**The role of risk factors across the spectrum of cardiovascular disease presentations.   
A UK Biobank study.**

**Statistical analysis plan**

Nathalie Conrad, Mark Woodward, 25 May 2025.

**Aims**

This study aims to perform a comprehensive analysis of risk factors and their associations across the spectrum of cardiovascular disease presentations.

**Methods**

**Data Source**

We will use data from the [UK Biobank](https://www.sciencedirect.com/topics/medicine-and-dentistry/uk-biobank" \o "Learn more about UK Biobank from ScienceDirect's AI-generated Topic Pages), a population-based [prospective cohort study](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/prospective-cohort-study" \o "Learn more about prospective cohort study from ScienceDirect's AI-generated Topic Pages) comprising 500 000 individuals from the UK, aged 40–69 years at recruitement between 2006 and 2010.

**Study population**

We will include all men and women from the UK Biobank cohort. Baseline date will be set as the date of first assessment center visit. Participants with prevalent cardiovascular disease at baseline will be excluded.

**Study endpoints**

The primary study endpoint is incident cardiovascular disease (CVD). To best characterise the spectrum of CVD, we will examine the following conditions: ischaemic heart disease (acute coronary syndrome and other ischaemic heart disease); aortic aneurysm; aortic stenosis; atrial fibrillation and flutter; heart failure; peripheral arterial disease; second- and third-degree heart block; stroke; and venous thromboembolism. We will consider diseases individually and as a composite outcome of all CVD combined (defined as the first recorded event across all conditions investigated).

A series of substudies will investigate secondary endpoints, focussing on specific disease subtypes including: i) acute coronary syndromes (ST-segment elevation myocardial infarction (STEMI) and non- ST-segment elevation myocardial infarction (NSTEMI)); ii) heart failure (following ischaemic heart disease and other); iii) heart block (following ischaemic heart disease and other); subtypes of stroke (ischaemic, intracerebral haemorrhage, subarachnoid haemorrhage); iv) venous thromboembolism (deep vein thrombosis and pulmonary embolism).

Cardiovascular events will be extracted from secondary care admissions data and death certificates, in any diagnostic position. Code lists for event identification will be established following previously established methods1–3 and include diagnoses (International Classification of Diseases, tenth revision (ICD-10)) and procedures (UK Office of Population Census and Surveys classification (OPCS-4)).

Two sets of analyses will be performed - first only considering events recorded during the patient lifetime (i.e. from hospital admissions data), and second, considering both fatal and non-fatal events (ie. also including events recorded on death certificates).

**Exposures**

Covariates will be selected to represent a range of well-established cardiovascular risk factors, and included all those used in the most recent cardiovascular risk scores endorsed by European and American heart associations or recommended for screening in the latest cardiovascular disease prevention guidelines.4–6 We will complement this list with a small set of additional markers (eg. Lp(a) and C-reactive protein), based on recent literature substantiating their importance in cardiovascular risk assesment.7,8

Demographic variables will include the participant's age, sex (male or female), and socioeconomic status. Socioeconomic status will be described using the Townsend deprivation index.

Clinical and lifestyle variables will include systolic blood pressure, body mass index (BMI), type 2 diabetes, and smoking status, and will be extracted by combining self-reported questionnaires, clinical measurements, and elctronic health record linkage. Biomarkers will consider non high-density lipoprotein (HDL) cholesterol, C-reactive protein (CRP), Lipoprotein A (Lp(a)), urine albumin-creatinine ratio (UACR), and eGFR. eGFR will be calculated using the Chronic Kidney Disease Epidemiology Collaboration 2021 creatinine–cystatin C equation.9,10

To facilitate clinical interpretation and comparisons with previous literature, risk factor variables will be categorised into relevant groups (eg. BMI will be categorised into "underweight" (<18.5 kg/m2), "normal" (18.5-24.9 kg/m2), "overweight" (25-29.9 kg/m2), and "obesity" (>=30 kg/m2), and biomarkers will be categorised into sex-specific fifths).

Information on medication use at baseline will be extracted from self-reported health questionnaires and include cholesterol lowering as well as blood pressure lowering medications. Among the subset of participants with linkage to primary care records, we will further extract drug prescriptions during follow-up, to allow for sensitivity analyses stratified by ‘ever-use’ of medication, and analyses censoring follow-up time at the date of first prescription.

**Statistical analyses**

We will present patient characteristics as frequencies (%) for categorical data, means and standard deviation (SD) for symmetrically distributed continuous data, or medians and interquartile interval for non-symmetrically distributed continuous data, over the whole cardiovascular disease cohort and stratified by sex.

For exposure variables with missing data, we will impute data using multiple imputation by chained equations, with 10 imputed datasets, 50 iterations, and predictive mean matching for continuous variables or proportional odds model for ordered categorical variables (smoking status).11 For variables categorised for the purpose of the study (biomarker measures, BMI or socioeconomic status), we will perform imputation on continuous measures before categorisation. Imputation models will include all risk factor exposure and CVD outcome variables as predictors. Estimates and standard errors will be obtained with use of Rubin’s rules to combine the results of the separate analyses of individual imputed datasets.

For each condition investigated, we will define time at risk to start at participants’ baseline date, and to stop at the earliest of death, incidence of the disease of interest, or follow-up end date (30/05/2022).

To calculate overall and category-specific hazard ratios (HR) and corresponding 95% confidence intervals (CI) for new-onset cardiovascular disease, we will use Cox proportional hazard models. We will examine the proportionality of the hazard ratio visually.

To facilitate clinical interpretation and comparisons with previous literature, we will report hazard ratios for models considering continuous variables as continuous, as well as categorised values. We will present results from models adjusted for age and sex; as well as fully adjusted models considering all risk factors and biomarkers examined in this study.

As series of substudies will investigate shapes of association for each individual risk factor and individual CVD using restricted cubic splines and both age-sex- and fully adjusted models.

We will further perform stratified analyses by subgroups (including by age, sex, risk factor exposure, comorbidity, and medication exposure). To examine sex-differences in cardiovascular risk factor associations, we will calculate the ratios of hazard ratios and corresponding 95% CIs by fitting an interaction term for sex in fully adjusted cox proportional hazard models with women as the reference group.

To examine the absolute risk of developing CVD, we will calculate incidence rates by dividing the number of incident cases by the number of patient years in the cohort, and stratify analyses by age, sex and absence/presence of risk factors. To estimate the impact of individual risk factors on CVDs (and estimate the reduction in CVD that would occur if risk factor exposure was eliminated), we will calculate the population- attributable fraction (PAF) for the 10-year incidence of cardiovascular disease and each risk factor using the methods from Laaksonen and colleagues.12

To examine the robustness of our results, we will conduct several sensitivity analyses:

1. Stratifying analyses by baseline- and ever-use of cholesterol and/or blood pressure lowering medications;
2. Censoring time at risk to the date of the first-reported cholesterol and/or blood pressure lowering drug prescription (subgroup analysis restricted to the cohort to individuals with functional linkage to primary care records);
3. Excluding events recorded in the first two years following study enrollment (to account for the possibility of reverse causation in observed associations)
4. Recalculating category-specific hazard ratios and confidence intervals for new-onset cardiovascular disease using Fine-Gray sub-distribution hazard models accounting for the competing risk of death from any cause;

We will report study findings according to the Reporting of studies Conducted using Observational Routinely-collected health Data (RECORD) recommendations.13

Preliminary access to the data from UK Biobank was available to the authors at the time this study protocol was established.

**Bibliography**

1 Conrad N, Judge A, Tran J, *et al.* Temporal trends and patterns in heart failure incidence: a population-based study of 4 million individuals. *The Lancet* 2017; **391**: 572–80.

2 Conrad N, Verbeke G, Molenberghs G, *et al.* Autoimmune diseases and cardiovascular risk: a population-based study on 19 autoimmune diseases and 12 cardiovascular diseases in 22 million individuals in the UK. *Lancet (London, England)* 2022; **0**. DOI:10.1016/S0140-6736(22)01349-6.

3 Conrad N, Molenberghs G, Verbeke G, *et al.* Trends in cardiovascular disease incidence among 22 million people in the UK over 20 years: population based study. *BMJ* 2024; **385**: e078523.

4 Visseren FLJ, Mach F, Smulders YM, *et al.* 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. *European Heart Journal* 2021; **42**: 3227–337.

5 Khan SS, Matsushita K, Sang Y, *et al.* Development and Validation of the American Heart Association’s PREVENT Equations. *Circulation* 2024; **149**: 430–49.

6 SCORE2 working group and ESC Cardiovascular risk collaboration. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *European Heart Journal* 2021; **42**: 2439–54.

7 Ridker PM, Moorthy MV, Cook NR, Rifai N, Lee I-M, Buring JE. Inflammation, Cholesterol, Lipoprotein(a), and 30-Year Cardiovascular Outcomes in Women. *N Engl J Med* 2024; : NEJMoa2405182.

8 Savarese G, von Haehling S, Butler J, Cleland JGF, Ponikowski P, Anker SD. Iron deficiency and cardiovascular disease. *European Heart Journal* 2023; **44**: 14–27.

9 Inker LA, Eneanya ND, Coresh J, *et al.* New Creatinine- and Cystatin C–Based Equations to Estimate GFR without Race. *New England Journal of Medicine* 2021; **385**: 1737–49.

10 Delgado C, Baweja M, Crews DC, *et al.* A Unifying Approach for GFR Estimation: Recommendations of the NKF-ASN Task Force on Reassessing the Inclusion of Race in Diagnosing Kidney Disease. *American Journal of Kidney Diseases* 2022; **79**: 268-288.e1.

11 White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* 2011; **30**: 377–99.

12 Laaksonen MA, Virtala E, Knekt P, Oja H, Härkänen T. *SAS* Macros for Calculation of Population Attributable Fraction in a Cohort Study Design. *J Stat Soft* 2011; **43**. DOI:10.18637/jss.v043.i07.

13 Benchimol EI, Smeeth L, Guttmann A, *et al.* The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLOS Medicine* 2015; **12**: e1001885.