**Supplemental Methods**

EPIC-CVD

EPIC-CVD is, a case-cohort study nested within the European Prospective Investigation into Cancer and Nutrition (EPIC) study (1). EPIC consists of 366,521 women and 153,457 men, aged between 35 and 70 years at baseline, and recruited from 23 centres across 10 European countries (Denmark, France, Germany, Greece, Italy, The Netherlands, Norway, Spain, Sweden and the UK) between 1992 and 2000. EPIC complies with the Declaration of Helsinki, and all participants gave written informed consent before participating in this study. The study was approved by the local ethics committees of the participating centres and the Institutional Review Board of the International Agency for Research on Cancer (IARC, Lyon). For EPIC-CVD, a representative subcohort of 17,634 participants was selected by simple random sampling, stratified by centre, from participants who had available stored blood and buffy coat samples (n=385,747) (2, 3). The subcohort was supplemented with additional stroke and CHD cases that occurred during follow-up.

For this analysis, we included data from 27,072 non-related individuals with information on hard call or imputed rs4988235 genotype. After exclusion of participants with no information on dairy consumption (n=114), non-cases for stroke or CHD outside the subcohort (n=450), participants with an inclusion date that is identical to censoring date (n=398), participants from Norway due to the low number of participants (n=52), and participants from Greece (n=1,318), 13,114 subcohort participants and 3,737 additional stroke cases and 8,985 CHD cases remained (Figure 1). Of the total stroke cases, 2,746 (66%) could be identified as ischaemic. By design, 307 stroke and 307 stroke cases were part of the subcohort that was drawn at baseline.



**Figure 1. Flowchart for participants of EPIC-CVD study.**

*Genotyping*

For EPIC-CVD, DNA was extracted from buffy coat from citrated blood plasma. EPIC-CVD participants were genotyped with the Illumina Human Core Exome Chip array, Human Core Exome array, Illumina 660 Quad array and Omni Exome Express array. Single nucleotide polymorphisms (SNPs) were removed with minor allele count <2 or Hardy Weinberg p-value <1e-6. Missing genotypes were imputed using the 1000 Genomes reference panel.

Information on rs4988235 genotype was available from hard call data for participants genotyped using the Illumina HumanCore Exome Chip array (n=16,685). Hard call data means that this variant was directly genotyped using the array. For participants genotyped to the Illumina HumanCoreExome, Illumina OmniExomeExpress, or Illumina Quad660 array, rs4988235 genotype was imputed (n=8,055, impute info~0.42). This means that information from directly genotyped SNPs and a reference panel was used to obtain a probable genotype for each individual. This probable genotype is reported as a continuous variable ranging from 0 (C/C, lactase non persistent) to 2 (T/T, lactase persistent) called a SNP dosage to reflect any uncertainty of the imputation.

Principal components were calculated and used to adjust for potential population stratification (4).

*Dietary intake*

Dietary intake over the previous twelve months before study inclusion was assessed at baseline through diet questionnaires, which varied per country or study centre. These questionnaires were developed to reflect local eating habits and validated locally (1, 5, 6). Dietary intake data was not calibrated.

Information on intake of milk ((semi-)skimmed or full-fat, regardless of fermentation), yoghurt and thick fermented milk (e.g. sour milk) and cheese was available for the full subcohort. Availability of consumption data for dairy creams (e.g. whipped cream), curd (e.g. quark, cottage cheese), milk based puddings (e.g. custard), milk beverages (e.g. chocolate milk), and milk for coffee and creamers, differed by country and/or centre, depending on the cohort-specific questionnaire. Information on consumption of non-milk dairy was calculated by summing intake of all dairy products other than milk.

*Assessment of covariates*

Baseline questionnaires included questions on demographic and cardiovascular risk factors. Smoking status was categorised as current, former, never. Education level was categorised into low (no schooling/primary school), medium (secondary school) and high (vocational education/university). Physical activity was categorized using the Cambridge Physical Activity Index into inactive, moderately inactive, moderately active and active (7). Prevalent high cholesterol and diabetes was self-reported.

A physical examination was also performed. Systolic and diastolic blood pressure (mmHg) were measured twice. If both measurements were available, the mean of two measurements was used. High blood pressure was defined as any of self-reported hypertension, self-reported use of anti-hypertensive medication, systolic blood pressure >140 mmHg or diastolic blood pressure >90 mmHg.

All centres used trained professionals to measure height and weight except the French centre, for which self-reported measures were used for a subset of participants, and Oxford, for which recalibrated self-reported measures were used based on a comparison between self-reported and measured data in a subset of participants. Both height and weight were adjusted for clothing worn (1, 8). Body mass index (BMI) was calculated as weight divided by the square of height in metres. Waist and hip circumference (cm) were measured, and waist-to-hip ratio (WHR) was calculated by dividing the mean waist by mean hip circumference.

Blood samples were collected for approximately 70% of the participants and stored in liquid nitrogen at –196°C. Serum biomarkers were measured in baseline non-fasted samples at Stichting Huisartsen Laboratorium (Etten-Leur, The Netherlands) and included total cholesterol, high-density lipoprotein cholesterol (HDL-c) and triglycerides. Biomarkers were measured using a Cobas enzymatic assay (Roche Diagnostics, Mannheim, Germany) on a Roche HitachiModular P analyzer.

*Endpoints*

Methods used in the recruitment centres to determine first non-fatal stroke and CHD events included self-report and linkage with morbidity or hospital registries. Non-fatal events were further validated by a review of medical records, imaging evidence (stroke) and/or linkage with registries. Fatal events were generally determined through mortality registries (3). Relevant ICD codes are noted below.

|  |  |  |
| --- | --- | --- |
| **Endpoint**  | **ICD9 codes** | **ICD10 codes** |
| Stroke | 430-434, 436-438 | I60-I69 |
|  Ischaemic  | 433, 434 | I63 |
| CHD | 410-414 | I20-I25 |

EPIC-NL cohort

The European Prospective Investigation into Cancer and Nutrition (EPIC)-NL cohort (n=40,011) consists of the Prospect-EPIC cohort and the Monitoring Project on Risk Factors for Chronic Diseases (MORGEN)-EPIC cohort. These cohorts were set-up simultaneously between 1993 and 1997. The Prospect cohort consists of 17,357 women aged 49-70, living in Utrecht or its vicinity, who participated in a Dutch breast cancer screening program. The MORGEN cohort consists of 22,654 men and women, aged 20-59 who were recruited through age stratified random samples of three Dutch towns (Amsterdam, Doetinchem and Maastricht).

All participants provided written informed consent before study inclusion. The study complied with the Declaration of Helsinki and was approved by the medical ethics committee of the Netherlands Organization for Applied Scientific Research (TNO) (MORGEN), and the institutional review board of the University Medical Centre Utrecht (Prospect). Detailed description of the design and rationale of this cohort can be found elsewhere (9).

A random sample (n=2,245) of the baseline EPIC-NL cohort for whom no previous genotyping was performed (i.e. excluding participants selected for the random subcohort or included as additional case for EPIC-InterAct or EPIC-CVD, and incident cancer cases who have been genotyped for EPIC cancer consortia) was selected for genotyping. In addition, 674 stroke and 1,107 CHD cases for whom no previous genotyping was attempted were identified.

We included participants from the EPIC-NL subcohort and additional stroke and CHD cases with information on rs4988235 genotype (n=3,331). Participants with missing data on intake of dairy products (n=14) or CVD endpoints (n=3) were excluded. From each pair of related individuals (identity by descent > 0.1875), we excluded 1 participant (n=44), leaving a total of 3,270 participants for analysis in this study, including 567 total stroke and 843 CHD cases. Of the total stroke cases, 410 (72%) could be identified as ischaemic. By design, 50 stroke cases and 67 CHD cases were a part of the subcohort that was drawn at baseline.



**Figure 2. Flowchart for participants of EPIC-NL study.**

*Genotyping*

DNA was extracted from buffy coat on an TECAN EVO/HSM instrument (TECAN Freedom EVO 2.0 and Promega HSM 2.0) extraction system with ReliaPrep (TM) Large Volume HT gDNA System (Art. Nr. A2751) chemistry (Prospect and MORGEN), or using a salting out method (MORGEN). Participant samples were genotyped with the GSA BeadArray. Sample exclusion criteria were low call rate (<97%), ambiguous sex (X chromosome homozygosity rate between 0.2 and 0.8), discordance between self-reported sex and sex based on X chromosome homozygosity, outlier for heterozygosity (mean heterozygosity ± 3 SD), duplicated individual or divergent ancestry. Marker exclusion criteria were deviation from HWE at p < 1\*10-6, MAF < 1% or marker missingness >3%. Imputation to the Haplotype Reference Consortium panel (Version r1.1 2016) using minimac3 was performed at the Michigan Imputation Server (10). Imputation quality of rs4988235 was high (R2=0.91). Principle components are a means to summarize multi-dimensional data in fewer new dimensions. Instead of plotting data on multiple axes (one axis for every variable in the dataset) new axes (principle components) are used that are perpendicular to each other and are chosen in a way that they reflect the maximal variation in the data. Principal components were calculated in Plink(11) to adjust for population stratification(4) using all hard call SNPs that passed quality control in the genetic EPIC-NL dataset. The first two principal components explained respectively 11.6% and 7.4% of genetic variation.

*Dietary intake*

The EPIC-NL cohort used a self-administered food frequency questionnaire (FFQ) to assess the food consumption (in times per day, per week, per month or per year, or as never) of 79 main food categories during the year before enrolment. The questionnaire also contained colour photographs with portion sizes of 21 foods. The FFQ gives an estimation of the average daily consumption of 178 food items (9, 12).

The FFQ provided information on intake of milk ((semi-)skimmed or full-fat, regardless of fermentation) and milk subtypes i.e. unfermented and unsweetened milk, fermented milk, sweetened milk. We also had information on consumption of curd (e.g. quark, cottage cheese), yogurt and cheese. Information on consumption of non-milk dairy was calculated by summing intake of all dairy products other than milk.

Validity of the FFQ was assessed through comparison with 12 24-hours recalls among 121 men and women. Spearman’s rank correlation coefficients showed good validity for milk and milk products in men (0.69) and women (0.77) (12). Dietary intake data was not calibrated.

*Assessment of covariates*

A general questionnaire was administered at baseline, which included questions on demographic characteristics and cardiovascular risk factors. Educational level was categorized into low (primary till intermediate vocational education), moderate (higher general secondary education completed or till 3rd year with success), and high (higher vocational education and university). Smoking status was categorized as never, former, or current smoker. Physical activity was assessed using a validated questionnaire (7). The Cambridge Physical Activity Score (13) was then calculated and used to categorise physical activity into inactive, moderately inactive, moderately active and active. Missing Cambridge physical activity scores (14%) were imputed by means of single linear regression modelling (SPSS MVA procedure). Presence of hypercholesterolemia was self-reported.

Furthermore, a physical examination was performed. Body weight (kg) and height (cm) were measured and BMI was calculated by dividing weight (kg) by height (m) squared (kg/m2). Waist and hip circumference (cm) were measured twice, and waist-to-hip ratio (WHR) was calculated by dividing the mean waist by mean hip circumference. Systolic- and diastolic blood pressure (mmHg) were measured twice on participants’ left arm, while in the supine position, with a Boso Oscillomat (Prospect) or with a random-zero sphygmomanometer (MORGEN). The mean of these two measurements was used. Hypertension was defined when one of the following criteria was met: systolic blood pressure > 140 mmHg; diastolic blood pressure >90 mmHg; self-reported use of antihypertensive medication, or self-reported physician-diagnosed hypertension.

Total cholesterol was measured using enzymatic method, and high-density lipoprotein (HDL)-cholesterol and low-density lipoprotein (LDL)-cholesterol were measured using a homogeneous assay with enzymatic endpoint on an autoanalyzer (LX20, Beckman Coulter, Mijdrecht, the Netherlands) (9).

*Assessment of CHD and stroke*

To obtain cases of incident stroke and CHD, participants were followed over time by linkage to mortality and hospital discharge registers. Data on hospital discharge diagnoses were provided by the National Medical Registry, using the Dutch Hospital Discharge Diagnosis Database. Vital status was obtained through linkage with the municipal population registries. Primary and secondary causes of death were obtained from Statistics Netherlands.

In the Dutch Hospital Discharge Diagnosis Database, endpoints were defined through International Classification of Diseases (ICD)-9 codes. Causes of death were coded according to ICD-10. Relevant ICD codes are noted in the table below.

|  |  |  |
| --- | --- | --- |
| **Endpoint**  | **ICD9 codes** | **ICD10 codes** |
| Stroke | 430-434, 436 | I60-I66 |
|  Ischaemic  | 433, 434 | I63, I65 |
| CHD | 410-414, 427.5, 798.1, 798.2, 798.9 | I20-I25, I46, R96 |

UK Biobank

The UK Biobank ([www.ukBiobank.ac.uk](http://www.ukbiobank.ac.uk)) is a population based study that recruited over 500,000 men and women (aged 37 to 73 years) between 2006 and 2010 (application number 29916). We used data from 367,643 participants (including 29,278 CHD and 9,652 stroke cases) that were successfully genotyped using the Affymetrix UK BiLEVE Axiom array and the Affymetrix UK Biobank Axiom Array (14). rs498825 was imputed to a merged reference panel of the UK10K haplotype reference panel and the 1000 Genomes Phase 3 reference panel in IMPUTE2, resulting in a high imputation quality (impute info=0.97). Cases were obtained both at baseline by touchscreen questionnaire and hospital records and validated by a nurse interview, and during follow-up from hospital records and death certificates. The UK Biobank study was approved by the North West Multi-Centre Research Ethics Committee and all participants provided written informed consent to participate.

CARDIoGRAM GWAS

CARDIoGRAM GWAS is a European ancestry meta-analysis of 22 GWAS case-control studies imputed to HapMap2 involving 22,233 CHD cases and 64,762 controls (15). Data have been downloaded for this analysis from [www.CARDIOGRAMPLUSC4D.ORG](http://www.CARDIOGRAMPLUSC4D.ORG) on November 6, 2018. EPIC-Norfolk, which is a part of EPIC-CVD, was included in CARDIoGRAM GWAS.

MEGASTROKE

MEGASTROKE provides single genetic variant associations from a fixed-effect meta-analysis among European-ancestry participants, including 406,111 controls and 40,585 stroke cases (16). Data on stroke risk were downloaded from <http://www.megastroke.org/> on May 15, 2018. EPIC-CVD was included in the MEGASTROKE analysis.

Literature Cited.

1. Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public health nutrition. 2002;5(6b):1113-24.

2. Langenberg C, Sharp S, Forouhi NG, Franks PW, Schulze MB, Kerrison N, et al. Design and cohort description of the InterAct Project: an examination of the interaction of genetic and lifestyle factors on the incidence of type 2 diabetes in the EPIC Study. Diabetologia. 2011;54(9):2272-82.

3. Danesh J, Saracci R, Berglund G, Feskens E, Overvad K, Panico S, et al. EPIC-Heart: the cardiovascular component of a prospective study of nutritional, lifestyle and biological factors in 520,000 middle-aged participants from 10 European countries. European journal of epidemiology. 2007;22(2):129-41.

4. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006;38(8):904-9.

5. Kroke A, Klipstein-Grobusch K, Voss S, Moseneder J, Thielecke F, Noack R, et al. Validation of a self-administered food-frequency questionnaire administered in the European Prospective Investigation into Cancer and Nutrition (EPIC) Study: comparison of energy, protein, and macronutrient intakes estimated with the doubly labeled water, urinary nitrogen, and repeated 24-h dietary recall methods. The American journal of clinical nutrition. 1999;70(4):439-47.

6. Margetts BM, Pietinen P. European Prospective Investigation into Cancer and Nutrition: validity studies on dietary assessment methods. International journal of epidemiology. 1997;26 Suppl 1:S1-5.

7. Haftenberger M, Schuit AJ, Tormo MJ, Boeing H, Wareham N, Bueno-de-Mesquita HB, et al. Physical activity of subjects aged 50-64 years involved in the European Prospective Investigation into Cancer and Nutrition (EPIC). Public health nutrition. 2002;5(6B):1163-76.

8. Haftenberger M, Lahmann PH, Panico S, Gonzalez CA, Seidell JC, Boeing H, et al. Overweight, obesity and fat distribution in 50- to 64-year-old participants in the European Prospective Investigation into Cancer and Nutrition (EPIC). Public health nutrition. 2002;5(6b):1147-62.

9. Beulens JW, Monninkhof EM, Verschuren WM, van der Schouw YT, Smit J, Ocke MC, et al. Cohort profile: the EPIC-NL study. International journal of epidemiology. 2010;39(5):1170-8.

10. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nature genetics. 2016;48(10):1284-7.

11. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. 2007;81(3):559-75.

12. Ocke MC, Bueno-de-Mesquita HB, Goddijn HE, Jansen A, Pols MA, van Staveren WA, et al. The Dutch EPIC food frequency questionnaire. I. Description of the questionnaire, and relative validity and reproducibility for food groups. International journal of epidemiology. 1997;26 Suppl 1:S37-48.

13. Wareham NJ, Jakes RW, Rennie KL, Schuit J, Mitchell J, Hennings S, et al. Validity and repeatability of a simple index derived from the short physical activity questionnaire used in the European Prospective Investigation into Cancer and Nutrition (EPIC) study. Public health nutrition. 2003;6(4):407-13.

14. UK Biobank. Protocol for a large-scale prospective epidemiological resource. Available at <www.ukbiobank.ac.uk/resources/> 2006.

15. Schunkert H, Konig IR, Kathiresan S, Reilly MP, Assimes TL, Holm H, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nature genetics. 2011;43(4):333-8.

16. Malik R, Chauhan G, Traylor M, Sargurupremraj M, Okada Y, Mishra A, et al. Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. Nature genetics. 2018;50(4):524-37.