Cardiovascular risk in autoimmune diseases: associations and interactions with traditional risk factors in a prospective population-based cohort of 450,000 individuals

**Statistical analysis plan**

Art Schuermans, Mark Woodward, Nathalie Conrad; September 17th, 2025.

**Aims**

This study aims to perform a comprehensive analysis of cardiovascular outcomes and risk-factor associations in autoimmune diseases.

**Methods**

**Study design and participants**

We will use data from the UK Biobank. The UK Biobank is a population-based study including roughly 500,000 volunteers aged 40 to 69 years at baseline.1 Participants were recruited from 22 assessment centres across the United Kingdom between 2006 and 2010. At enrolment, participants provided informed consent; underwent routine physical examination; provided details on social and demographic details, lifestyle factors, medical history, and medication use; and donated blood samples. Ascertainment of incident events occurred via linkage to electronic health records through October 2022.

Among the full UK Biobank cohort, we will define an autoimmune disease cohort consisting of individuals with an autoimmune disease (any of the 19 conditions investigated) and free of cardiovascular disease at enrolment. UK Biobank participants free of autoimmune diseases at enrolment will constitute the reference group. Similar to the autoimmune disease cohort, individuals with cardiovascular disease at enrolment will be excluded.

**Autoimmune diseases**

We will investigate 19 of the most common autoimmune diseases.2 These include: Addison’s disease; ankylosing spondylitis; Coeliac; Graves’ disease; Hashimoto thyroiditis; inflammatory bowel disease; type 1 diabetes; multiple sclerosis; myasthenia gravis; psoriasis; polymyalgia rheumatica; primary biliary cholangitis; pernicious anaemia; rheumatoid arthritis; Sjögren’s; systemic lupus erythematosus; systemic sclerosis; vasculitis; and vitiligo. Diseases will be considered individually and as a composite including all 19 autoimmune diseases combined. Individuals with autoimmune diseases will be identified based on self-reported baseline diagnoses and corresponding ICD-10 (*International Classification of Diseases, Tenth Revision*) codes from linked electronic health records.

**Cardiovascular diseases**

To best characterize the spectrum of cardiovascular diseases, we will examine the following twelve conditions as outcomes: aortic aneurysm; atrial fibrillation/flutter; conduction system disease; heart failure; infective endocarditis; ischemic heart disease; myocarditis and pericarditis; peripheral arterial disease; stroke; supraventricular arrhythmias; valve disorders; and venous thromboembolism.2 Diseases will be considered individually and as a composite outcome of all twelve cardiovascular diseases combined.

Incident outcomes will be identified using qualifying ICD-10 and OPCS-4 (*Office of Population Censuses and Surveys' Classification of Surgical Operations, Version 4*) codes from linked secondary care admissions data and death certificates, in any diagnostic position.

In addition to linked diagnosis and procedure codes, we will also use self-reported data from the baseline study visit to identify and exclude individuals with prevalent cardiovascular diseases at enrolment.

**Risk factors**

In the UK Biobank, demographic characteristics, health behaviours, medical history, and medication use were collected through self-report during a standardized baseline assessment.1 Physical measurements were taken and blood and urine samples were obtained at baseline by trained study staff. Conventional cardiovascular risk factors were chosen based on their use in the most recent cardiovascular risk scores endorsed by *European Society of Cardiology* and *American Heart Association* or recommended for screening in the latest cardiovascular disease prevention guidelines. We complemented this list with four additional markers: lipoprotein(a) and high-sensitivity C-reactive protein (hsCRP).

Therefore, the following variables were selected to represent the range of well-established cardiovascular risk factors: age, sex, smoking status, BMI, blood pressure, type 2 diabetes, non-HDL cholesterol, socioeconomic status, lipoprotein(a), hsCRP, eGFR, and UACR.

Age and sex will be ascertained from self-report at enrolment. Smoking status will be derived from self-reported history. BMI will be calculated from baseline height and weight measurements. Blood pressure was recorded twice at one-minute intervals after five minutes of seated rest using an *Omron 705 IT* electronic monitor (*OMRON Healthcare Co., Ltd.*; Kyoto, Japan); the mean of the first and the second automated readings will be used for data analysis when available. Type 2 diabetes at baseline will be defined as either self-reported history of type 2 diabetes, a diagnosis of type 2 diabetes in secondary care, or a measurement of glycated haemoglobin (HbA1c) ≥48 mmol/mol (6.5%) at first assessment.

Socioeconomic status will be quantified using the Townsend deprivation index, which incorporates local area-based data on employment, car ownership, home ownership, and household overcrowding.3 Non-high-density lipoprotein (HDL) cholesterol will be calculated by subtracting HDL cholesterol from total cholesterol, both measured in baseline blood samples using *AU5800* analysers (*Beckman Coulter [UK] Ltd.*; High Wycombe, UK). Serum lipoprotein(a), hsCRP, and creatinine were also quantified at baseline using *AU5800* analysers (*Beckman Coulter [UK] Ltd.*; High Wycombe, UK). Urine creatinine was quantified using an *AU5400* analyser (*Beckman Coulter [UK] Ltd.*; High Wycombe, UK). Urine microalbumin was quantified using immune-turbidimetric analysers from *Randox Bioscience* (Crumlin, UK). Estimated glomerular filtration rate (eGFR) will be calculated using the 2021 CKD-EPI creatinine-cystatin C equation, incorporating serum creatinine, cystatin C, sex, and age.4 Urine albumin creatinine ratio (UACR) will be calculated by dividing urine microalbumin by urine creatinine.

Biomarker measurements will be categorized into fifths based on sex-specific study cohort quintiles, with the highest fifth serving as the reference group for eGFR, and the lowest fifth serving as the reference group for all other biomarkers. Socioeconomic status will be categorised into fifths from the UK population census data, with the highest fifth serving as the reference group (i.e., the fifth with the lowest Townsend deprivation index values).

**Cardiovascular preventive therapies**

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

**Statistical analysis**

Biomarker measurements outside the reportable range will be set to fixed values, either half the lower limit of detection (for values below the reportable range) or the upper limit of detection (for values above the reportable range). For biomarker measurements missing due to other reasons and other exposure variables with missing data, we will impute missing values using multiple imputation by chained equations, with ten imputed datasets, 50 iterations, and predictive mean matching for continuous variables and proportional odds models for ordered categorical variables (smoking status). Imputation models will use all risk factor exposures and cardiovascular outcome variables, as well as self-reported medication use, as predictors. Estimates and standard errors will be obtained using Rubin’s rules to combine the results of the separate analyses of individual imputed datasets.

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

Time at risk will be calculated separately for every outcome of interest, and will start on participants’ baseline date, and stop at the earliest of death, incidence of the disease of interest, or follow-up end date. To visualize differences in cardiovascular event rates over time, we will generate cumulative incidence curves using the Kaplan-Meier method.

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

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 models adjusted for all risk factors examined in this study and baseline use of cholesterol lowering and blood pressure lowering medications. Continuous variables with skewed distributions will be log-transformed prior to analysis.

We will further explore the shapes of associations between continuous risk factors (age, BMI, systolic blood pressure, non-HDL cholesterol, socioeconomic status, lipoprotein[a], hsCRP, eGFR, and UACR) and incident cardiovascular outcomes using restricted cubic spline models, stratified by autoimmune disease status.

To assess whether associations between the risk factors and cardiovascular outcomes are modified by autoimmune disease status, we will calculate ratios of HRs and corresponding 95% CIs by fitting an interaction term for autoimmune disease status in fully adjusted Cox proportional hazard models.5

To evaluate whether any of the risk factors investigated have differential effects in sociodemographic subgroups, we will perform Cox proportional hazards models stratified by age, sex and Townsend deprivation index.

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

1. To examine whether observed associations could be biased by increased use of preventive therapies in high-risk individuals, we will perform (i) stratified analyses by baseline-use of lipid and/or blood pressure lowering medications; and (ii) analyses censoring time at risk to the date of the first-reported lipid and/or blood pressure lowering drug prescription (subgroup analysis restricted to the cohort to individuals with functional linkage to primary care records);
2. To test whether observed associations could be due to increased medical attention leading to higher rates of diagnoses in individuals regularly followed-up for a chronic condition, we will perform sensitivity analyses restricting cardiovascular diagnoses recorded as the primary reason for hospital admission or as the primary cause of death.
3. To examine the impact of missing data management techniques, we will perform analyses restricted to individuals with no missing data on risk factor exposure variables (complete case analyses);
4. To examine the potential impact of the competing risk of death, we will calculate 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.

All statistical analyses will be performed in R (version 4.3.3).

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

We uploaded this analysis plan on GitHub on September 17th, 2025.

**References**

1. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209. doi:10.1038/s41586-018-0579-z

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. *The Lancet*. 2022;400(10354):733-743. doi:10.1016/S0140-6736(22)01349-6

3. Jordan H, Roderick P, Martin D. The Index of Multiple Deprivation 2000 and accessibility effects on health. *J Epidemiol Community Health (1978)*. 2004;58(3):250-257. doi:10.1136/JECH.2003.013011

4. Inker LA, Eneanya ND, Coresh J, et al. New Creatinine- and Cystatin C–Based Equations to Estimate GFR without Race. *N Engl J Med*. 2021;385(19):1737-1749. doi:10.1056/NEJMOA2102953/SUPPL\_FILE/NEJMOA2102953\_DISCLOSURES.PDF

5. Woodward M. Rationale and tutorial for analysing and reporting sex differences in cardiovascular associations. *Heart*. 2019;105(22):1701-1708. doi:10.1136/HEARTJNL-2019-315299