)

18:1n-9 \(\rightarrow\) 18:1n-7

Elongation

16:1n-9

Beta-oxidation
Online Supplemental Figure 1

A. Plasma phospholipids 16:0
B. Plasma phospholipids 16:1 n-9

C. Plasma phospholipid 16:1n7
D. Plasma phospholipid 18:1n7

Survival Free of CHD

Year

First Quartile
Second Quartile
Third Quartile
Fourth Quartile
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Participants (n = 2746)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, y</td>
<td>73 (70-77)</td>
</tr>
<tr>
<td>Male, %</td>
<td>39</td>
</tr>
<tr>
<td>White Race, %</td>
<td>90</td>
</tr>
<tr>
<td>Education ≥ high school diploma</td>
<td>47</td>
</tr>
<tr>
<td>Current smoker, %</td>
<td>9</td>
</tr>
<tr>
<td>Treated diabetes, %</td>
<td>15</td>
</tr>
<tr>
<td>Treated Hypertension, %</td>
<td>42</td>
</tr>
<tr>
<td>Systolic blood pressure, mm Hg</td>
<td>134 (122-148)</td>
</tr>
<tr>
<td>BMI, kg/m²</td>
<td>26 (24-29)</td>
</tr>
<tr>
<td>Physical activity, kcal/wk</td>
<td>908 (315-1935)</td>
</tr>
<tr>
<td>LDL-C, mg/dL</td>
<td>126 (105-148)</td>
</tr>
<tr>
<td>HDL-C, mg/dL</td>
<td>52 (44-62)</td>
</tr>
<tr>
<td>Triglycerides, mg/dL</td>
<td>124 (90-173)</td>
</tr>
<tr>
<td>C-reactive protein, mg/L</td>
<td>2.4 (1.1-5.4)</td>
</tr>
<tr>
<td>Caloric intake, kcal/d</td>
<td>1938 (1618-2324)</td>
</tr>
<tr>
<td>Saturated fat, % calories</td>
<td>10 (9-12)</td>
</tr>
<tr>
<td>Carbohydrate, % calories</td>
<td>54 (49-58)</td>
</tr>
<tr>
<td>Alcohol, drinks/wk</td>
<td>0 (0-1)</td>
</tr>
</tbody>
</table>

1 Values are median (25th to 75th percentile) for continuous variables, or percentages for categorical variables.

2 Abbreviations: BMI, body mass index (calculated as weight in kilograms divided by the square of height in meters); DHA, docosahexaenoic acid; EPA, eicosapentaenoic acid; HDL-C, high density lipoprotein cholesterol; LDL-C, low density lipoprotein cholesterol.
Table 2. Spearman correlation coefficients among plasma phospholipids endogenous fatty acid biomarkers in 2746 men and women in the Cardiovascular Health Study\textsuperscript{1}

<table>
<thead>
<tr>
<th></th>
<th>14:0</th>
<th>16:0</th>
<th>16:1 n-7</th>
<th>16:1 n-9</th>
<th>18:0</th>
<th>18:1 n-7</th>
<th>18:1 n-9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean ± SD</td>
<td>0.28 ± 0.08</td>
<td>25.4 ± 1.6</td>
<td>0.50 ± 0.21</td>
<td>0.09 ± 0.03</td>
<td>13.4 ± 1.1</td>
<td>1.30 ± 0.20</td>
<td>7.6 ± 1.1</td>
</tr>
</tbody>
</table>

as percent of total fatty acids

<table>
<thead>
<tr>
<th></th>
<th>14:0</th>
<th>16:0</th>
<th>16:1 n-7</th>
<th>16:1 n-9</th>
<th>18:0</th>
<th>18:1 n-7</th>
<th>18:1 n-9</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:0</td>
<td>0.39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:1 n-7</td>
<td>0.60</td>
<td>0.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:1 n-9</td>
<td>0.23</td>
<td></td>
<td>0.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:0</td>
<td>-0.16</td>
<td>-0.44</td>
<td>-0.23</td>
<td>-0.16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:1 n-7</td>
<td>-0.21</td>
<td></td>
<td>0.17</td>
<td>0.31</td>
<td>-0.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:1 n-9</td>
<td>0.29</td>
<td>0.32</td>
<td>0.56</td>
<td>0.40</td>
<td>-0.28</td>
<td>0.24</td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{1}All shown correlations are significant at P<0.0001

Histograms of fatty acids:
Table 3 Hazard ratio (95% CI) for total CHD associated with plasma phospholipids endogenous fatty acids (n=2746, 485 cases)\(^1\)

<table>
<thead>
<tr>
<th>Fatty acids</th>
<th>Unadjusted(^2)</th>
<th>Adjusted for multiple variables(^2,3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16:0</td>
<td>1.10 (0.78-1.55)</td>
<td>0.92 (0.64-1.34)</td>
</tr>
<tr>
<td>16:1 n7c</td>
<td>0.81 (0.58-1.12)</td>
<td>0.95 (0.71-1.28)</td>
</tr>
<tr>
<td>16:1 n9c</td>
<td>2.04 (1.63-2.55)</td>
<td>2.06 (1.60-2.65)</td>
</tr>
<tr>
<td>18:1 n7c</td>
<td>1.39 (0.92-2.10)</td>
<td>1.14 (0.74-1.77)</td>
</tr>
</tbody>
</table>

\(^1\)Hazard ratio associated with inter-decile difference (between 10\(^{th}\) and 90\(^{th}\) percentile) for each fatty acid. The inter-decile range is calculated by calculating the inter-decile range (90\(^{th}\) percentile subtract 10\(^{th}\) percentile value for each exposure variable), then dividing each subject’s value of the exposure variable by this range. The calculated variable is then used in the Cox regression models. We calculated this to make the analysis comparable to previous work done in the group.

\(^2\)Separate models were fitted for each fatty acid.

\(^3\)Hazard ratios were calculated with the use of Cox proportional hazards regression models. Analysis adjusted for age (years), gender, race (whites versus non-whites), education (<high school, high school, >high school), smoking (non, previous and current smokers), diabetes (treated versus non-treated), hypertension (treated versus non-treated) and history of stroke or transient ischemic attack (yes versus no), BMI (kg/m\(^2\)), physical activity (kcal/wk), alcohol intake (number of drinks/wk), total fat intake (% energy), and energy per day (Kcal).
Table 4. Multivariate hazard ratio (95% CI) of total CHD, fatal CHD and non fatal myocardial infarction associated with phospholipids 16:1 n9 concentration

<table>
<thead>
<tr>
<th>Quartiles of 16:1 n9c</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>P for linear trend</th>
<th>Increase in 16:1 n9c (10th to 90th percentile)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total CHD Baseline to 5 y (n=268)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (reference)</td>
<td>0.52 (0.16-1.72)</td>
<td>2.12 (0.69-6.54)</td>
<td>9.59 (3.63-25.3)</td>
<td>&lt;0.001</td>
<td>2.66 (2.15-3.28)</td>
<td></td>
</tr>
<tr>
<td>2 (10th to 90th percentile)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total CHD 5 to 10 y (n=217)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (reference)</td>
<td>0.33 (0.12-0.91)</td>
<td>0.21 (0.07-0.62)</td>
<td>0.27 (0.08-0.89)</td>
<td>0.05</td>
<td>0.23 (0.07-0.74)</td>
<td></td>
</tr>
<tr>
<td><strong>CHD death Baseline to 5 y (n=128)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (reference)</td>
<td>0.28 (0.05-1.45)</td>
<td>0.91 (0.16-5.09)</td>
<td>4.82 (1.28-18.2)</td>
<td>0.002</td>
<td>2.26 (1.51-3.37)</td>
<td></td>
</tr>
<tr>
<td>2 (10th to 90th percentile)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CHD death 5 to 10 y (n=127)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (reference)</td>
<td>0.51 (0.11-2.45)</td>
<td>0.79 (0.17-3.71)</td>
<td>0.94 (0.19-4.73)</td>
<td>0.87</td>
<td>0.64 (0.19-2.21)</td>
<td></td>
</tr>
<tr>
<td><strong>Non-fatal MI Baseline to 5 y (n=163)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (reference)</td>
<td>1.08 (0.21-5.62)</td>
<td>4.12 (0.92-18.5)</td>
<td>18.3 (4.7-71)</td>
<td>&lt;0.001</td>
<td>2.59 (2.05-3.28)</td>
<td></td>
</tr>
<tr>
<td>2 (10th to 90th percentile)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Non-fatal MI 5 to 10 y (n=126)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (reference)</td>
<td>0.24 (0.07-0.78)</td>
<td>0.06 (0.01-0.24)</td>
<td>0.11 (0.03-0.44)</td>
<td>0.003</td>
<td>0.08 (0.02-0.38)</td>
<td></td>
</tr>
</tbody>
</table>

Hazard ratios were calculated with the use of Cox proportional hazards regression models. Analysis adjusted for age (years), gender, race (whites versus non-whites), education (< high school, high school, > high school), smoking (non, previous and current smokers), diabetes (treated versus non-treated), hypertension (treated versus non-treated) and history of stroke or transient ischemic attack (yes versus no), BMI (kg/m²), physical activity (kcal/wk), alcohol intake (number of drinks/wk), total fat intake (% energy), and energy per day (Kcal).
Table 5 Multivariable adjusted relationships between potential DNL inducing factors and plasma phospholipids levels of fatty acids (n = 2746)¹

<table>
<thead>
<tr>
<th></th>
<th>16:0</th>
<th>16:1 n7c</th>
<th>16:1 n9c</th>
<th>18:1 n7c</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD difference</td>
<td>P</td>
<td>SD difference</td>
<td>P</td>
<td>SD difference</td>
</tr>
<tr>
<td>Carbohydrate intake (5% energy)²</td>
<td>0.06</td>
<td>0.005</td>
<td>0.08</td>
<td>0.001</td>
</tr>
<tr>
<td>Protein intake (5% energy)²</td>
<td>0.21</td>
<td>&lt;0.001</td>
<td>0.13</td>
<td>0.003</td>
</tr>
<tr>
<td>Alcohol intake (drinks/day)</td>
<td>0.26</td>
<td>0.003</td>
<td>0.27</td>
<td>0.009</td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td>0.03</td>
<td>&lt;0.001</td>
<td>0.02</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Fasting insulin (IU/mL)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

¹Standardized differences in plasma phospholipids fatty acid concentrations for each potential DNL inducer. Associations were evaluated using linear regression with each biomarker as the dependent outcome. Only relationships with P<0.05 are presented. Analysis adjusted for all factors in the table and additionally adjusted for age (years), gender, recruitment site (4 sites), education (< high school, high school, > high school), race (white versus non-white) smoking (non, previous and current smokers) and energy per day (Kcal).

²The association between carbohydrate and protein intake with fatty acid biomarkers represent substitution for an equal amount of energy from total fat, in accordance to the multivariate nutrient density method (42).
Appendix A: Stata codes to generate new variables

*obtained database from Dr Mozaffarian*

use "C:\Documents and Settings\JASONWU\My Documents\Dataset and analysis\CHS_SCD analysis_2009\database\From Dary\CHS_FA_events_Dary_110210.dta", clear

*generate sampling weights*

*** Sampling Weights ***

****we exclude those with prevalent CHD at baseline****

**prior indicates if fatty acid is measured in previous assay run (coded=2) whereas new assay run is coded=1****

*****case indicates if the participant was a case from a previous case control study nested within the CHS(coded=1) and the controls are coded=0*****

egen newmeas = count(prior) if prior==1
egen oldmeasecase = count(prior) if prior==2 & case==1
egen oldmeascont = count(prior) if prior==2 & case==0

*there were 237 cases from the prior case control nested in the CHS study*****
*cases from case control study had probability =1 to be sampled*

gen psample = 1 if prior==2 & case==1

*There were 1411 possible controls identified for the previous case control study*  
*of these, we measured fatty acids from 276 people*

replace psample = oldmeascont/1411 if prior==2 & case==0

*****the probability of sampling for these 276 people was 276/1411=0.196*****
 *****all together, there were 513 fatty acid measurements from previous case control study***

*we have 2905 stored blood samples available from baseline*

replace psample = newmeas/2905 if prior==1

******there were 2233 new measurements, there probability of sampling was 2233/2905 = 0.77*****

label var psample "Probability of sampling"
gen sw = 1/psample
label var sw "Sampling weight"

The sampling scheme is summarized in the flow diagram below:
n = 5226
With stored plasma samples available for fatty acid measurement

Excluding n=673 with prevalent CHD
Left with n=4553

Cases for prior case-control study
n=237
Also identified 1411 possible controls, measured fatty acids in n=276
Total = 513

2905 stored plasma samples left, measured fatty acids in n=2233
Total fatty acid measurements in 2233 + 513 = 2746 people

Fatty acid measurements in the CHS for the current study

*generate new dietary intake variables*
*for those with 2 estimates use mean otherwise retain available measure*

egen prot_per_new = rowmean (prot_per prot_peryr8)
egen carb_per_new = rowmean (carb_per carb_peryr8)
egen tfat_per_new = rowmean (tfat_per tfat_peryr8)
egen satfat_per_new = rowmean (satfat_per satfat_peryr8)
egen monfat_per_new = rowmean (monfat_per monfat_peryr8)
egen poly_per_new = rowmean (poly_per poly_peryr8)
sum prot_per_new carb_per_new tfat_per_new
egen carboC_new = rowmean (carboC carboCyr8)
egen protC_new = rowmean (protC protCyr8)
egen tfatC_new = rowmean (tfatC tfatCyr8)
egen satfatC_new = rowmean (satfatC satfatCyr8)
egen monfatC_new = rowmean (monfatC monfatCyr8)
egen polyC_new = rowmean (polyC polyCyr8)

carboC_new protC_new tfatC_new
egen calor_new = rowmean (calor2 calor65)

drop if ihdbase==1
*generate standardized FA and other continuous variables*

impute gluy3 agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(gluy3_i)
impute bmiy3 agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(bmiy3_i)
impute insy3 agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(insy3_i)
impute kcaiy3 agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(kcaiy3_i)
impute ldladjy3 agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(ldladjy3_i)
impute hdly3 agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(hdly3_i)
impute trigy3 agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(trigy3_i)
impute crpy3 agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(crpy3_i)
impute fiby3 agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(fiby3_i)
impute income01 agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(income01_i)
impute tfat_per_new agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(tfat_per_new_i)
impute satfat_per_new agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(satfat_per_new_i)
impute carb_per_new agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(carb_per_new_i)
impute prot_per_new agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(prot_per_new_i)
impute monfat_per_new agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(monfat_per_new_i)
impute poly_per_new agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(poly_per_new_i)
impute carboC_new agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(carboC_new_i)
impute protC_new agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(protC_new_i)
impute tfatC_new agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(tfatC_new_i)
impute satfatC_new agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(satfatC_new_i)
impute monfatC_new agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(monfatC_new_i)
impute polyC_new agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(polyC_new_i)
impute calor_new agebl gend01 race smoker alcohy3 grade01 kcal2y3 bmiy3 diabbase strokebase, gen(calor_new_i)

*generate standardized FA and other continuous variables*
egen pf140_stnd=std(pf140)
egen pf160_stnd=std(pf160)
egen pf161n7c_stnd=std(pf161n7c)
egen pf161n9c_stnd=std(pf161n9c)
egen pf180_stnd=std(pf180)
egen pf181n7c_stnd=std(pf181n7c)
egen pf181n9c_stnd=std(pf181n9c)

egen age_stnd=std(age)
egen gluy3_i_stnd=std(gluy3_i)
egen insy3_i_stnd = std(insy3_i)
egen abpy3_stnd=std(abpy3)
egen kcsaly3_i_stnd=std(kcsaly3_i)
egen hdyly3_i_stnd= std(hdyly3_i)
egen crpy3_i_stnd=std(crpy3_i)
egen fiby3_i_stnd=std(fiby3_i)
egen epadha_stnd=std(epadha)
egen trans182_stnd=std(trans182)
egen trans181_stnd=std(trans181)
egen tfat_per_new_i_stnd =std(tfat_per_new_i)

***********************************************************************
*generate categorized FA, replace with median value in each category*
***********************************************************************

egen pf140_cat = cut(pf140), group(4) icode
bys pf140_cat: egen pf140_cat_med = median(pf140)

egen pf160_cat = cut(pf160), group(4) icode
bys pf160_cat: egen pf160_cat_med = median(pf160)

egen pf161n7c_cat = cut(pf161n7c), group(4) icode
bys pf161n7c_cat: egen pf161n7c_cat_med = median(pf161n7c)

egen pf161n9c_cat = cut(pf161n9c), group(4) icode
bys pf161n9c_cat: egen pf161n9c_cat_med = median(pf161n9c)

egen pf180_cat = cut(pf180), group(4) icode
bys pf180_cat: egen pf180_cat_med = median(pf180)

egen pf181n7c_cat = cut(pf181n7c), group(4) icode
bys pf181n7c_cat: egen pf181n7c_cat_med = median(pf181n7c)

egen pf181n9c_cat = cut(pf181n9c), group(4) icode
bys pf181n9c_cat: egen pf181n9c_cat_med = median(pf181n9c)

****************************************************
*generate new follow up time variable for each outcome*
****************************************************

gen tihdyrs_2=tihdyrs-year3
gen lhdyrs_2=lhdyrs-year3
gen nfmiyrs_2=nfmiyrs-year3

****************************************************
*generate inter-decile range*
****************************************************

global enlip pf160 pf161n7c pf161n9c pf181n7c pf140 pf180 pf181n9c

foreach x of global enlip{
egen temp10= pctile(`x'),p(10)
egen temp90= pctile(`x'),p(90)
gen intd_`x' = `x'/(temp90-temp10)
drop temp10 temp90
}

*sav DATABASE WITH NEW VARIABLES*

******************************************************************************
Appendix B: Stata codes for Cox regression models

***************open database**************
use "C:/Documents and Settings\JASONWU\My Documents\Dataset and analysis\CHS_SCD\analysis_2009\database\CHS_FA_events_Dary_Jason_dropIHD_110210.dta"

*************************************************************************************************
*  
*generate macro for covariates in each model and for inter-decile range for each endogenous lipid*
*************************************************************************************************
* 
*global cov_1 age_stnd i.gend01 i.race i.educ3 i.smoker i.diabbase i.htnmedy3 i.strokebase bmiy3_i 
 kcaly3_i_stnd alcohy3_i tfat_per_new_i_stnd calor_new_i 
*global cov_2 age_stnd i.gend01 i.race i.educ3 i.smoker i.diabbase i.htnmedy3 i.strokebase bmiy3_i 
 kcaly3_i_stnd alcohy3_i tfat_per_new_i_stnd calor_new_i epadha_stnd trans181_stnd trans182_stnd 
*global cov_3 age_stnd i.gend01 i.race i.educ3 i.smoker i.diabbase i.htnmedy3 i.strokebase bmiy3_i 
 kcaly3_i_stnd alcohy3_i tfat_per_new_i_stnd calor_new_i epadha_stnd trans181_stnd trans182_stnd 
crpy3_i_stnd lldadjy3_i hdly3_i_stnd trigy3_i fiby3_i_stnd sbpy3_stnd 

************set data for total IHD limit follow up to 10 years*****************
stset tihdyrs_2 [pw=sw], failure(tihd) enter(time 0) exit(time 10)

********************TIHD , model one include traditional MI and lifestyle risk factors************
xi:stcox $cov_1 intd_pf160, nohr schoenfeld(sch*) scaledsch(sca*)
estat phtest, d
drop sch* sca*

********
*pf161n7c*
********
********************TIHD , model one include traditional MI risk factors************
xi:stcox $cov_1 intd_pf161n7c, nohr schoenfeld(sch*) scaledsch(sca*)
estat phtest, d
drop sch* sca*

********
*pf161n9c*
********
********************TIHD , model one include traditional MI risk factors************
xi:stcox $cov_1 intd_pf161n9c, nohr schoenfeld(sch*) scaledsch(sca*)
estat phtest, d
drop sch* sca*

********
*pf181n7c*
********
********************TIHD , model one include traditional MI risk factors************
xi:stcox $cov_1 intd_pf181n7c, nohr schoenfeld(sch*) scaledsch(sca*)
estat phtest, d
drop sch* sca*

********
*pf181n9c*
********
********************TIHD , model one include traditional MI risk factors************
xi:stcox $cov_1 intd_pf181n9c, nohr schoenfeld(sch*) scaledsch(sca*)
estat phtest, d
drop sch* sca*

********
*pf140*
********
********
**********TIHD , model one include traditional MI risk factors**********
xi:stcox $cov_1 intd_pf140, nohr schoenfeld(sch*) scaledsch(sca*) robust
estat phtest, d
drop sch* sca*

********
*pf180*
********
**********TIHD , model one include traditional MI risk factors**********
xi:stcox $cov_1 intd_pf180, nohr schoenfeld(sch*) scaledsch(sca*)
estat phtest, d
drop sch* sca*

********
*pf181n9c*
********
**********TIHD , model one include traditional MI risk factors**********
xi:stcox $cov_1 intd_pf181n9c, nohr schoenfeld(sch*) scaledsch(sca*)
estat phtest, d
drop sch* sca*

**************************************************************************
*found Cox PH assumption violated for 160,161n9c, 140*
*repeat analysis using 2 follow up periods*
**************************************************************************
*follow up time split into 0-5, and 5-10 years*
**************************************************************************
forvalues x =5(5)10 {
  stset tihdyrs_2 [pw=sw], failure(tihd) enter(time `x'-5) exit(time `x')
  xi:stcox $cov_1 intd_pf160, nohr schoenfeld(sch*) scaledsch(sca*)
estat phtest, d
drop sch* sca*
}
forvalues x =5(5)10 {
  stset tihdyrs_2 [pw=sw], failure(tihd) enter(time `x'-5) exit(time `x')
  xi:stcox $cov_1 intd_pf161n9c, nohr schoenfeld(sch*) scaledsch(sca*)
estat phtest, d
drop sch* sca*
}
forvalues x =5(5)10 {
  stset tihdyrs_2 [pw=sw], failure(tihd) enter(time `x'-5) exit(time `x')
  xi:stcox $cov_1 intd_pf140, nohr schoenfeld(sch*) scaledsch(sca*)
estat phtest, d
drop sch* sca*
}

*******************************************************
*only 140 and 161n9c still associated with risk when*
*split up into 0-5 and 5-10 years*
*******************************************************
forvalues x =5(5)10 {
  stset tihdyrs_2 [pw=sw], failure(tihd) enter(time `x'-5) exit(time `x')
  xi:stcox $cov_1 intd_pf161n9c intd_pf140, nohr schoenfeld(sch*) scaledsch(sca*)
estat phtest, d
drop sch* sca*
}

*************************************************************************
*check for linear interaction with follow up time for 161n9c*
*************************************************************************
*set tihdyrs_2 [pw=sw], failure(tihd) enter(time 0) exit(time 10)
*xi: stcox $cov_1 intd_pf161n9c, tvc(intd_pf161n9c) texp(ln(_t))
*************************************************************************
*check if pf161n9c linearly associated with risk after multivariate adjustment*
*split analyses into 5 year intervals*
*check effect of additional covariates*
******************************************************************************
drop pf161n9c_cat pf161n9c_cat_med
egen pf161n9c_cat = cut(pf161n9c), group(4) icode
bys pf161n9c_cat: egen pf161n9c_cat_med = median(pf161n9c)
char pf161n9c_cat[omit] 0
forvalues x =5(5)10 {
    stset tihdyrs_2 [pw=sw], failure(tihd) enter(time `x'-5) exit(time `x')
    xi:stcox $cov_2 intd_pf161n9c, nohr
}
forvalues x =5(5)10 {
    stset tihdyrs_2 [pw=sw], failure(tihd) enter(time `x'-5) exit(time `x')
    xi:stcox $cov_3 intd_pf161n9c, nohr
}
forvalues x =5(5)10 {
    stset tihdyrs_2 [pw=sw], failure(tihd) enter(time `x'-5) exit(time `x')
    xi:stcox $cov_1 i.pf161n9c_cat, nohr
}
forvalues x =5(5)10 {
    stset tihdyrs_2 [pw=sw], failure(tihd) enter(time `x'-5) exit(time `x')
    xi:stcox $cov_1 pf161n9c_cat_med, nohr
}
******************************************************************************
*check for interaction between age, gender and 16:1 n9*
******************************************************************************
forvalues x =5(5)10 {
    stset tihdyrs_2 [pw=sw], failure(tihd) enter(time `x'-5) exit(time `x')
    xi3:stcox $cov_1 age_stnd*intd_pf161n9c, nohr
}
forvalues x =5(5)10 {
    stset tihdyrs_2 [pw=sw], failure(tihd) enter(time `x'-5) exit(time `x')
    xi3:stcox $cov_1 i.gend01*intd_pf161n9c, nohr
}
******************************************************************************
*repeat analysis for other outcomes*
******************************************************************************
***************IHD death************
***********************check linear effect********************
forvalues x =5(5)10 {
    stset ihdyrs_2 [pw=sw], failure(ihddea) enter(time `x'-5) exit(time `x')
    xi:stcox $cov_1 intd_pf161n9c, nohr
}
forvalues x =5(5)10 {
    stset ihdyrs_2 [pw=sw], failure(ihddea) enter(time `x'-5) exit(time `x')
    xi:stcox $cov_2 intd_pf161n9c, nohr
}
forvalues x =5(5)10 {
    stset ihdyrs_2 [pw=sw], failure(ihddea) enter(time `x'-5) exit(time `x')
    xi:stcox $cov_3 intd_pf161n9c, nohr
}
forvalues x =5(5)10 {
    stset ihdyrs_2 [pw=sw], failure(ihddea) enter(time `x'-5) exit(time `x')
    xi:stcox $cov_1 i.pf161n9c_cat, nohr
}
forvalues x =5(5)10 {
    stset ihdyrs_2 [pw=sw], failure(ihddea) enter(time `x'-5) exit(time `x')
    xi:stcox $cov_1 pf161n9c_cat_med, nohr
}
***************NFMI***************
*check linear effect*

**forvalues x =5(5)10 {**
  stset nfmiyrs_2 [pw=sw], failure(nfmi) enter(time `x'-5) exit(time `x')
  xi:stcox $cov_1 intd_pf161n9c, nohr
}**

**forvalues x =5(5)10 {**
  stset nfmiyrs_2 [pw=sw], failure(nfmi) enter(time `x'-5) exit(time `x')
  xi:stcox $cov_2 intd_pf161n9c, nohr
}**

**forvalues x =5(5)10 {**
  stset nfmiyrs_2 [pw=sw], failure(nfmi) enter(time `x'-5) exit(time `x')
  xi:stcox $cov_3 intd_pf161n9c, nohr
}**

**forvalues x =5(5)10 {**
  stset nfmiyrs_2 [pw=sw], failure(nfmi) enter(time `x'-5) exit(time `x')
  xi:stcox $cov_1 i.pf161n9c_cat, nohr
}**

**forvalues x =5(5)10 {**
  stset nfmiyrs_2 [pw=sw], failure(nfmi) enter(time `x'-5) exit(time `x')
  xi:stcox $cov_1 pf161n9c_cat_med, nohr
}**

**forvalues x =5(5)10 {**
  stset scdyrs_2 [pw=sw], failure(scd) enter(time `x'-5) exit(time `x')
  xi:stcox $cov_1 intd_pf161n9c, nohr
}**

**forvalues x =5(5)10 {**
  stset scdyrs_2 [pw=sw], failure(scd) enter(time `x'-5) exit(time `x')
  xi:stcox $cov_2 intd_pf161n9c, nohr
}**

**forvalues x =5(5)10 {**
  stset scdyrs_2 [pw=sw], failure(scd) enter(time `x'-5) exit(time `x')
  xi:stcox $cov_3 intd_pf161n9c, nohr
}**

**forvalues x =5(5)10 {**
  stset scdyrs_2 [pw=sw], failure(scd) enter(time `x'-5) exit(time `x')
  xi:stcox $cov_1 i.pf161n9c_cat, nohr
}**

**forvalues x =5(5)10 {**
  stset scdyrs_2 [pw=sw], failure(scd) enter(time `x'-5) exit(time `x')
  xi:stcox $cov_1 pf161n9c_cat_med, nohr
}**

---

**Appendix C: Stata codes for linear regression and multiple imputation methods**

use "C:\Documents and Settings\JASONWU\My Documents\Dataset and analysis\CHS_SCD analysis_2009\database\CHS_FA_events_Dary_Jason_dropIHD_110210.dta"

*******generate global macro for endogenous FA************

global enlip pf160_stnd pf161n7c_stnd pf161n9c_stnd pf181n7c_stnd

foreach x of global enlip{
xi: reg `x' age_stnd i.gend01 i.smoker i.educ3 i.clinic i.race calor_new_i
carb_per_new_i prot_per_new_i alcohy3_i bmiy3_i insy3_i [pw=sw]
lincom 5*carb_per_new_i
lincom 5*prot_per_new_i
}

foreach x of global enlip{
xi: reg `x' bmiy3_i insy3_i alcohy3_i prot_per_new_i carb_per_new_i calor_new_i [pw=sw]
}

***************check linear effect of non fat intake and alcohol intake*******
gen non_fat_per=carb_per_new_i+prot_per_new_i
egen non_fat_cat = cut( non_fat_per), group(4) icode
bys non_fat_cat: egen nonfat_cat_med = median(non_fat_per)
egen alcoh_cat = cut(alcohy3_i), group(4) icode
bys alcoh_cat: egen alcoh_cat_med = median(alcohy3_i)

foreach x of global enlip{
xi: reg `x' age_stnd i.gend01 i.smoker i.educ3 i.clinic i.race bmiy3_i insy3_i alcohy3_i prot_per_new_i carb_per_new_i calor_new_i [pw=sw]
}

*******set data as multiple imputation******

mi set wide

*******register variables as imputed or regular****

mi register imputed kcaly3 tfat_per_new calor_new prot_per_new bniy3
insy3 alcohy3 crpy3 ldldajy3 hdiy3 trigy3 fiby3

mi register regular age_stnd pf160_stnd pf161n7c_stnd pf161n9c_stnd pf181n7c_stnd race
clinic alcohy3d agebl gend01 smoker grade01 kcal2y3 diabada chdlmod sthlmmod pf160_cat
pf161n7c_cat pf181n9c_cat pf160_cat_med pf161n7c_cat_med pf161n9c_cat_med pf181n7c_cat_med
pf161n9c_cat_med angiainc angyrs2 ihyrs_2 ihddea sdyrs_2 sdiyrs_2 sctihhyrs_2 tihd
nfiyrs_2 nfmi pf160 pf161n7c pf161n9c pf181n7c pf140 pf180 pf181n9c educ3
income_recode diabbase htnmedy3 strokebase epadha_stnd trans181_stnd trans182_stnd
tfat_per_new_i_stnd kcaly3_i_stnd sbpy3_stnd intd_pf160 intd_pf161n9c intd_pf140
intd_pf161n7c intd_pf180 intd_pf181n7c intd_pf181n9c

*******clear original stset*****

mi stset, clear

*****impute variables with missing data*******

mi impute mvn calor_new carb_per_new prot_per_new tfat_per_new bniy3 kcaly3 insy3
alcohy3 crpy3 ldldajy3 hdiy3 trigy3 fiby3, add(3)

************************************************************************************
*generate macro for covariates in each model and for inter-decile range for each endogenous lipid**

**********************************************************************************
************

```plaintext
global cov_1 age_stnd i.gend01 i.race i.educ3 i.income_recode i.smoker i.diabbase
    i.htnmedy3 i.strokebase bmiy3 kcaly3 alcohy3 tfat_per_new calor_new

global cov_2 age_stnd i.gend01 i.race i.educ3 i.income_recode i.smoker i.diabbase
    i.htnmedy3 i.strokebase bmiy3 kcaly3 alcohy3 tfat_per_new calor_new epadha_stnd
    trans181_stnd trans182_stnd

global cov_3 age_stnd i.gend01 i.race i.educ3 i.income_recode i.smoker i.diabbase
    i.htnmedy3 i.strokebase bmiy3 kcaly3 alcohy3 tfat_per_new calor_new epadha_stnd
    trans181_stnd trans182_stnd crpy3 ldladjy3 hdly3 trigy3 fiby3 sbpy3_stnd i.anginainc
```

****stset data****

```plaintext
mi stset tihdyrs_2 [pw=sw], failure(tihd) enter (time 0) exit (time 10)
```

******run cox regressions*****

```plaintext
mi estimate: stcox $cov_1 intd_pf160
mi estimate: stcox $cov_1 intd_pf161n7c
mi estimate: stcox $cov_1 intd_pf161n9c
mi estimate: stcox $cov_1 intd_pf161n9c, tvc(intd_pf161n9c) texp(ln(_t))
mi estimate: stcox $cov_1 intd_pf181n7c
```

**************set data for total CHD first 5 years**************

```plaintext
mi stset tihdyrs_2 [pw=sw], failure(tihd) enter (time 0) exit (time 5)
```

******run cox regressions*****

```plaintext
mi estimate: stcox $cov_1 intd_pf161n9c
mi estimate: stcox $cov_2 intd_pf161n9c
mi estimate: stcox $cov_3 intd_pf161n9c
```

******run linear regressions******

```plaintext
global enlip pf160_stnd pf161n7c_stnd pf161n9c_stnd pf181n7c_stnd
foreach x of global enlip{
    xi: reg `x' age_stnd i.gend01 i.clinic i.smoker i.educ3 i.race calor_new
    carb_per_new prot_per_new alcohy3d bmiy3 insy3 [pw=sw]
    lincom 5*carb_per_new
    lincom 5*prot_per_new
}
```
Appendix D: Additional notes on statistical concepts developed for WPP

A. Inverse Probability of sampling weights

For statistical inference, we assume that the sample observations represent the population distribution. Departure from this assumption may lead to biased inference. For the current study, the measurement of plasma phospholipids fatty acids biomarkers used a sampling scheme which resulted in unequal selection probabilities for the target population (rather than simple random selection where all individuals in the entire CHS cohort with blood samples have an equal probability of being selected). This is because some of the measurements were carried out in a prior case-control study, where if the subject is a case, their probability of inclusion equals 1. On the other hand, the controls from this prior study had selection probability of 276/1411 = 0.1956, and the rest of the measurements (n = 2233) came from a pool of 2905 subjects (selection probability = 0.7687). Using sampling weights allows for correction for the unequal probability of selection and reduces bias in regression modeling (45). The weights are calculated as 1/probability of selection, for example, for the cases from the prior case-control study the weight associated with each individual is 1/1 = 1. The sample weights act as ‘inflation’ factors to represent the number of people in the original population that are accounted for by the sampling unit to which the weight is assigned; and a subject with a small probability of selection is seen as representing more individuals than those with higher selection probability. The result of applying sampling weights to the regression is that the estimators should be unbiased and consistent for the corresponding population quantities, however, the variance of the regression estimates also increase (45). We believe it is more important to obtain unbiased and consistent estimators, and hence applied the sampling weights in our analysis. This is shown using the phospholipids 18:1 n-7 as an example:

No sampling weights applied:

Cox regression -- Breslow method for ties
No. of subjects = 2744                     Number of obs   =      2744
No. of failures =          485
Time at risk    =  21835.25666
Log likelihood  =   -3616.7496                     Prob > chi2     =    0.0000
------------------------------------------------------------------------------
               _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 age_stnd |    1.42992   .0668715     7.65   0.000     1.304682     1.56718
 _Igend01_1 |   1.847396   .1863987     6.08   0.000     1.515917    2.251359
 _Irace_0 |   1.100066   .1605102     0.65   0.513     .8264569    1.464257
 _Ieduc3_1 |   1.059797   .1694782     0.47   0.639     .8315476    1.350699
 _Ieduc3_2 |   1.002593   .1194872     0.02   0.983     .7937436    1.263936
 _Ismoker_2 |   1.191836   .1225370     1.71   0.088     .9743191    1.457913
 _Ismoker_3 |   1.976892   .3178008     4.24   0.000     1.442599    2.709070
 _Dlabbase_1 |   1.79629   .2022897     5.20   0.000     1.440517    2.239932
 _Ihtnmedy3_1 |   1.515805   .1428594     4.41   0.000     1.260145    1.823334
 _Istrokeba~1 |   1.758522   .2676953     3.71   0.000     1.304883    2.369867
 bmiy3_i |      1.002   .0110602     0.18   0.856     .9805549    1.023913
 kcaly3_i_s~d |   .9195095    .048341    -1.60   0.110     .8294809    1.019309
 alcohy3_i |   1.00062   .0062087     0.10   0.920     .9885244    1.012863
 tfat_per_n~d |   .9957201   .0502914    -0.08   0.932     .9018724    1.099333
 calor_new_i |     1.0001   .0000843     1.18   0.236     .9999346    1.000265
 intd_pf18~7c |   1.025879   .1206643     0.22   0.828     .8146625    1.291859
------------------------------------------------------------------------------
With sampling weights:
Cox regression -- Breslow method for ties
No. of subjects      =  4267.856892                Number of obs   =      2744
No. of failures      =   665.852416
Time at risk         =  34605.83233
Wald chi2(18)   =    148.92
Log pseudolikelihood =   -3198.7282                Prob > chi2     =    0.0000
------------------------------------------------------------------------------
               _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 age_stnd |   1.418044   .0812432     6.10   0.000     1.267425    1.586562
 _Igend01_1 |   1.722843   .2124028     4.41   0.000     1.353021    2.193757
 _Irace_0 |    0.802992   .1453444    -0.80   0.425     .5998149    1.099561
 _Ieduc3_1 |   1.033039   .1501429     0.22   0.823     .7769668    1.373508
 _Ieduc3_2 |   1.019199   .1391050     0.14   0.889     .7959588    1.313234
 _Ismoker_2 |   1.198475   .1456745     1.49   0.137     .8768258    1.683933
 _Ismoker_3 |   2.063452   .3408009     3.64   0.000     1.396359    3.035933
 _Dlabbase_1 |   1.709209   .2096896     3.99   0.000     1.331433    2.224245
 _Ihtnmedy3_1 |   1.498848   .1722408     3.52   0.000     1.206579    1.887473
 _Istrokeba~1 |   2.055371   .3532413     3.95   0.000     1.419899    2.832251
 bmid3_i |    0.997130   .0124166    -0.23   0.817     .9730088    1.021766
 kcaly3_i_s~d |   .9611682   .0733889    -0.52   0.604     .8275741    1.116328
 alcohy3_i |   1.002719   .0054155     0.52   0.602     .9922604    1.013469
 tfat_per_n~d |   .9824841   .0596666    -0.29   0.771     .8722315    1.106673
 calor_new_i |   1.000059   .0001075     0.55   0.578     .9998483    1.000270
 intd_pf18~7c |   1.079965   .1431938     0.59   0.557     .8356514    1.395706
------------------------------------------------------------------------------
As expected, the standard error increased slightly from 0.12 to 0.14.
B. Imputation of missing values

Missing data is a common problem in all types of research studies. Several strategies have been developed to deal with missing data in regression analysis. The common approaches include: complete case analysis (exclude any subjects with at least one missing value on any of the predictors or outcome analyzed); indicator methods (create a new variable to indicate whether the predictor of interest is observed or missing); single imputation (essentially ‘fill in’ the missing value by replacing the missing value with the mean that is estimated from the observed value of other covariates) and multiple imputation (MI, same as single imputation but ‘filling in’ of the missing value is done multiple times) (46). The performance of these methods partly depends on the underlying assumption about the reason for missing data. For example: under missing completely at random (MCAR) assumption; the complete case analysis yields unbiased estimates but it is inefficient (reduced number of subjects available for analysis); whereas if the data are missing at random (MAR), the complete case analysis estimates are biased and inefficient. Recent studies suggest that both single imputation and MI are superior to complete case and indicator methods; as they provide unbiased estimates under the MAR assumption. Although single imputation could provide an over-estimation of the precision (standard error) compared to MI (47), recent studies suggest that single imputation method performed equally as well as the multiple imputation method when the percentage of missing data was low, (48). The percent of missing data for covariates used in the regression models in the current study are listed below:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Subjects with observed data (total n=2746)</th>
<th>% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>2746</td>
<td>0.0</td>
</tr>
<tr>
<td>Race</td>
<td>2746</td>
<td>0.0</td>
</tr>
<tr>
<td>Gender</td>
<td>2746</td>
<td>0.0</td>
</tr>
<tr>
<td>Education level</td>
<td>2746</td>
<td>0.0</td>
</tr>
<tr>
<td>Smoking status</td>
<td>2746</td>
<td>0.0</td>
</tr>
<tr>
<td>Treated diabetes</td>
<td>2746</td>
<td>0.0</td>
</tr>
<tr>
<td>Treated hypertension</td>
<td>2744</td>
<td>0.1</td>
</tr>
<tr>
<td>EPA</td>
<td>2746</td>
<td>0.0</td>
</tr>
<tr>
<td>DHA</td>
<td>2746</td>
<td>0.0</td>
</tr>
<tr>
<td>Trans fatty acids</td>
<td>2746</td>
<td>0.0</td>
</tr>
<tr>
<td>SBP</td>
<td>2746</td>
<td>0.0</td>
</tr>
<tr>
<td>DBP</td>
<td>2743</td>
<td>0.1</td>
</tr>
<tr>
<td>BMI</td>
<td>2728</td>
<td>0.7</td>
</tr>
<tr>
<td>Variable</td>
<td>Value 1</td>
<td>Value 2</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>physical activity</td>
<td>2735</td>
<td>0.4</td>
</tr>
<tr>
<td>LDL-C</td>
<td>2695</td>
<td>1.9</td>
</tr>
<tr>
<td>HDL-C</td>
<td>2741</td>
<td>0.2</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>2745</td>
<td>0.0</td>
</tr>
<tr>
<td>CRP</td>
<td>2715</td>
<td>1.1</td>
</tr>
<tr>
<td>Fibrinogen</td>
<td>2564</td>
<td>6.6</td>
</tr>
<tr>
<td>Alcohol intake</td>
<td>2745</td>
<td>0.0</td>
</tr>
<tr>
<td>Protein intake (% energy)</td>
<td>2681</td>
<td>2.4</td>
</tr>
<tr>
<td>Carbohydrate intake (% energy)</td>
<td>2681</td>
<td>2.4</td>
</tr>
<tr>
<td>Total energy intake</td>
<td>2681</td>
<td>2.4</td>
</tr>
</tbody>
</table>

Given the low level of missing data, we used single imputation which is unlikely to differ substantially from multiple imputation for the standard error estimates.

Stata obtain multiple imputation by simulating from a Bayesian posterior predictive distribution of the missing data under the conventional (or chosen) prior distribution. For our model we chose the mvn (multivariate normal) option because we were imputing multiple variables simultaneously and assume the pattern of missing is arbitrary. We carry out 3 multiple imputations (which should be sufficient due to the low percentage of missing data). Once we have multiply imputed data, we perform our primary analysis on each complete dataset and then use Rubin’s combination rule (Rubin, 1987) to form one set of results, i.e. each analysis will produce an association with standard error. Averaging the estimates should provide an unbiased combined estimate. The pooled standard error reflects both the uncertainty in the sampling (i.e. choosing random sample from the underlying population) as well as imputation (i.e. the uncertainty in the estimated underlying distributions of the variable with missing values), and will provide a more conservative estimate of the standard error than if single imputation method was used.
Table 3 Hazard ratio (95% CI) for total CHD associated with plasma phospholipids endogenous fatty acids (n=2746, 485 cases)\(^1\)

<table>
<thead>
<tr>
<th>Fatty acids</th>
<th>Adjusted for multiple variables- single imputation(^2,3)</th>
<th>Adjusted for multiple variables- multiple imputation(^2,3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16:0</td>
<td>0.92 (0.64-1.34)</td>
<td>1.11 (0.69-1.77)</td>
</tr>
<tr>
<td>16:1 n7c</td>
<td>0.95 (0.71-1.28)</td>
<td>0.97 (0.69-1.34)</td>
</tr>
<tr>
<td>16:1 n9c</td>
<td>2.06 (1.60-2.65)</td>
<td>1.73 (1.25-2.40)</td>
</tr>
<tr>
<td>18:1 n7c</td>
<td>1.14 (0.74-1.77)</td>
<td>1.12 (0.61-2.06)</td>
</tr>
</tbody>
</table>

\(^1\)Hazard ratio associated with inter-decile difference (between 10\(^{\text{th}}\) and 90\(^{\text{th}}\) percentile) for each fatty acid

\(^2\)Separate models were fitted for each fatty acid.

\(^3\)Hazard ratios were calculated with the use of Cox proportional hazards regression models. Analysis adjusted for age (years), gender, race (whites versus non-whites), education (<high school, high school, >high school), smoking (non, previous and current smokers), diabetes (treated versus non-treated), hypertension (treated versus non-treated) and history of stroke or transient ischemic attack (yes versus no), BMI (kg/m\(^2\)), physical activity (kcal/wk), alcohol intake (number of drinks/wk), total fat intake (% energy), and energy per day (Kcal).

**Conclusion**

As expected, the multiple imputation method did not significantly alter the results of our analysis except for slightly wider confidence intervals for some of the estimates. This is likely due to the low proportion of missing data in our dataset.
C. Adjustment for regression dilution

Prospective follow up studies often rely on a single ‘baseline’ exposure assessment, and investigate these characteristics with disease development over time. Due to a combination of measurement error and real biological variability (e.g. day to day variation in blood pressure), this approach results in mis-classification of the subjects ‘usual’ level of the risk factors under consideration. As a result, the observed association between baseline risk factor level and disease risk is smaller than if we were able to obtain the true ‘usual’ level of exposure (this is called ‘regression dilution bias’) (40). To account for this bias, the hazard ratio estimates from this study were adjusted according to published methods (40, 41), using repeat fatty acids measurements available among a subset of subjects taken 3 years apart (n = 75). All calculations were carried out in Microsoft Excel using beta-coefficients and standard errors obtained from Cox regression models. An example of the calculation is outlined here. From Cox regression, the coefficient associated with 16:0 and total CHD is -0.0587 (HR= 0.94) with standard error 0.14. The Spearman correlation coefficient for 16:0 based on the repeated measurement was 0.73. To calculate the regression dilution adjusted hazard ratio estimates:

\[ \beta_{\text{adjusted}} = \frac{\beta}{r} \]

\[ \text{SE}_{\text{adjusted}} = \frac{\text{SE}}{r} \]

95% CI = \( e^{(\beta_{\text{adjusted}} \pm \text{SE}_{\text{adjusted}})} \)

Numeric example (for 16:0):

\[ -0.0587/0.73 = -0.08, 0.14/0.73 = 0.19 \]

HR =\( e^{-0.08} = 0.92 \)

95% CI = \( (e^{-0.08-1.96*0.19}, e^{-0.08+1.96*0.19}) = (0.64, 1.34) \)
As expected, the adjustment decreased the hazard ratio estimate (from 0.94 to 0.92) but also widens the confidence interval, as a result, the $P$-value associated with the null hypothesis that the estimated coefficient equals 0 remains the same.
D. Multivariate nutrient density method

In epidemiological studies assessing the effects of nutrient intake on disease risk or other outcomes, it is important to take into account total energy intake (49). Firstly, total energy intake is largely a consequence of variations in physical activity, body size, and metabolic efficiency. Total energy intake is also associated with nutrient intake (they contribute to energy and those who eat more likely to have more of each nutrient). As physical activity, lean mass and metabolic efficiency may affect disease risk, adjusting for total energy intake allows us to control for the potential confounding effects due to these factors. Furthermore, in experimental studies if subjects gained weight it is not possible to figure out the effect of the specific nutrient. So the most common approach is to substitute nutrient intake while holding total energy the same (iso-energetic studies). By controlling for total energy we simulate this approach and allow assessment of the effect of changes in a particular dietary factor.

In our analyses, we adopted the multivariate nutrient density model (49). This includes both nutrient density (expressed as percent of total daily energy intake) and total energy, expressed in the equation below. The interpretation is that an increase in 1% of daily energy from carbohydrate (replacing 1% of energy from fat), while total energy intake is kept constant, is associated with $\beta_2$ amount of change in the outcome variable.

\[
\text{Outcome} = \beta_1 + \beta_2 \text{carbohydrate intake expressed as percent daily energy} + \beta_3 \text{protein intake expressed as percent daily energy} + \beta_4 \text{total energy} + \beta_5 \text{other covariates} \ldots
\]
E. Checking Cox model

In Cox proportional models it is important to assess the assumption that a variable is linearly associated with the log hazard. We did this by 3 different approaches: categorical analysis of each exposure (e.g. into quartiles), restricted cubic spline analysis, and trend test (by assigning each participant the median value in their quartile categories, and fitting this as a continuous variable in the model). The association between 16:1n-9 and total CHD in the first 5 years is shown here as an example to explain these methods.

We initially look at the relative risks by quartile categories of 16:1n-9:

```
| _t | Haz. Ratio | Robust Std. Err. | z | P>|z| | [95% Conf. Interval] |
|----|------------|------------------|---|-----|---------------------|
| age_stnd | 1.459796 | .112813 | 4.90 | 0.000 | 1.254617 | 1.698529 |
| _Igend01_1 | 1.691778 | .2999221 | 2.97 | 0.003 | 1.1952 | 2.394674 |
| _Irace_0 | .7068774 | .1572863 | -1.56 | 0.119 | .4570297 | 1.093311 |
| _Ieduc3_1 | 1.105265 | .2206311 | 0.50 | 0.616 | .7473967 | 1.634489 |
| _Ieduc3_2 | 1.196629 | .282861 | 0.94 | 0.347 | .8233265 | 1.73919 |
| _Iincome_r~0 | 1.444584 | .2604501 | 2.04 | 0.041 | 1.014556 | 2.056882 |
| _Iincome_r~2 | 1.171196 | .2503587 | 0.74 | 0.460 | .7703254 | 1.780678 |
| _Ismoker_2 | 1.062718 | .1823335 | 0.35 | 0.723 | .7592329 | 1.487514 |
| _Ismoker_3 | 2.144707 | .177238 | 3.16 | 0.002 | 1.33626 | 3.442268 |
| _Idiabbase_1 | 1.62938 | .274806 | 2.90 | 0.004 | 1.17118 | 2.268115 |
| _Ihtnmedy3_1 | 1.306744 | .195457 | 1.79 | 0.074 | .9747178 | 1.751871 |
| _Istrokeba~1 | 2.432155 | .5297221 | 4.08 | 0.000 | 1.587084 | 3.727199 |
| _bmi3_i | .9984974 | .015353 | -0.10 | 0.922 | .968555 | 1.029047 |
| kcaly3_i_s~d | .9486458 | .0792867 | -0.63 | 0.528 | .8053073 | 1.177497 |
| alcohol3_i | .9953661 | .011455 | -0.40 | 0.687 | .973166 | 1.018073 |
| tfat_per_n~d | .8853249 | .0721193 | -1.50 | 0.135 | .756804 | 1.036586 |
| calor_new_i | 1.000159 | .0001523 | 1.04 | 0.298 | .9998602 | 1.000457 |
| _Ipf161n9c~1 | .7016815 | .1537739 | -1.62 | 0.106 | .4566661 | 1.078155 |
| _Ipf161n9c~2 | .9143227 | .198462 | -0.41 | 0.680 | .9750088 | 1.399138 |
| _Ipf161n9c~3 | 1.434255 | .2980537 | 1.74 | 0.083 | .9544158 | 2.155338 |
```

We see a the first 2 quartiles of 16:1n9 appear to have similar risk with risk of total CHD.

However it is important to keep in mind that when assessing the model fit using this approach, we need to look at confidence intervals associated with each quartile and not just look at the point estimate. We look
at the plot of the median value of 16:1n-9 in each quartile against the beta-coefficients:

![Plot of median value of 16:1n-9 against beta-coefficients]

Which is a visual representation of the results in the Table above.

Next, we tested the linear trend by assigning all participants the median value of their respective quartile and test as a continuous variable:

|                | Haz. Ratio | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|----------------|------------|-----------|------|------|----------------------|
| age_stnd       | 1.470419   | .1124678  | 5.04 | 0.000 | 1.265714   1.708232 |
| _Igend01_1     | 1.719029   | .3069948  | 3.03 | 0.002 | 1.211352   2.439473 |
| _Irace_0       | .747266    | .1632276  | -1.33| 0.182 | .4870177  1.146584 |
| _Ieduc3_1      | 1.103304   | .2206262  | 0.49 | 0.623 | .745579   1.632709 |
| _Ieduc3_2      | 1.176176   | .2252704  | 0.85 | 0.397 | .8080597  1.711989 |
| _Iincome_r-0   | 1.439271   | .2579265  | 2.03 | 0.042 | 1.012982  2.044953 |
| _Iincome_r-2   | 1.171802   | .250806   | 0.74 | 0.459 | .7703139  1.782546 |
| _Ismoker_2     | 1.059504   | .1807934  | 0.34 | 0.735 | .7583225  1.480305 |
| _Ismoker_3     | 2.141485   | .5120095  | 3.18 | 0.001 | 1.340295  3.421602 |
| _Idiabase_1    | 1.601208   | .268197   | 2.81 | 0.005 | 1.153121  2.223418 |
| _Ihtnmedy3_1   | 1.331161   | .1995733  | 1.91 | 0.056 | .99236   1.785856 |
| _Istrokeba-1   | 2.428043   | .5239244  | 4.11 | 0.000 | 1.590683  3.706202 |
| bmi3_i         | .9979762   | .0152097  | -0.13| 0.894 | .9686066  1.028236 |
| kcal3_i_s-d    | .9452386   | .0796282  | -0.67| 0.504 | .8013738  1.114931 |
| alcohol3_i     | .9948378   | .0112104  | -0.46| 0.646 | .9731068  1.017054 |
| tfat_per_n-d   | .8853641   | .0719275  | -1.50| 0.134 | .7550397  1.038183 |
| calor_new_i    | 1.000153   | .0001519  | 1.01 | 0.313 | .9988556  1.000451 |
| pf16in9c_c-d   | 56710.28   | 251174.8  | 2.47 | 0.013 | 9.629777  3.34e+08 |

We reject the null hypothesis that there is no relationship (β associated with variable =0). The drawback of this approach this way is we will misclassify people and attenuate towards the null
(essentially ‘forcing’ people to take on the same values for the exposure variable) so might not detect true relationship which could be detected using variable as continuous variable.

Next we look at the restricted cubic spline graph of the associations:

The vertical red lines indicate the 10th, 25th, 50th, 75th and 90th percentiles of 16:1n-9. The grey areas are the 95% confidence intervals.

Essentially we see a graded response between the association of 16:1n-9 and risk of total CHD. We also checked if the non-linear terms in the restricted cubic spline model are significant, which again indicates lack of evidence to reject the null hypothesis that the relationship is linear:

**Stata output:**
```
. test rcs2 rcs3
( 1)  rcs2 = 0
( 2)  rcs3 = 0

chi2(  2) =    3.28
Prob > chi2 =    0.1944
```
For all the analysis presented in this study, we used these 3 approaches to assess the association between each of the primary fatty acids of interest and the primary outcomes. As overall these demonstrated relatively graded relationships we therefore focused on the continuous analyses.

**Check proportional hazards assumption**

The Cox model assumes a ‘proportional hazards model’, i.e. the β coefficient associated with each covariate remains constant over the follow up time. We can check this assumption by fitting a model which includes the interaction term between covariate ‘x’ with the natural logarithm of study time ‘ln(t)’. The resulting Wald statistic tests the null hypothesis that the proportional hazard is constant. An example is shown here using phospholipid 16:1n-9 as an example using Stata:

<table>
<thead>
<tr>
<th></th>
<th>Robust</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Coef.</td>
<td>Std. Err.</td>
</tr>
<tr>
<td>-------------</td>
<td>----------------------</td>
<td>--------------</td>
<td>--------------</td>
</tr>
<tr>
<td>rh</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age_stnd</td>
<td>0.3793708</td>
<td>0.0606372</td>
<td>6.26</td>
</tr>
<tr>
<td>_Igend01_1</td>
<td>0.4856701</td>
<td>0.1242698</td>
<td>3.91</td>
</tr>
<tr>
<td>_Irace_0</td>
<td>-0.464446</td>
<td>0.17944</td>
<td>-2.59</td>
</tr>
<tr>
<td>_Ieduc3_1</td>
<td>0.034304</td>
<td>0.1527903</td>
<td>0.22</td>
</tr>
<tr>
<td>_Ieduc3_2</td>
<td>0.0061507</td>
<td>0.1371721</td>
<td>0.04</td>
</tr>
<tr>
<td>_Iincome_r~0</td>
<td>0.321621</td>
<td>0.1326155</td>
<td>2.43</td>
</tr>
<tr>
<td>_Indiabbase_1</td>
<td>0.5155413</td>
<td>0.137226</td>
<td>3.76</td>
</tr>
<tr>
<td>_Ihtnmedy3_1</td>
<td>0.3058754</td>
<td>0.1135225</td>
<td>2.69</td>
</tr>
<tr>
<td>_Istrokeba~1</td>
<td>0.6976557</td>
<td>0.1968262</td>
<td>3.58</td>
</tr>
<tr>
<td>bmiy3_i</td>
<td>-.000624</td>
<td>0.0115836</td>
<td>-0.05</td>
</tr>
<tr>
<td>kcal_y3_i</td>
<td>-.0368853</td>
<td>0.0709064</td>
<td>-0.52</td>
</tr>
<tr>
<td>alcoy3_1</td>
<td>0.0012456</td>
<td>0.0069066</td>
<td>0.18</td>
</tr>
<tr>
<td>tfat_per_n~d</td>
<td>-0.051638</td>
<td>0.0573538</td>
<td>-1.03</td>
</tr>
<tr>
<td>color_new_i</td>
<td>0.0009886</td>
<td>0.0001077</td>
<td>8.82</td>
</tr>
<tr>
<td>intd_pf16--9c</td>
<td>0.2767334</td>
<td>0.0454224</td>
<td>6.09</td>
</tr>
<tr>
<td>-------------</td>
<td>----------------------</td>
<td>--------------</td>
<td>--------------</td>
</tr>
</tbody>
</table>

Note: second equation contains variables that continuously vary with respect to time; variables are interacted with current values of ln(_t).

The test statistics for 16:1n-9’s interaction with time was highly significant (P<0.001) we therefore reject the null hypothesis that the hazards are constant over time.
A plot of the scaled Schoenfeld residual and log of follow up time is also useful for checking these assumptions:

(The black line at y=0) is drawn for reference.

The plot indicates a possible slight negative slope over time, supporting potential non-proportional hazards for this covariate. The negative slope is also consistent with this variable being less important over time (which supports the possibility it is associated with CHD initially but not later on in the follow-up, possibly due to high temporal and analytical variation of this fatty acid).
Results for other fatty acids of interest:

16:0, \( P \) for interaction with log of study time = 0.02.

(The black line at y=0) is drawn for reference.

The plot indicate a possible slight positive slope over time. Together with the observed significant interaction with time, supports potential non-proportional hazards for this covariate.
16:1n7, $P$ for interaction with log of study time = 0.66.

(The black line at $y=0$) is drawn for reference. No significant evidence for non-proportional hazards.
18:1n7, $P$ for interaction with log of study time = 0.20.

(The black line at $y=0$) is drawn for reference. No significant evidence for non-proportional hazards.

**Overall model fit**

Next we check the model fit using Cox-Snell residuals. Using total incident CHD over the entire follow up period as an example. If the Cox proportional hazard model is appropriate, then the Cox-Snell residuals should have the exponential distribution with hazard function equal to 1, for all follow up times. So we compute Cox-Snell residuals, regard them as failure times, and compute their cumulative hazard function which should follow a 45 degree line.
As the points all follow the 45 degree line, the graphs suggest reasonably satisfactory overall fit.

This was also found to be the case for other Cox models in our analyses.
References


49. Willett WC, Howe GR, Kushi LH. Adjustment for total energy intake in epidemiologic studies. Am J Clin Nutr 1997;65:1220S-1228S; discussion 1229S-1231S.