BMJ Open

Cohort profile: Copenhagen Hospital Biobank - Cardiovascular Disease Cohort (CHB-CVDC): Construction of a large-scale genetic cohort to facilitate a better understanding of heart diseases

Ina H Laursen ,[1] Karina Banasik ,[2] Amalie D Haue ,[2] Oscar Petersen,[1] Peter C Holm,[2] David Westergaard,[2] Henning Bundgaard,[3,4] Søren Brunak,[2] Ruth Frikke-Schmidt,[4,5] Hilma Holm,[6] Erik Sørensen,[1] Lise W Thørner,[1] Margit A H Larsen,[1] Michael Schwinn,[1] Lars Køber,[3,4] Christian Torp-Pedersen,[7] Sisse R Ostrowski,[1,4] Christian Erikstrup,[8] Mette Nyegaard,[9] Hreinn Stefánsson,[6] Arnaldur Gylfason,[6] Florian Zink,[6] G Bragi Walters ,[6,10] Asmundur Oddsson,[6] Guðmar Þorleifsson,[6] Gisli Másson,[6] Unnur Thorsteinsdottir,[6,10] Daniel Gudbjartsson,[6,11] Ole B Pedersen,[12] Kári Stefánsson,[6,10] Henrik Ullum[1,4]

To cite: Laursen IH, Banasik K, Haue AD, *et al*. Cohort profile: Copenhagen Hospital Biobank - Cardiovascular Disease Cohort (CHB-CVDC): Construction of a large-scale genetic cohort to facilitate a better understanding of heart diseases. *BMJ Open* 2021;**11**:e049709. doi:10.1136/bmjopen-2021-049709

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2021-049709).

Received 04 February 2021
Accepted 03 December 2021

For numbered affiliations see end of article.

Correspondence to
Dr Karina Banasik;
karina.banasik@cpr.ku.dk

ABSTRACT

Purpose The aim of Copenhagen Hospital Biobank-Cardiovascular Disease Cohort (CHB-CVDC) is to establish a cohort that can accelerate our understanding of CVD initiation and progression by jointly studying genetics, diagnoses, treatments and risk factors.

Participants The CHB-CVDC is a large genomic cohort of patients with CVD. CHB-CVDC currently includes 96 308 patients. The cohort is part of CHB initiated in 2009 in the Capital Region of Denmark. CHB is continuously growing with ~40 000 samples/year. Patients in CHB were included in CHB-CVDC if they were above 18 years of age and assigned at least one cardiovascular diagnosis. Additionally, up-to 110 000 blood donors can be analysed jointly with CHB-CVDC. Linkage with the Danish National Health Registries, Electronic Patient Records, and Clinical Quality Databases allow up-to 41 years of medical history. All individuals are genotyped using the Infinium Global Screening Array from Illumina and imputed using a reference panel consisting of whole-genome sequence data from 8429 Danes along with 7146 samples from North-Western Europe. Currently, 39 539 of the patients are deceased.

Findings to date Here, we demonstrate the utility of the cohort by showing concordant effects between known variants and selected CVDs, that is, >93% concordance for coronary artery disease, atrial fibrillation, heart failure and cholesterol measurements and 85% concordance for hypertension. Furthermore, we evaluated multiple study designs and the validity of using Danish blood donors as part of CHB-CVDC. Lastly, CHB-CVDC has already made major contributions to studies of sick sinus syndrome and the role of phytosterols in development of atherosclerosis.

Future plans In addition to genetics, electronic patient records, national socioeconomic and health registries extensively characterise each patient in CHB-CVDC and provides a promising framework for improved understanding of risk and protective variants. We aim to include other measurable biomarkers for example, proteins in CHB-CVDC making it a platform for multiomics cardiovascular studies.

Strengths and limitations of this study

- ► Genetic data on 96 308 patients allows for extensive studies of both common and rare cardiovascular diseases.
- ► Extensive information of medical history allows for stratification on a high level of granularity, which can be combined with sociodemographic and other environmental factors.
- ► There is a potential bias caused by the selection of healthy blood donors from the Danish Blood Donor Study as controls in case–control studies.

INTRODUCTION

Cardiovascular diseases (CVDs) include diseases of the heart and the major blood vessels and range from asymptomatic disease entities to myocardial infarction, heart failure (HF), sudden cardiac death and stroke. Modifiable risk factors, such as obesity, smoking, hypertension and dyslipidaemia, are shared across several CVDs. Genetic risk factors are also shared across several CVDs, for example, the earliest and most robust genetic marker for coronary artery disease (CAD), the chromosome 9p21 locus, is also associated with stroke, aneurysms and

myocardial infarction.[1–4] Genetic studies of CAD started with linkage studies that identified monogenic causes of CAD and small candidate gene studies with dubious findings. The field has since evolved with increasing cohorts and large, international consortia such as CARDIoGRAM and CAD C4D Genetics Consortium being established.[5–8] These initiatives combined with individual efforts across multiple ancestries have led to the discovery of 171 genome-wide significant CAD risk variants,[9] 12 HF variants[10] and at least 138 atrial fibrillation (AF) loci,[11] which have been replicated in independent populations. It is expected that there are many more risk variants to be discovered. Recently, biobanks such as the UK Biobank, Japan Biobank, FinnGen and the Trøndelag Health Study (The HUNT Study)[12–15] have contributed to the ethnic diversity within genetic research and increased the number of both participants and phenotypes for large-scale genetic studies including studies within CVDs. Not only have numbers increased, but the detailed information captured on the individual enables a deeper phenotyping of the biobank participants with great potential to pave the way for precision cardiology. Improving our understanding of the complex interactions between genetics, lifestyle factors and individual CVDs necessitates large cohorts, such as Copenhagen Hospital Biobank (CHB)-Cardiovascular Disease Cohort (CVDC), enabling stratification of individuals into more refined and detailed phenotypes.

Here, we present the CHB-CVDC, a potent resource for genetic research within CVDs. The CHB-CVDC is part of the CHB, initiated in 2009 and funded by the Department of Clinical Immunology, Copenhagen University Hospital, Denmark. The CHB takes advantage of the pre-existing blood banking system that in addition to treatment obligations (eg, blood transfusions) aims to support medical research. The data foundation of CHB is a collection of leftover EDTA blood samples from patients hospitalised in the Capital Region of Denmark, who were subject to blood typing or red blood cell antibody screening. Details about data collection, quality assessment and storage can be found in Sørensen *et al*.[16] Presently, CHB contains samples from more than 450 000 individuals.[16] Each patient is identified by a Central Person Registry (CPR) number, which facilitates linkage to nationwide registries and electronic patient records. Nationwide registries and electronic patient records are updated continuously. Patients in CHB are over 18 years old and only included once.[16]

The aim of the CHB-CVDC is to provide a platform for studying genetic, environmental, medical and other factors to further our knowledge of initiation, progression and manifestation of CVDs. CHB-CVDC facilitate studies of individual diseases as well as the shared risk factors, pathophysiology and disease trajectories. The ability to link CHB samples to individual data available from local hospital databases, the Danish National Health Registries, Electronic Patient Records and the Clinical Quality Databases in Denmark facilitates fine-grained stratification of patients into subpopulations. This information can be used to develop for example, risk prediction models for CAD and other CVDs.

The objectives of the present study are to present the features of CHB-CVDC and to evaluate the use of the cohort as a resource for genetic studies by replicating established genetic variants associated with CAD, AF, HF, essential hypertension and cholesterol levels in European populations. Furthermore, we also investigate the potential bias of including the Danish blood donors in the cohort.

COHORT DESCRIPTION

Population characteristics

The CHB is a hospital-driven biobank with a broad collection scheme covering a wide range of diseases, designed to facilitate research in health and disease by enabling researchers access to a large resource of well-defined patient samples.[16] The CHB-CVDC inclusion criteria contain two components: individuals have (1) to be included in CHB and (2) been assigned at least one of the hospital admission codes presented in table 1. For a more detailed overview of population characteristics, see online supplemental table 1. Only the first assigned CVD was counted. Currently, the CHB-CVDC comprises 96 308 individuals (55% males), and the cohort is increasing as patients are being included on a continuous basis.

Females were older when diagnosed with their first CVD (63.2 years, 95% CI 63.1 to 63.4) compared with males (59.8 years, 95% CI 59.7 to 60.0).

We evaluated the prevalence of comorbidities in the CHB-CVDC using the Charlson Comorbidity Index.[17 18] See online supplemental material for a full description. We found that 34 375 individuals (53% males) have at least one comorbidity, and 23 656 individuals (61% males) have more than five comorbidities (online supplemental table 2). Online supplemental figure 1 gives an overview of the number of comorbidities in different age groups. Consequently, the cohort is well powered to study CVDs in the context of comorbidities. An overview of diagnosis assigned prior to the first CVD is presented in online supplemental table 3. The co-occurrence of CVDs within the cohort is also pronounced, see online supplemental table 4 for an overview.

The most prevalent disease was hypertension, which affected 64 455 CHB-CVDC patients. Cross-referencing with the Danish National Prescription Registry, 63 431 patients had redeemed one or more prescriptions for antihypertensive medication. In addition, 26 581 have been diagnosed with hypercholesterolaemia (online supplemental table 5).

In studies of binary disease traits there is also the opportunity to use blood donors from the Danish Blood Donor Study (DBDS), adding 110 000 individuals to the total cohort population (n=206 308). The DBDS is a large prospective study of blood donors recruited from the blood bank infrastructure across Denmark.[19]

Table 1 Cohort characteristics

	Women	Men	Total
No of patients in CHB-CVDC (%)	43 479 (45)	52 829 (55)	96 308 (100)
Year of birth, mean (SD)	1942.3 (14.8)	1945.3 (13.0)	
Age at first cardiovascular disease, mean (SD)	63.2 (15.6)	59.8 (13.6)	
Age at inclusion, mean (SD)	70.5 (14.9)	67.4 (13.0)	
Cardiovascular inclusion ICD-10 codes from the National Patient Registry*			
Hypertension and hypertensive cardiac diseases ICD-10: I10-15	16 229	13 317	29 546
Coronary artery diseases and atherosclerosis ICD-10: I20-25, I70	8823	15 972	24 795
Lipid disorders ICD-10: E78	2024	2022	4046
Cardiac arrhythmia ICD-10: I44-49	7177	9031	16 208
Heart failure, cardiac valve disorders, and myocardial diseases ICD-10: I50, I34-39, I05-09, I40-44	3001	4263	7264
Vascular disorders and aneurysms ICD-10: I71-79	1238	1883	3121
Cerebrovascular diseases and cerebral haemorrhage ICD-10: I60-69	3691	4710	8401
Pulmonary heart diseases and diseases of the pulmonary circulation ICD-10: I26-28	751	770	1521
Vascular kidney disease ICD-10: N17-19	545	861	1406

*Patients are stratified by their first assigned cardiovascular diagnosis.
CHB-CVDC, Copenhagen Hospital Biobank-Cardiovascular Disease Cohort; ICD-10, International Statistical Classification of Diseases and Related Health Problems 10th Revision.

Sociodemographic details of the Danish blood donors, including sex and age distributions, education, labour market affiliation and level of urbanisation can be found in Burgdorf *et al*.[19] As part of the DBDS, the DBDS Genomic Cohort was later established with the aim to identify genetic predictors important for the healthy donor phenotype.[20] Genotyping and imputation of samples in DBDS were performed in the same way as for samples in CHB.[20]

Follow-up

The 96 308 genotyped individuals currently in the CHB-CVDC are being followed up through Danish national registries, that contain detailed longitudinal information on every contact with the Danish primary and secondary care service. We will on a regular basis update the data collection by extracting information from the many registries to augment the registry-based phenotyping of the cohort. The linked registries are constantly being updated and to date the cohort contains 56 769 patients that are still alive and 39 539 that have died as per information from the Danish Registry of Causes of Death.[21] The median time from inclusion (after 2009) to end of follow-up (before 2020) or death is 3.9 years (IQR 1.70–5.73) (online supplemental figure 2). However, because phenotypic data are also included prior to inclusion, the median follow-up time is 41 years. This correspond to the majority of the patients have been followed since the inception of the Danish National Patient Registry in 1977.

The most common underlying cause of death is cancer (36%) followed by heart disease (17%) and respiratory disease (10%). However, only 2.6% (n=969) of those that died from a medical reason has undergone autopsy and the exact cause is therefore subject to considerable uncertainty.[21 22] Among those who have undergone autopsy a CVD is the most common cause of death (52%). Online supplemental table 6 presents a summary of causes of death for all patients who have died in CHB-CVDC.

Patient and public involvement

Patients and public were not involved in the design of this study.

Genetics

As part of a collaboration between the Copenhagen University Hospital, Rigshospitalet, Denmark and deCODE genetics, Iceland, leftover EDTA blood samples from the patients in CHB-CVDC are sent to deCODE genetics for DNA extraction, genotyping and subsequent imputation. SNP genotyping is performed on the Infinium Global Screening Array from Illumina. Approximately 660 000 common variants are genotyped. A reference panel backbone consisting of whole-genome sequence data from 8429 Danes along with 7146 samples from North-Western Europe from participants in various research projects at deCODE genetics is used for imputation.[23–25] The genotyped samples were long-ranged phased using Eagle2 together with 171 298 genotyped samples from Northwestern Europe. The process used to whole-genome sequence the reference panel backbone, and the subsequent imputation has been extensively described previously.[23–25] Genotyping and imputation procedures are identical to the general procedures applied in DBDS and are described in detail previously.[20]

Databases

Denmark has one of the world's most comprehensive population registry systems, integrated via the social security number (CPR number) that was established in 1968.[26] Nationwide individual-level data that is linkable via the CPR number includes birthdate, place of residence, emigration, immigration, family relationships, education, labour market affiliation, cause and date of death, and much more enabling a deep phenotyping of patients in CHB-CVDC.[26–30] The Danish National Patient Registry contains information of all hospitalisations, since 1977. This registry includes date of hospitalisation, diagnoses related to the hospitalisation, hospital examinations and procedures.[31] Diagnoses in this registry are classified according to the international classification of diseases, V.8 (1977–1993) and version 10 (1994-).[32] Procedural codes are classified using a country-specific coding until 1995, and then according to the Nordic Medico-Statistical Committee Classification of Surgical Procedures.[33]

Furthermore, electronic patient records that contain clinical notes, laboratory results, images and treatments are available. From the clinical notes we are in the process of extracting smoking, alcohol intake, height, weight and blood pressure measurements using a text mining approach. The electronic health records cover all hospitals in the Capital Region and Region Zealand in the period 2006–2016, and 87% of patients in CHB-CVDC data related to at least one hospital admission.

The laboratory database contains laboratory results from the Departments of Clinical Biochemistry and Clinical Immunology laboratories in Denmark. This database includes results from inpatient, outpatient and emergency encounters. The database includes data dating back to 2008. We are working on making the data from different laboratories comparable.

The Danish National Prescription Registry contains data on all prescription drugs dispensed in Denmark since 1995 and it is mandatory by law for all pharmacies to report to this registry.[34] Consequently, it is possible to trace medication use and compliance, assess associations between medications and paraclinical outcomes and through well-established drug-associated phenotypes to investigate diseases that do not require a hospitalisation, such as hypertension and diabetes.[34] In combination with the electronic patient records that cover drugs administered during hospital admissions this database is a valuable source for pharmacogenomic studies, from where we can extract information about drug dosage prescribed/administered, dosage changes and adherence over time, adverse reactions and polypharmacy.[35]

The Danish healthcare system has developed several National Clinical Registries such as the Danish Heart Registry, the AF Database, the Familial Hypercholesterolaemia Database and the Danish Heart Failure Database.[36 37] These databases contribute information such as body mass index, smoking, alcohol consumption, diabetes history, previous cardiac surgery, follow-up and outcome data and other related variables.[36 37] An extensive overview of registries and databases and their content are given in online supplemental table 7.

Findings to date

To validate CHB-CVDC as a resource for genetic research, we set out to replicate genetic variants associated with CAD, AF, HF, essential hypertension and cholesterol levels identified in prior large meta-analyses of multiple populations.

For each phenotype, we selected a reference study based on the date of publication, the number of individuals in the genetic analysis and the genetic ancestry of the population (preferably European) (table 2). Independent significant genetic variants were retrieved from the reference study and compared with association results from CHB-CVDC. Generalised linear mixed models (Scalable and Accurate Implementation of GEneralized mixed model, SAIGE) and linear mixed models (Bayesian mixed model association method, BOLT-LMM) were applied to obtain association results for the included genetic variants.[38 39] For CAD, AF, HF and hypertension, we use cases and controls from the cohort containing both CHB-CVDC and DBDS. Additionally, we compared how the choice of control group impacts the findings, (1) using only patients from CHB-CVDC as cases or controls (2) using only cases from CHB-CVDC and only controls from DBDS. Further details on the methods are available in online supplemental material.

CAD, AF and HF

Cases and controls for CAD, AF and HF were defined by International Statistical Classification of Diseases and Related Health Problems 10th/8th Revision (ICD-10/ICD-8) codes (table 2). A positive correlation was found between the observed and published effect sizes for all these phenotypes (figure 1A–C). We observed that we, in general, had smaller effect sizes compared with other studies (see online supplemental table 8). The proportion of replicated variants compared with the number we had power to replicate were 66% for CAD, 88% for AF and 90% for HF (table 2).

Cholesterol measurements

A patient's first measured level of High-Density Lipoprotein (HDL), Low-Density Lipoprotein (LDL), Total Cholesterol (TC) and Triglyceride (TG) were extracted based on NPU codes from the Laboratory Database in Denmark. Detailed information on methods is found in online supplemental material. More than 80 000 corresponding to more than 83% of the CHB-CVDC had at least one cholesterol measurement. We observed that the effect sizes for TG were largely concordant compared with the prior study (figure 1G and online supplemental table 8). For HDL, LDL and TC the effect sizes in this study are lower compared with the effect sizes in the reference study (online supplemental table 8 and figure 1D–F). The proportion of replicated variants compared with the number we had power to replicate exceeds 85% for all

Table 2 Overview of reference studies, number of cases and controls and number of variants investigated

Phenotype	Reference study (reference)	Cases/controls in reference studies	Replication in current study Cases	Controls	No of variants with concordant direction of effect	Replicated/total	Replicated/ power to replicate
Coronary artery disease	Van der Harst 2018[46]	122 733/424 528	33 746	154 311	236/241 (98%)	90/241 (37%)	90/137 (66%)
Atrial fibrillation	Nielsen 2018[11]	60 620/970 216	30 229	157 669	137/140 (98%)	96/140 (69%)	96/109 (88%)
Heart failure	Shah 2020/Arvanitis 2020[10 49]	47 309/930 014 10 976/437 573	21 443	167 068	14/15 (93%)	9/15 (60%)	9/10 (90%)
High density lipoprotein	Global Lipids Genetic Consortium 2013[50]	188 577/*	85 435	*	67/68 (99%)	55/68 (81%)	55/60 (92%)
Low density lipoprotein	Global Lipids Genetic Consortium 2013[50]	188 577/*	81 435	*	55/57 (96%)	35/57 (61%)	35/41 (85%)
Total cholesterol	Global Lipids Genetic Consortium 2013[50]	188 577/*	86 297	*	71/72 (99%)	44/72 (61%)	44/52 (85%)
Triglycerides	Global Lipids Genetic Consortium 2013[50]	188 577/*	83 087	*	39/40 (98%)	29/40 (73%)	29/32 (91%)
Hypertension versus SBP	Evangelou 2018[40]	1 006 863/*	63 431	87 752	220/258 (85%)	39/258 (15%)	39/59 (66%)
Hypertension versus DBP	Evangelou 2018[40]	1 006 863/*	63 431	87 752	260/307 (85%)	33/307 (11%)	33/80 (41%)

If the variant had the same direction of effect and p<0.05 (Bonferroni adjusted), we considered it replicated.
*Case/control setup not applicable. Instead, the total number of samples are listed under cases. A variant was replicated if the effect size of the risk allele had the same direction of effect and a p<0.05 (Bonferroni adjusted). The power to replicate was estimated from the SE and the effect size. The power was set at 80%.
DBP, diastolic blood pressure; SBP, systolic blood pressure.

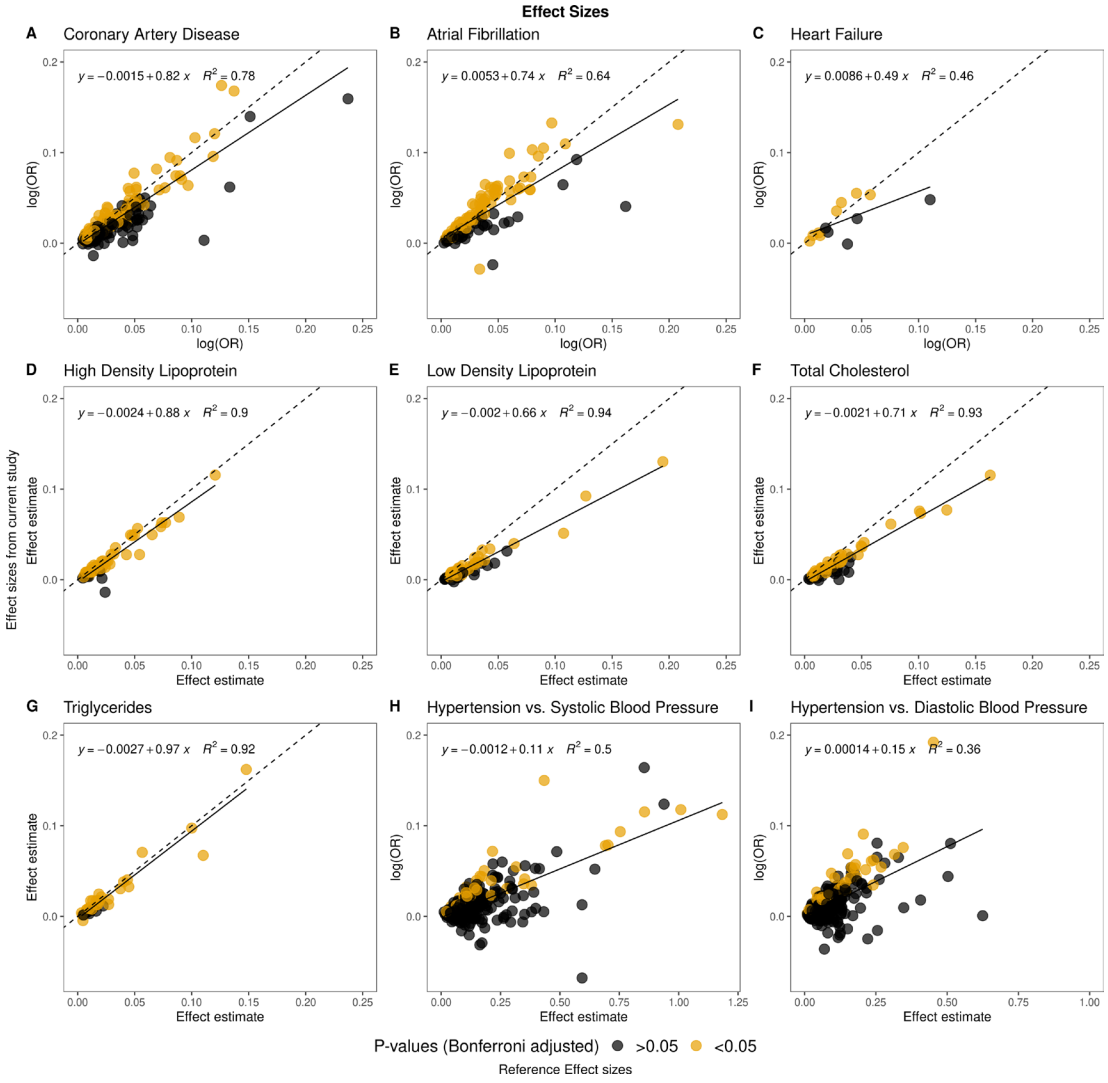


Figure 1 Comparison of effect sizes and reference effect sizes. The effect sizes weighted by risk allele frequency of the reference study (X-axis) are compared with the effect sizes weighted by the risk allele frequency from this study (y-axis). The dotted lines correspond to a correlation of 1 and the dense lines to the observed trendlines.

the cholesterol analyses. The HDL association analysis where the analysis with most replicated variants (81%).

Essential hypertension

Hypertension cases were identified using ICD-10/ICD-8 codes and prescriptions of antihypertensive drugs. This is in contrast to the reference study, where systolic and diastolic blood pressure were measured directly.[40] The observed effect sizes and the effect sizes in the reference study were positively correlated, although a smaller proportion of variants had a concordant direction of effect (85%) compared with the phenotypes described above (>90%) (figure 1H,I and table 2). The number of variants we had power to replicate were low (59/258 and 80/307) and the proportion of replicated variants compared with the number of variants we had power to replicate were also low (<67%).

Additional comparisons with the reference studies

All variants had similar allele frequencies as the frequencies in the reference studies (online supplemental figure 3). Overall, association p values were lower in the reference studies, likely due to greater sample sizes (online supplemental figure 4). The variants that showed discordant direction of effect had small effect sizes and large SEs.

We were able to replicate >85% of the variants for six out of nine phenotypes, given the estimated power (power=0.8, α=0.05) (table 2).

Comparison of different study designs

We repeated the association analyses for CAD, AF and HF with only patients from CHB-CVDC as controls and only participants from DBDS as controls, respectively. For all the phenotypes the control group had no effect on the number of variants with concordant direction of effect, but the number of replicated variants decreased when only DBDS was used as controls (online supplemental table 9 and online supplemental figures 5–8). Furthermore, we employed LD Score regression to investigate residual confounding (online supplemental table

10).[41] Residual confounding was evaluated through the LD Score regression intercept and the attenuation ratio (the ratio between the intercept and the mean χ^2 statistic). An intercept close to one indicates no additional confounding, and a small ratio indicates that the genomic inflation arises from the polygenicity of the phenotype and not residual confounding. Lastly, we also evaluated the genetic correlations to prior studies using LD Score Regression (online supplemental table 11). We used the 1000Genomes EUR v3 LD reference panel for both analyses. Using only DBDS as controls yielded a high level of residual confounding. The genetic correlation analysis indicated the setup using only CHB-CVDC was not significantly different from the main analysis. We could not calculate a genetic correlation for the last study design as the observed heritability was negative.

Other contributions

In a recent study, samples from CHB-CVDC were used together with Icelandic samples and samples from UK biobank, to examine the effects of ABCG5/8 variants on dietary cholesterols and phytosterols and the risk of CAD.[42] The authors conclude that phytosterols may be involved in the development of atherosclerosis and clinical trials are needed to investigate whether phytosterols increase risk of CVD.[42 43]

A genome-wide association study of 6 469 sick sinus syndrome cases and more than a million controls revealed six loci associated to the disease.[44] Mendelian randomisation suggests a direct role of AF and lower heart rate in the development of sick sinus syndrome.[44]

Angioedema is a known adverse drug reaction in individuals treated with antihypertensive ACE inhibitors. A Danish study has found common variants located close to the bradykinin receptor B_2 gene to be associated with increased risk of developing angioedema related to treatment with ACE inhibitors.[45]

DISCUSSION

In this study, we present the features of CHB-CVDC and use the genetic data to validate the registry-based phenotyping in CHB-CVDC. We show a high proportion of variants with a concordant direction of effect between the respective reference studies and the current study. For CAD, AF, HF and cholesterol measurements the proportion exceeded 93%. For the diagnosis of hypertension compared with blood pressure measurements the proportion of variants with the same direction was lower but still 85%.

The proportion of replicated variants exceeded 60% for AF, HF and cholesterol measurements compared with the number we had power to replicate. For CAD and hypertension, this proportion were <67%. The lower proportion could be a consequence of different phenotyping, analysis methods, population differences and the number of cases compared with the number of samples in the reference studies. Furthermore, a high proportion of patients in CHB-CVDC have a hypertension diagnosis, hence it was almost only possible to compare with DBDS that is a cohort of younger and typically healthier individuals (see text below).

It is very likely that having systematic blood pressure measurements available for the present study would have increased the concordance with the studies we set out to replicate. However, we do show that the proxies for blood pressure used that is, diagnosis codes for hypertension and prescription of antihypertensive drugs capture enough signal to replicate some of the strongest signals with comparable effect sizes. We are currently in the process of extracting blood pressure measurement from the unstructured clinical notes using a text mining approach, and they will be available in future studies.

The analyses with DBDS as controls only, show that this setup is decreasing the number of replicated variants and the residual confounding is high. For AF and HF, it seems that DBDS together with patients from CHB-CVDC as controls and CHB-CVDC as controls can be used interchangeably. Nonetheless, careful considerations of study design are necessary and depends on the phenotype under investigation. Here, we have outlined one way of doing so.

For AF, HF and the cholesterol measurements the proportion of replicated variants compared with the number we had power to replicate were >85%. Overall, these results are promising regarding future collaborations with CHB-CVDC as either a discovery or replication cohort.

The comparison of age at their first cardiovascular diagnosis and age at the blood typing test demonstrates that the majority of the patients already had a cardiovascular diagnosis at time of inclusion in the cohort. Such a study design where the event has occurred before the inclusion could lead to a survivor bias, as individuals who died from diseases before the inclusion are not part of the analyses. Nonetheless, as the cohort increases in size over time, this potential issue will disappear.

Strengths and limitations

As a hospital-driven biobank with a comprehensive collection scheme, the CHB covers a wide range of diseases, thereby facilitating large-scale studies with high clinical importance across many patient groups including the patients with CVDs in CHB-CVDC. The collection of routine blood samples from the clinic is a cost-effective biobanking practice, and as the data used for phenotyping already are available through the registries the time and cost of studies using a biobank platform like CHB-CVDC are less than studies using primary data collection.

With genetic data on 96 308 patients, CHB-CVDC comprises a significant increase in the number of cardiovascular patients with genetic data available.[10 11 40 46] This is a strength in terms of discovery of new genetic and environmental risk factors and interactions as well as a major strength in the cooperation with other cohorts to increase power of future meta analyses.

The potential to link genetic data to individual-level data from the Danish National Registries is a huge advantage in epidemiological and genetic research. Denmark has one of the world's oldest registry systems with information of nationwide hospitalisations from the Danish National Patient Registry coupled with the Danish Civil Registration System.[26 31] Thus, Denmark is, similar to other Nordic countries, a forerunner of the digitalisation of clinical systems for quality and research purposes. Since these registries are nationwide and the healthcare system is financed by state taxes in Denmark, there is a high population-based coverage. The long-term temporal registration of all hospital admissions, procedures and treatments also enables studies of trajectories of the different diseases and their temporal relations.[31]

In the last decades, many clinical quality databases relevant for cardiovascular research has been created, for example, Karbase, a nationwide database of vascular surgery from 1993, the Danish Heart Registry with information of all patients with CAD and with a coronary angiography, a percutaneous coronary intervention or a coronary artery bypass graft from 2004.[36 47] Hence, we will be able to stratify patients at a high level of detail combined with sociodemographic and other environmental factors.

CHB-CVDC is comparable to other major hospital-driven cohorts where patients are not recruited based on a specific disease. However, as patients were subject to blood typing or red blood cell antibody screening the health condition of the patients are perhaps worse than in other hospital-driven cohorts.

A challenge in designing genome-wide association studies in CHB-CVDC is the selection of controls. There is a potential bias caused by the selection of blood donors from the DBDS as controls. Compared with the patients from CHB-CVDC the DBDS participants are much younger.[19] It is inevitable that some of these participants will develop CVDs in the future. Furthermore, the participants from the DBDS are implicitly healthier than the general population and thereby also much healthier than the patients from CHB-CVDC.[19 48] This is exemplified by our finding that using only DBDS as controls the number of replicated variants decrease, and the residual confounding is high. Ways to overcome these challenges must be considered in future studies and reflected in the study designs.

Author affiliations
[1]Department of Clinical Immunology, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark
[2]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark
[3]Department of Cardiology, The Heart Center, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark
[4]Department of Clinical Medicine, University of Copenhagen, Copenhagen, Denmark
[5]Department of Clinical Biochemistry, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark
[6]deCODE genetics, Reykjavik, Iceland
[7]Department of Clinical Investigation and Cardiology, Nordsjællands Hospital, Hillerød, Denmark
[8]Department of Clinical Immunology, Aarhus University Hospital, Aarhus, Denmark
[9]Department of Biomedicine, Aarhus University, Aarhus, Denmark
[10]Faculty of Medicine, University of Iceland, Reykjavik, Iceland
[11]School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland
[12]Department of Clinical Immunology, Zealand University Hospital Køge, Køge, Denmark

Twitter Karina Banasik @Karina_Banasik, Amalie D Haue @fhq750, Søren Brunak @s_brunak and Mette Nyegaard @MetteNyegaard

Acknowledgements We would like to thank the patients in CHB-CVDC and the participants in DBDS. We thank deCODE genetics/Amgen and Department of Clinical Immunology, Copenhagen University Hospital for financial support of this study. Furthermore, we thank the staff working in the biobank facilities and in the hospitals whose work make the existence of this cohort sustainable.

Collaborators We encourage scientific collaborations based on data generated in CHB-CVDC and summary statistics from published analyses will be made available upon request. Any collaboration needs first approval by the CHB-CVDC steering committee.

Contributors IHL, KB, HH and HU conceived and planned the experiments. IHL carried out the analyses. KB, HB, SB, RF-S, HH, ES, LK, CT-P, SRO, CE, MN, HS, UT, DG, OBP, KS and HU contributed to cohort and research design. KB, DW, LT, MAHL, MS, ES, HU and SB established the data infrastructure and data governance design. HB, RF-S, ES, LK, CT-P, OBP, HS, AG, FZ, GBW, AO, GÞ, GM and HU were instrumental to data capture. IHL, KB, ADH, OP, PCH and DW contributed to analyses and interpretation of the results. IHL, KB and HU took the lead in writing the manuscript. HU is the guarantor of this work. All authors provided critical feedback, helped shape the analysis and manuscript, and approved the final version.

Funding This work was supported by the Novo Nordisk Foundation (NNF) (grant numbers: NNF17OC0027594, NNF14CC0001 and NNF18SA0034956), Innovation Fund Denmark (grant number: 5153-00002B) and Nordforsk, Precision Medicine Heart (grant number: 90580 PM Heart).

Competing interests The authors affiliated with deCODE genetics/Amgen are employed by the company.

Patient consent for publication Not applicable.

Ethics approval CHB is classified as a 'biobank for future research'. It is part of the Danish National Biobank and has been approved by the Danish Data Protection Agency (general approval number 2012-58-0004, and local number: RH-2007--30–4129/I-suite 00678). Patients included in CHB were informed about their right to refuse the use of their samples for research via the Danish Tissue Utilisation Registry.[16] Patients from CHB who have been assigned at least one of the CVD inclusion codes (table 1) are included in CHB-CVDC. Studies under CHB-CVDC, including the use of DBDS as controls, are approved by The National Ethical Committee (1708829, 'Genetics of CVD'—a genome-wide association study on repository samples from CHB).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. The study will adhere to the FAIR (http://datafairport.org/: Findable, Accessible, Interoperable and Reusable) concepts. For further access possibilities and contact details please see: https://www.regionh.dk/blodbanken/afdelingen/enheder-paa-rigshospitalet/Sider/biobank.aspx. The data were handled in accordance with 'the Danish Act on Data Protection following the EU Regulation (EU) 2016/ 679, 27 April 2016 pf the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC'. All data are stored in a private secure cloud of the Danish National Supercomputer for Life Sciences-Computerome.

ORCID iDs
Ina H Laursen http://orcid.org/0000-0002-9534-8581
Karina Banasik http://orcid.org/0000-0003-2489-2499
Amalie D Haue http://orcid.org/0000-0001-7656-7976
G Bragi Walters http://orcid.org/0000-0002-5415-6487

REFERENCES

1 Samani NJ, Erdmann J, Hall AS, *et al*. Genomewide association analysis of coronary artery disease. *N Engl J Med* 2007;357:443–53.
2 McPherson R, Pertsemlidis A, Kavaslar N, *et al*. A common allele on chromosome 9 associated with coronary heart disease. *Science* 2007;316:1488–91.
3 Helgadottir A, Thorleifsson G, Manolescu A, *et al*. A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* 2007;316:1491–3.
4 Wellcome Trust Case Control Consortium. Genome-Wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;447:661–78.
5 Preuss M, König IR, Thompson JR, *et al*. Design of the coronary artery disease genome-wide replication and meta-analysis (cardiogram) study: a genome-wide association meta-analysis involving more than 22 000 cases and 60 000 controls. *Circ Cardiovasc Genet* 2010;3:475–83.
6 Schunkert H, König IR, Kathiresan S, *et al*. Large-Scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet* 2011;43:333–8.
7 Coronary Artery Disease (C4D) Genetics Consortium. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nat Genet* 2011;43:339–44.
8 IBC 50K CAD Consortium. Large-Scale gene-centric analysis identifies novel variants for coronary artery disease. *PLoS Genet* 2011;7:e1002260.
9 Roberts R, Chang CC, Hadley T. Genetic risk stratification: a paradigm shift in prevention of coronary artery disease. *JACC Basic Transl Sci* 2021;6:287–304.
10 Shah S, Henry A, Roselli C, *et al*. Genome-Wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nat Commun* 2020;11:163.
11 Nielsen JB, Thorolfsdottir RB, Fritsche LG, *et al*. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat Genet* 2018;50:1234–9.
12 Bycroft C, Freeman C, Petkova D, *et al*. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018;562:203–9.
13 Nagai A, Hirata M, Kamatani Y, *et al*. Overview of the Biobank Japan project: study design and profile. *J Epidemiol* 2017;27:S2–8.
14 FinnGen. FinnGen documentation of R5 release, 2021. Available: https://github.com/FINNGEN/finngen-documentation
15 Krokstad S, Langhammer A, Hveem K, *et al*. Cohort profile: the HUNT study, Norway. *Int J Epidemiol* 2013;42:968–77.
16 Sørensen E, Christiansen L, Wilkowski B, *et al*. Data resource profile: the Copenhagen Hospital Biobank (Chb). *Int J Epidemiol* 2021;50:719–20.
17 Quan H, Sundararajan V, Halfon P, *et al*. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care* 2005;43:1130–9.
18 Gasparini A. Comorbidity: an R package for computing comorbidity scores. *J Open Source Softw* 2018;3:648.
19 Burgdorf KS, Simonsen J, Sundby A, *et al*. Socio-Demographic characteristics of Danish blood donors. *PLoS One* 2017;12:e0169112.
20 Hansen TF, Banasik K, Erikstrup C, *et al*. Dbds genomic cohort, a prospective and comprehensive resource for integrative and temporal analysis of genetic, environmental and lifestyle factors affecting health of blood donors. *BMJ Open* 2019;9:e028401.
21 Helweg-Larsen K. The Danish register of causes of death. *Scand J Public Health* 2011;39:26–9.
22 Svendsen MT, Bøggild H, Skals RK, *et al*. Uncertainty in classification of death from fatal myocardial infarction: a nationwide analysis of regional variation in incidence and diagnostic support. *PLoS One* 2020;15:e0236322.
23 Kong A, Masson G, Frigge ML, *et al*. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* 2008;40:1068–75.
24 Gudbjartsson DF, Helgason H, Gudjonsson SA, *et al*. Large-Scale whole-genome sequencing of the Icelandic population. *Nat Genet* 2015;47:435–44.
25 Helgadottir A, Thorleifsson G, Gretarsdottir S, *et al*. Genome-Wide analysis yields new loci associating with aortic valve stenosis. *Nat Commun* 2018;9:987.
26 Schmidt M, Pedersen L, Sørensen HT. The Danish civil registration system as a tool in epidemiology. *Eur J Epidemiol* 2014;29:541–9.
27 Thygesen LC, Daasnes C, Thaulow I, *et al*. Introduction to Danish (nationwide) registers on health and social issues: structure, access, legislation, and archiving. *Scand J Public Health* 2011;39:12–16.
28 Laugesen K, Ludvigsson JF, Schmidt M, *et al*. Nordic health registry-based research: a review of health care systems and key registries. *Clin Epidemiol* 2021;13:533–54.
29 Jensen VM, Rasmussen AW. Danish education registers. *Scand J Public Health* 2011;39:91–4.
30 Petersson F, Baadsgaard M, Thygesen LC. Danish registers on personal labour market affiliation. *Scand J Public Health* 2011;39:95–8.
31 Schmidt M, Schmidt SAJ, Sandegaard JL, *et al*. The Danish national patient registry: a review of content, data quality, and research potential. *Clin Epidemiol* 2015;7:449–90.
32 WHO. Classification of diseases (ICD). Available: https://www.who.int/standards/classifications/classification-of-diseases [Accessed 5 Sep 2021].
33 Nordisk Medicinal-Statistisk Komité. *NOMESCO classification of surgical procedures*. Copenhagen: Nordic Medico-Statistical Committee, 2010.
34 Kildemoes HW, Sørensen HT, Hallas J. The Danish national prescription registry. *Scand J Public Health* 2011;39:38–41.
35 Rodríguez CL, Kaas-Hansen BS, Eriksson R. Drug interactions in hospital prescriptions in Denmark: prevalence and associations with adverse outcomes 2021.
36 Özcan C, Juel K, Flensted Lassen J, *et al*. The Danish heart registry. *Clin Epidemiol* 2016;8:503–8.
37 Schjødt I, Nakano A, Egstrup K, *et al*. The Danish heart failure registry. *Clin Epidemiol* 2016;8:497–502.
38 Loh P-R, Tucker G, Bulik-Sullivan BK, *et al*. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* 2015;47:284–90.
39 Zhou W, Nielsen JB, Fritsche LG, *et al*. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* 2018;50:1335–41.
40 Evangelou E, Warren HR, Mosen-Ansorena D, *et al*. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat Genet* 2018;50:1412–25.
41 Bulik-Sullivan BK, Loh P-R, Finucane HK, *et al*. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015;47:291–5.
42 Helgadottir A, Thorleifsson G, Alexandersson KF, *et al*. Genetic variability in the absorption of dietary sterols affects the risk of coronary artery disease. *Eur Heart J* 2020;41:2618–28.
43 Weingärtner O, Patel SB, Lütjohann D. It's time to personalize and optimize lipid-lowering therapy. *Eur Heart J* 2020;41:2629–31.
44 Thorolfsdottir RB, Sveinbjornsson G, Aegisdottir HM, *et al*. Genetic insight into sick sinus syndrome. *Eur Heart J* 2021;42:1959–71.
45 Ghouse J, Ahlberg G, Andreasen L, *et al*. Association of Variants Near the Bradykinin Receptor B_2 Gene With Angioedema in Patients Taking ACE Inhibitors. *J Am Coll Cardiol* 2021;78:696–709.
46 van der Harst P, Verweij N. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circ Res* 2018;122:433–43.
47 Eldrup N, Cerqueira C, de la Motte L, *et al*. The Danish vascular Registry, Karbase. *Clin Epidemiol* 2016;8:713–8.
48 Rigas AS, Skytthe A, Erikstrup C, *et al*. The healthy donor effect impacts self-reported physical and mental health - results from the Danish Blood Donor Study (DBDS). *Transfus Med* 2019;29 Suppl 1:65–9.
49 Arvanitis M, Tampakakis E, Zhang Y, *et al*. Genome-Wide association and multi-omic analyses reveal ACTN2 as a gene linked to heart failure. *Nat Commun* 2020;11:1122.
50 Willer CJ, Schmidt EM, Sengupta S, *et al*. Discovery and refinement of loci associated with lipid levels. *Nat Genet* 2013;45:1274–83.

Supplementary Material

Methods

Descriptive statistics were presented as numbers, percentages, mean, medians and standard deviations. The patients were grouped according to their first cardiovascular diagnosis within the hospital admission ICD10 codes presented in Table 1 in the main document. If more than one diagnosis was given at the same day the A diagnosis, the primary cause of the health care contact, was used.

Charlson Comorbidity Index was calculated with the R package 'comorbidity'.[1] It is based on the Charlson score proposed by Quan et al. in 2005.[2] and includes the following comorbid conditions: acute myocardial infarction, congestive heart failure, peripheral vascular disease, cerebrovascular disease, dementia, chronic obstructive pulmonary disease [COPD], rheumatoid disease, peptic ulcer disease, mild and moderate/severe liver disease, diabetes mellitus with and without complications, hemiplegia/paraplegia, renal disease, cancer (any malignancy) and metastatic solid tumour, AIDS/HIV. ICD-8 codes were mapped to ICD-10 codes by an internal ICD-8 to ICD-10 mapping file. The diagnoses prior to the date of a patient's first cardiovascular disease, to inclusion date and end of follow-up or death were included in the Charlson Comorbidity Index calculations.

Cases and controls

In this study cases and controls were defined according to ICD10/ICD8 codes and prescription medicine from the National Patient Register and the Danish Prescription Database.[3,4] The population constitutes of participants from the Danish Blood Donor Study (n = 110 000) and patients from the Copenhagen Hospital Biobank - Cardiovascular Disease Cohort (CHB-CVDC) (n = 96 308).[5] A more detailed description of the phenotypes are presented below.

Phenotype descriptions is inspired by previously described phenotypes.[6,7]

Heart Failure

Laursen IH, *et al. BMJ Open* 2021; 11:e049709. doi: 10.1136/bmjopen-2021-049709

Hospitalization for ICD-10 code for heart failure (I50 and sub codes); or hospitalization for ICD-8 code for heart failure (42709, 42710, 42711, 42719, 78249).

Coronary Artery Disease

Hospitalization for ICD-10 code for coronary artery disease (I21, I22, I23, I24, I25); or hospitalization due to ICD-8 code for coronary artery disease (41199, 41099, 41409, 41499, 41299, 412909, 41009, 41109).

Atrial Fibrillation

Hospitalization for ICD-10 code for atrial fibrillation or atrial flutter (I48); or hospitalization for ICD-8 code for atrial fibrillation or atrial flutter (42793, 42794).

Hypertension

Hospitalization for ICD10 code for essential hypertension (I10), excluding following ICD10 codes from controls I11, I12, I13, I15; or hospitalization for ICD-8 code for essential hypertension (40009, 40299, 40199); and use of at least one antihypertensive drug, the following ATC-codes included: Renin–angiotensin system inhibitors C09; calcium channel blockers C08; b-blockers C07; diuretics C03; antiadrenergic drugs C02A, C02B, and C02C; and other antihypertensives C02DA, C02DB, C02DD, C02DG, and C02L. Controls with prescriptions of beforementioned atc codes were removed from the control group.

Cholesterol measurements

A patient's first measure of High-Density Lipoprotein (HDL), Low-Density Lipoprotein (LDL), Triglycerides (TG) and Total Cholesterol (TC) were retrieved from the Danish Laboratory Database. The following NPU codes were used: NPU01567 (HDL), NPU01568 (LDL), NPU01566 (TC) and NPU04094 (TG). The data for each NPU code used was checked for differences between laboratories.

Inclusion of previously published genetic variants

An extensive literature search was performed for each phenotype. The newest and most updated genome wide association study available, in the period this manuscript was written, of a given disease was included if it was published in English in a high impact peer reviewed journal and include populations of European ancestry. It was assumed that the most updated study was replicating previously found variants.

A genetic variant was included from the study if it was a significant independent variant and if minor allele frequency >1%. If a variant was not included in our dataset a proxy variant ($R^2 > 0.8$) was used, if possible. LDlink proxy tool with European population as reference was used to find proxy variants.[8]

Replication of included variants

For coronary artery disease, atrial fibrillation, heart failure and essential hypertension, the association between the included variants and disease were calculated using a logistic mixed model implemented by SAIGE assuming an additive genetic model and adjusted for year of birth, sex and the first 10 principal components (PCs) to correct for population structure. A variant was replicated if the effect size of the risk allele had the same direction of effect and a $P < 0.05$ (Bonferroni adjusted).

For low-density lipoprotein, high density lipoprotein, total cholesterol and triglyceride, Bolt-LMM was used to calculate associations to disease. The patients first measurement values were inverse rank normalized and adjusted for age at measurement, place of measurement, lipid lowering medications (ATC – C10 within 180 days before measurement) and sex. Residuals were used as the independent variable in association analysis adjusted for year of birth and 10 PCs. Effect sizes were calculated by regressing variants to the inverse rank normalized measurement values adjusted for sex, age at measurement, year of birth, lipid lowering medication, significant places of measurements and 10 PCs.

Effect sizes were retrieved from the included studies. Previously found effect sizes, risk allele frequencies and p-values were plotted against the obtained values from this study. Effect sizes were weighted by allele frequencies. Power calculations for each SNP were based on standard error and effect size and performed in R[9].

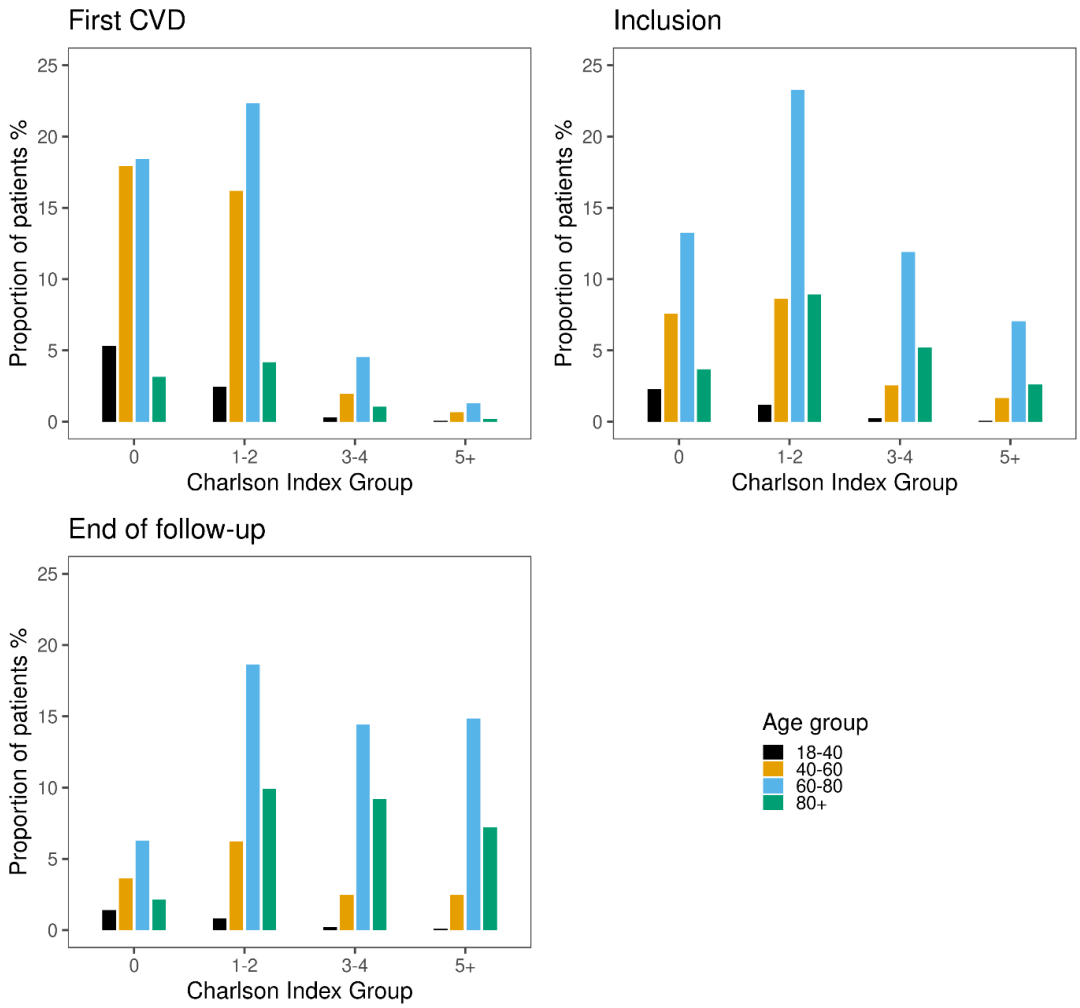
The following tools were used: PLINK 2[10,11], SAIGE 36.3.3[12], BOLT-LMM 2.3.4[13], flashpca 2.0[14], LDlink[8], Python 3 , Rstudio 1.2.1335[15].

Comparison of different study designs

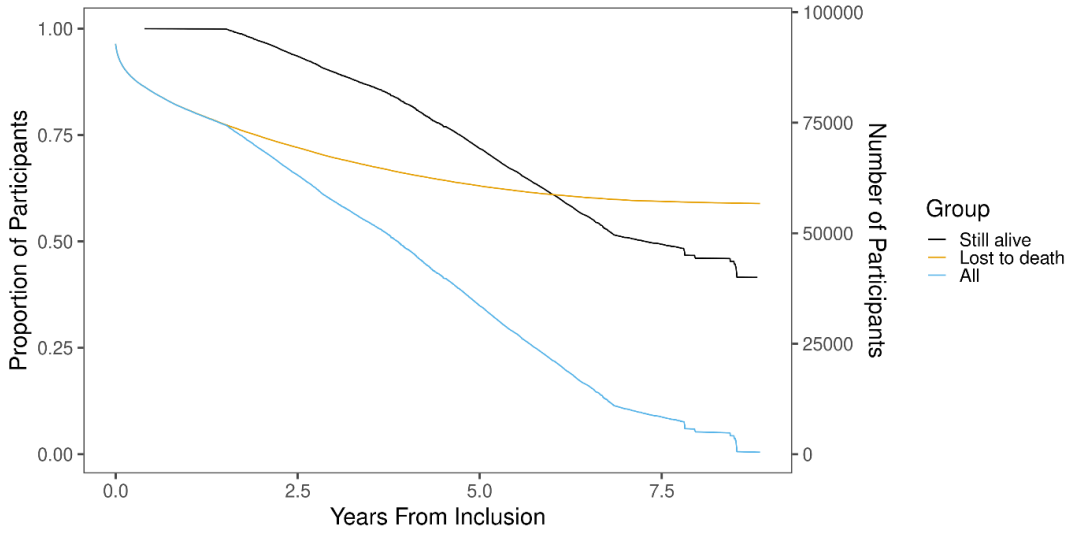
We evaluated the study designs with CHB-CVDC as the only controls and DBDS as the only controls respectively.

We employed LD Score regression to investigate residual confounding. Residual confounding was evaluated through the LD Score regression intercept and the attenuation ratio (the ratio between the intercept and the mean χ^2 statistic).[16] Genetic correlations were estimated using LD Score Regression with the 1000Genomes EUR v3 LD reference panel. Summary stats for a pre-selected list of phenotypes were retrieved from the GWAS catalog[17]. See Table 2 for a complete list.

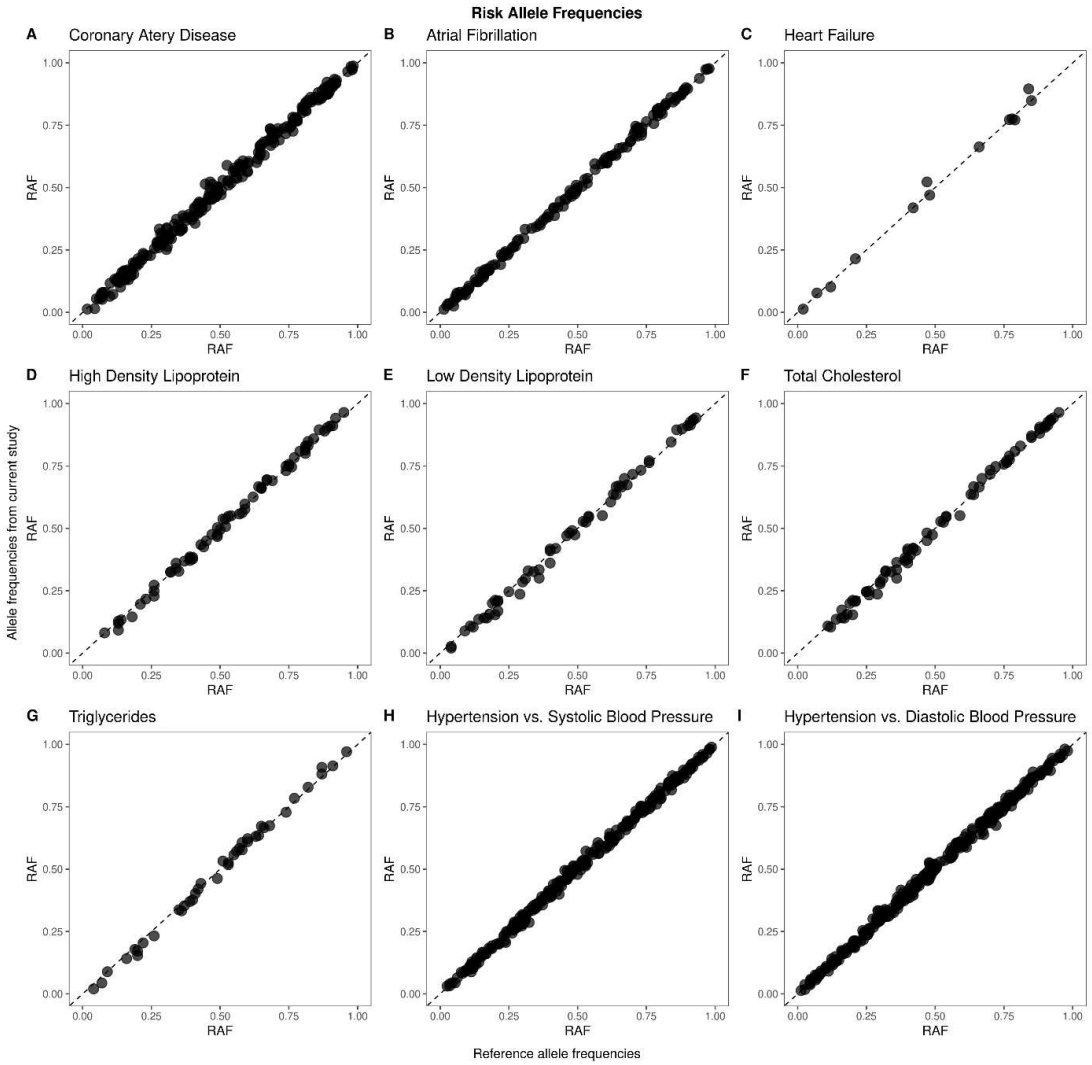
Supplementary Figures



Supplementary Figure 1: The patients in CHB-CVDC were classified by the Charlson Comorbidity Index prior to the date of the first cardiovascular diagnosis, the inclusion date and at the end of follow-up or death. The patients are stratified in age groups. CVD: cardiovascular disease

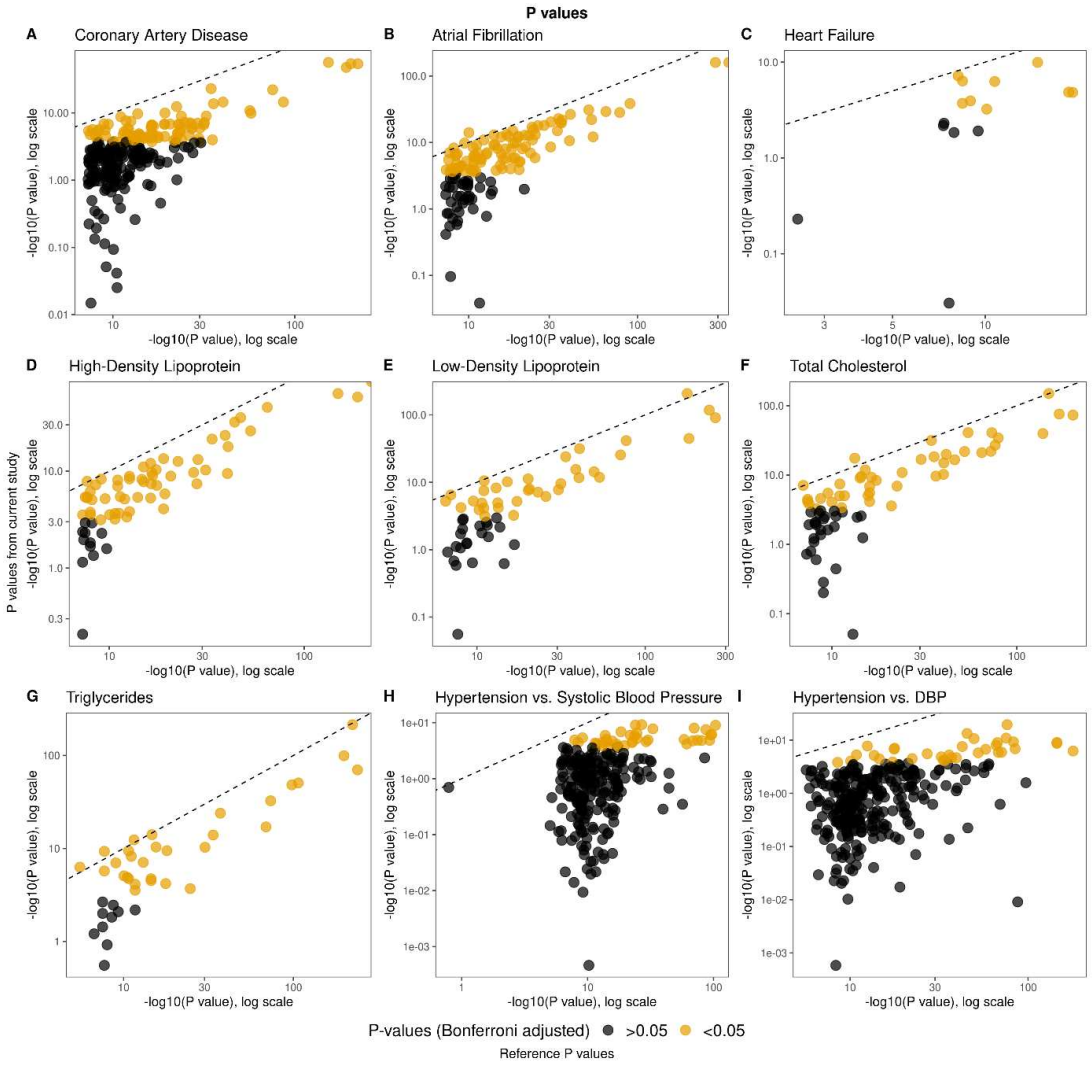


Supplementary Figure 2: The follow-up time from inclusion to either death or present was plotted. When a patient was either dead or the data extraction date (present) was reached the patient is subtracted from the total number of individuals. The black line shows the follow-up time for those still alive at present (56 259). The yellow line shows the follow-up time for those who are dead (39 539). The blue line shows the total follow-up time. More patients die within the first two years and then the curve flattens. The irregular part in the end of the survivor curve corresponds to some delays in inclusion times from the start of inclusion in 2009.



Supplementary Figure 3: This figure shows the correlation between the risk allele frequencies between the reference studies and this work. The dotted lines are the expected correlations of 1.

1



2

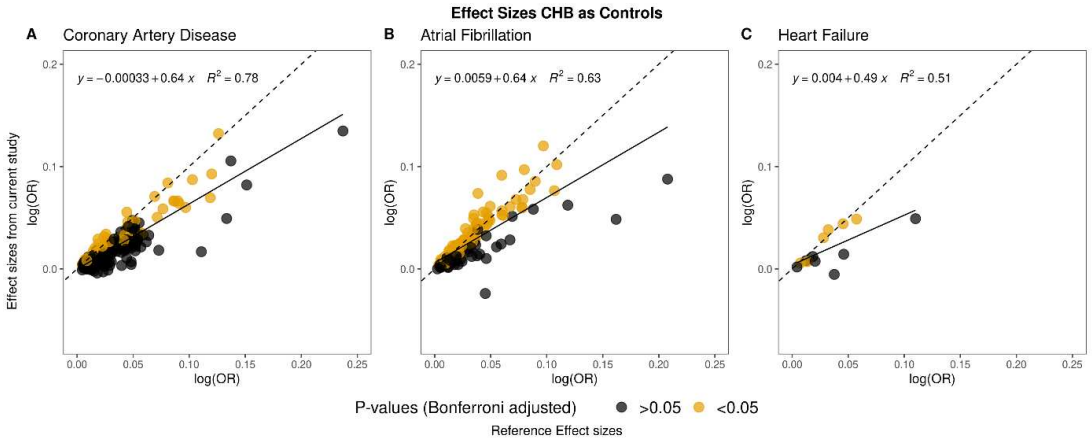
3 ***Supplementary Figure 4: For each trait the -log10(p-values) from this study were plotted against the -log(p-***

4 ***values) from the reference studies. The axes are on logarithmic scales. The dotted lines correspond to a***

5 ***correlation of 1.***

6

8

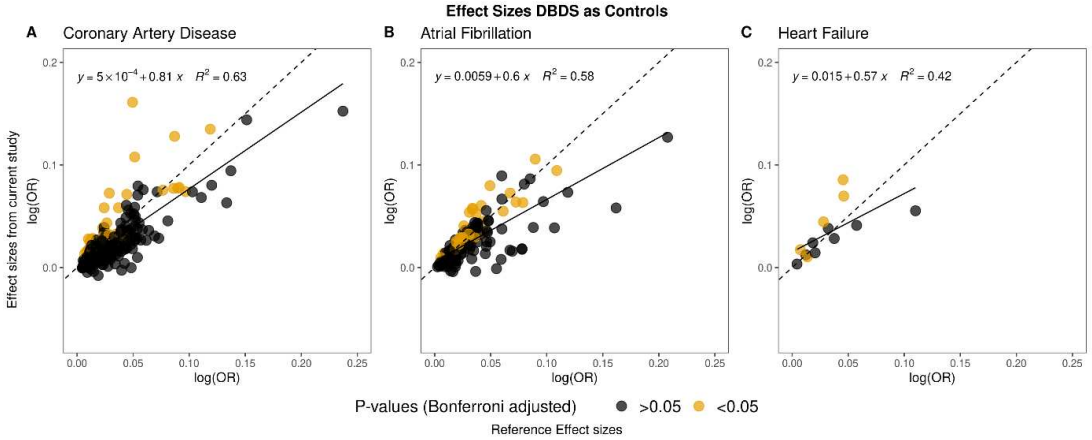


1

2 ***Supplementary Figure 5: Comparison of effect sizes between reference studies and the CHB-CVDC as controls***

3 ***study design.***

4

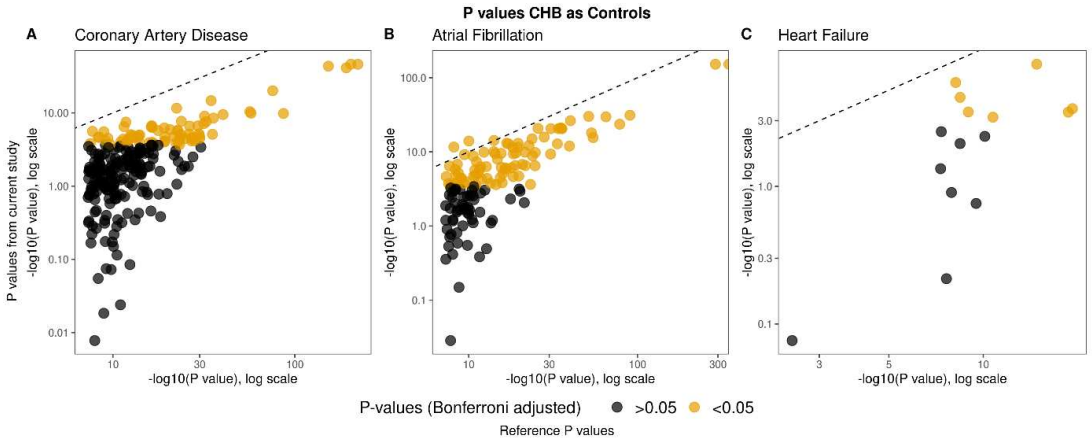


5

6 ***Supplementary Figure 6: Comparison of effect sizes between reference studies and the DBDS as controls study***

7 ***design.***

9

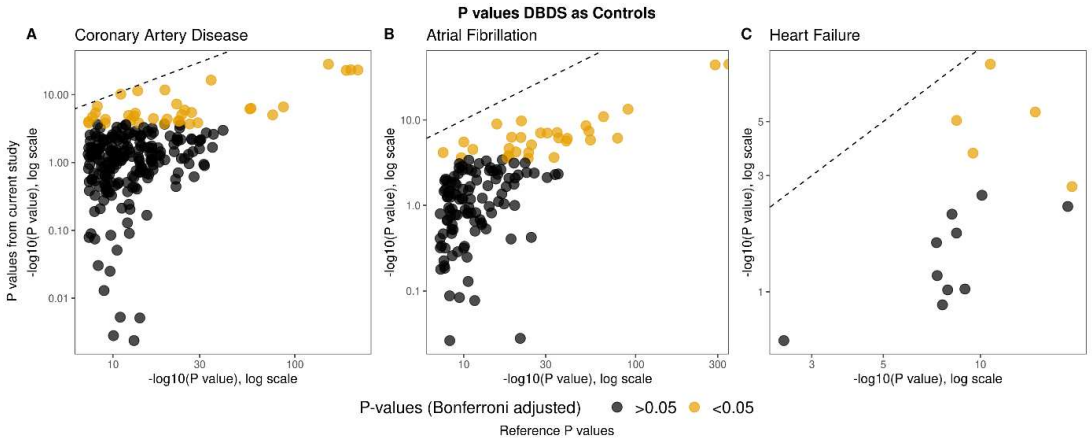


1

2 ***Supplementary Figure 7: Comparison of p-values between reference studies and the CHB-CVDC as controls***

3 ***study design.***

4



5

6 ***Supplementary Figure 8: Comparison of p-values between reference studies and the DBDS as controls study***

7 ***design.***

10

Supplementary Tables

Supplementary Table 1: Population characteristics

	ICD-10 Codes	N (% females)	Year of birth, Mean (SD)	Age at inclusion, Mean (SD)	Age at first CVD, Mean (SD)	Hemoglobin Complete %	Troponin T Complete %	Creatinine Complete %	CRP Complete %	LDL Complete %	HDL Complete %	Na^+ Complete %	K^+ Complete %
Hypertension and hypertensive cardiac diseases	I10-15	29 546 (55%)	1943.8 (13.1)	69.6 (13.2)	62.5 (14.3)	98.97 %	21.54 %	98.26 %	97.35 %	85.02 %	89.90 %	98.99 %	98.98 %
Coronary artery diseases and atherosclerosis	I20-25, I70	24 795 (36%)	1943.6 (12.7)	69.5 (12.8)	60.0 (12.4)	98.48 %	37.27 %	97.67 %	96.64 %	87.49 %	90.78 %	98.57 %	98.55 %
Lipid disorders	E78	4046 (50%)	1948.3 (13.1)	65.2 (13.2)	58.5 (13.8)	98.91 %	20.27 %	98.15 %	97.21 %	90.29 %	94.07 %	98.84 %	98.84 %
Cardiac arrhythmia	I44-49	16 208 (44%)	1945.5 (16.7)	67.9 (16.8)	60.5 (17.2)	98.65 %	22.77 %	97.80 %	96.99 %	80.81 %	85.48 %	98.53 %	98.51 %
Heart failure, cardiac valve disorders, and myocardial diseases	I50, I34-39, I05-09, I40-44	7264 (41%)	1943.9 (16.1)	69.2 (16.2)	62.5 (17.6)	98.73 %	36.65 %	97.88 %	97.94 %	82.30 %	86.38 %	98.61 %	98.61 %
Vascular disorders and aneurysms	I71-79	3121 (40%)	1942.6 (12.2)	70.3 (12.3)	62.3 (13.8)	98.08 %	31.66 %	97.18 %	97.53 %	85.61 %	88.72 %	98.17 %	98.14 %
Cerebrovascular diseases and cerebral hemorrhage	I60-69	8401 (44%)	1941.1 (12.2)	72.1 (12.4)	63.1 (13.2)	98.76 %	23.90 %	97.74 %	97.82 %	85.69 %	89.85 %	98.70 %	98.68 %
Pulmonary heart and pulmonary circulation diseases	I26-28	1521 (49%)	1942.4 (13.6)	70.7 (13.8)	59.7 (15.8)	98.75 %	31.62 %	98.09 %	97.50 %	81.53 %	87.44 %	98.69 %	98.69 %
Vascular kidney disease	N17-19	1406 (39%)	1945.2 (14.2)	67.4 (14.4)	61.0 (16.9)	98.44 %	26.53 %	97.51 %	98.08 %	79.16 %	83.85 %	98.44 %	98.36 %
Above diseases combined		**96 308 (45%)**	**1943.9 (13.9)**	**69.3 (14.0)**	**61.3 (14.5)**	**98.71 %**	**27.65 %**	**97.90 %**	**97.21 %**	**84.90 %**	**89.12 %**	**98.71 %**	**98.69 %**

N: Number of cases, SD: Standard Deviation, CRP: C-Reactive Protein, LDL: Low Density Lipoprotein, HDL: High Density Lipoprotein, Na^+: Sodium ion, K^+: Potassium ion

Supplementary Table 2: Charlson Comorbidity Index

Charlson index calculated at first cardiovascular diagnosis

sex	0	1-2	3-4	>=5	total
Counts	43 130	43 506	7520	2152	96 308

	M	22 287	24 845	3956	1185
	F	22 843	18 661	3564	967
% of total		45	45	8	2
	M	52	57	53	55
	F	53	43	47	45

Charlson index calculated at inclusion

	sex	0	1-2	3-4	>=5	total
Counts		25 745	40 361	19 070	10 913	96 089
	M	13 213	22 102	10 735	6616	
	F	12 532	18 259	8335	4297	
% of total		27	42	20	11	
	M	51	55	56	61	
	F	49	45	44	39	

Charlson index calculated at the end of follow up

	sex	0	1-2	3-4	>=5	total
Counts		12 977	34 375	25 300	23 656	96 308
	M	6878	18 258	14 045	14 427	
	F	6099	16 117	11 255	9229	
% of total		13	36	26	25	
	M	53	53	56	61	
	F	47	47	44	39	

Supplementary Table 3: An overview of diagnosis assigned prior to the first cardiovascular disease. If a patient has received multiple diagnosis within a ICD10 chapter they were only counted as one diagnosis. E.g 5 465 have a digestive disorder prior to a cardiac arrythmia diagnosis as their first cardiovascular disease.

First assigned cardiovascular diagnosis	-	Certain conditions originating in the perinatal period	Certain infections and parasitic diseases	Codes for special purposes	Congenital malformations, deformations and chromosomal abnormalities	Diseases of the blood and blood-forming organs and certain disorders involvning the immune mechanism	Diseases of the circulatory system included in CHB-CVDC	Diseases of the circulatory system included in CHB-CVDC diagnosed the same day
Cardiac arrythmia	0	236	2187	133	688	837	16208	2039
Cerebrovascular diseases and cerebral haemorrhage	0	16	855	45	266	296	8401	2672
Heart failure, cardiac valve disorders and myocardial diseases	0	118	1058	43	467	474	7264	1639
Hypertensive cardiac diseases	<10	139	3867	273	1049	1722	29546	2176
Ischemic heart diseases and atherosclerosis	0	80	2435	176	726	711	24795	4319
Lipid disorders	0	42	547	52	160	188	4046	1185
Pulmonary heart diseases	0	<10	202	<10	64	88	1521	232
Vascular disorders and aneurysms	0	12	359	17	127	129	3121	307
Vascular kidney diseases	0	12	394	13	107	215	1406	323

Supplementary Table 3 continued:

First assigned cardiovascular diagnosis	Diseases of the circulatory system not included in CHB-CVDC	Diseases of the digestive system	Diseases of the eye and adnexa, Diseases of the ear and mastoid process	Diseases of the genitourinary system	Diseases of the musculoskeletal system and connective tissue	Diseases of the nervous system	Diseases of the respiratory system	Diseases of the skin and subcutaneous tissue
Cardiac arrythmia	1751	5465	4294	4973	6296	2145	4258	1836
Cerebrovascular diseases and cerebral haemorrhage	682	2635	2252	2495	3033	1941	1556	898
Heart failure, cardiac valve disorders and myocardial diseases	1083	2589	2183	2175	2876	973	2188	878
Hypertensive cardiac diseases	3177	11929	8869	11674	13411	4777	7100	3914
Ischemic heart diseases and atherosclerosis	2180	7983	5281	6903	9721	2809	4803	2825
Lipid disorders	442	1665	1138	1492	1976	1015	913	635
Pulmonary heart diseases	428	505	346	506	604	189	524	164
Vascular disorders and aneurysms	356	1128	829	920	1308	401	665	399
Vascular kidney diseases	191	597	442	675	608	218	499	233

Supplementary Table 3 continued:

First assigned cardiovascular diagnosis	Endocrine, nutritional and metabolic diseases	External causes of morbidity and mortality	Factors influencing health status and contact with health services	Injury, poisoning and certain other consequences of external causes	Mental and behavioural disorders	Neoplasms	Pregnancy, childbirth and the puerperium	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
Cardiac arrythmia	2452	194	11302	10193	2253	4158	1963	6539
Cerebrovascular diseases and cerebral haemorrhage	1286	78	5487	4910	1305	1873	887	3121
Heart failure, cardiac valve disorders and myocardial diseases	1232	66	5254	4440	1019	1783	663	3050
Hypertensive cardiac diseases	8609	574	21820	18895	4786	10230	4765	12924
Ischemic heart diseases and atherosclerosis	3223	184	15421	14076	2905	4890	2417	8150
Lipid disorders	1584	78	3129	2728	742	1254	742	1820
Pulmonary heart diseases	198	15	935	868	206	427	189	566
Vascular disorders and aneurysms	391	19	2180	1924	409	745	297	1110
Vascular kidney diseases	493	21	1032	887	280	449	140	694

Supplementary Table 4: Patients in CHB-CVDC were grouped according to their first cardiovascular disease. The number of patients with diagnoses within the other cardiovascular diseases were calculated. If a patient has received multiple diagnoses within a disease group, they were only counted as one diagnosis. E.g. 2 403 patients receive a cardiac arrythmia diagnosis as their first cardiovascular disease and later in life one or more diagnoses within the group cerebrovascular diseases and cerebral haemorrhage.

First assigned cardiovascular diagnosis	Cardiac arrythmia	Cerebrovascular diseases and cerebral haemorrhage	Heart failure, cardiac valve disorders and myocardial diseases	Hypertensive cardiac diseases	Ischemic heart diseases and atherosclerosis	Lipid disorders	Pulmonary heart diseases	Vascular disorders and aneurysms	Vascular kidney diseases
Cardiac arrythmia	0	2403	4630	6610	4274	2334	747	1245	1421
Cerebrovascular diseases and cerebral haemorrhage	2897	0	2126	6203	2931	4050	431	1205	1050
Heart failure, cardiac valve disorders and myocardial diseases	3187	1077	0	3448	2844	1634	419	877	1078
Hypertensive cardiac diseases	6751	4951	5523	0	7110	6419	1367	2877	3821
Ischemic heart diseases and atherosclerosis	7778	4257	8639	13777	0	10581	1328	4696	2907
Lipid disorders	699	672	637	2128	1062	0	149	383	352
Pulmonary heart diseases	674	276	574	945	638	331	0	194	212
Vascular disorders and aneurysms	1055	654	967	2081	1911	944	206	0	493
Vascular kidney diseases	558	255	432	989	467	251	80	207	0

Supplementary Table 5: Top 100 diagnosis in CHB-CVDC

	ICD10 code	Number of patients with diagnosis	Medical condition
1	I109	64 247	Essential hypertension, unspecified
2	I489	27 315	Atrial fibrillation and flutter, unspecified
3	E780	26 581	Hypercholesterolemia
4	I259	22 305	Chronic ischemic heart disease, unspecified
5	I209	21 409	Angina pectoris, unspecified
6	I509	18 294	Heart failure, unspecified
7	I251	12 825	Atherosclerotic heart disease
8	I649	11 272	Apoplexia cerebri, unspecified
9	I639	11 076	Cerebral infarction, unspecified
10	I219	10 383	Acute myocardial infarction, unspecified
11	I694	9972	Sequelae of stroke, not specified as haemorrhage or infarction
12	I489B	8639	Atrial fibrillation and atrial flutter, unspecified
13	I252	8611	Old myocardial infarction
14	I214	8486	Acute subendocardial myocardial infarction
15	I200	7801	Unstable angina
16	I500	7536	Congestive heart failure
17	N189	7256	Chronic kidney disease, unspecified
18	I350	7242	Aortic (valve) stenosis
19	I702	6920	Atherosclerosis of arteries of extremities
20	E785	6864	Hyperlipidaemia, unspecified
21	I739A	6122	Intermittent claudication
22	I471	5356	Supraventricular tachycardia
23	N179	5158	Acute renal failure, unspecified
24	I480	4917	Paroxysmal atrial fibrillation
25	I213	4182	Acute transmural myocardial infarction of unspecified site
26	N199	3964	Unspecified kidney failure
27	I501	3829	Left ventricular failure
28	I482	3423	Chronic atrial fibrillation
29	I693	3216	Sequelae of cerebral infarction
30	I269	2982	Pulmonary embolism without mention of acute cor pulmonale
31	I709	2967	Generalized and unspecified atherosclerosis
32	I340	2931	Mitral (valve) insufficiency
33	I739C	2929	Peripheral vascular disease, unspecified +
34	I714	2919	Abdominal aortic aneurysm, without mention of rupture
35	I702A	2651	Atherosclerotic gangrene
36	I499	2622	Cardiac arrhythmia, unspecified
37	I442	2473	Atrioventricular block, complete
38	I10	2294	Essential (primary) hypertension
39	I351	2288	Aortic (valve) insufficiency
40	I460	2167	Cardiac arrest with successful resuscitation

41	I269A	2155	Pulmonary embolism NOS
42	I119	2121	Hypertensive heart disease without (congestive) heart failure
43	I469	2108	Cardiac arrest, unspecified
44	I159	2107	Secondary hypertension, unspecified
45	I479	2094	Paroxysmal tachycardia, unspecified
46	I619	1979	Intracerebral haemorrhage, unspecified
47	I208	1977	Other forms of angina pectoris
48	I493	1890	Ventricular premature depolarization
49	I489BB	1876	Atrial fibrillation and atrial flutter, unspecified
50	I210	1860	Acute transmural myocardial infarction of anterior wall
51	I258	1762	Other forms of chronic ischaemic heart disease
52	I489A	1670	Atrial fibrillation and atrial flutter, unspecified
53	I652	1609	Occlusion and stenosis of carotid artery
54	I495	1579	Sick sinus syndrome
55	I481	1477	Persistent atrial fibrillation
56	I359	1469	Aortic valve disorder, unspecified
57	I249	1465	Acute ischaemic heart disease, unspecified
58	I472	1454	Ventricular tachycardia
59	I420	1442	Dilated cardiomyopathy
60	I211B	1350	Acute transmural myocardial infarction of inferior wall
61	I429	1332	Cardiomyopathy, unspecified
62	I211	1319	Acute transmural myocardial infarction of inferior wall
63	I110	1304	Hypertensive heart disease with (congestive) heart failure
64	I210B	1162	Acute transmural myocardial infarction of anterior wall
65	I491	1143	Atrial premature depolarization
66	N185	1137	Chronic kidney disease, stage 5
67	I739	1086	Peripheral vascular disease, unspecified
68	I609	1061	Subarachnoid haemorrhage, unspecified
69	I633	1047	Cerebral infarction due to thrombosis of cerebral arteries
70	N180	1025	Hypertensive renal disease
71	I719	969	Aortic aneurysm of unspecified site, without mention of rupture
72	I743	948	Embolism and thrombosis of arteries of lower extremities
73	I495B	946	Tachycardia-bradycardia syndrome
74	I255	919	Ischaemic cardiomyopathy
75	I201	856	Angina pectoris with documented spasm
76	I712	853	Thoracic aortic aneurysm, without mention of rupture
77	I389	811	Endocarditis, unspecified
78	I501C	791	Left ventricular failure, pulmonary congestion
79	I443	784	Other and unspecified atrioventricular block
80	E782	768	Mixed hyperlipidaemia
81	E789	767	Disorder of lipoprotein metabolism, unspecified
82	I270	751	Other adrenocortical overactivity
83	I48	744	Atrial fibrillation and flutter
84	I279	740	Pulmonary heart disease, unspecified

85	I691	721	Sequelae of intracerebral haemorrhage
86	I708	705	Atherosclerosis of other arteries
87	I441	678	Atrioventricular block, second degree
88	I21	673	Acute myocardial infarction
89	E780B	651	Familial hypercholesterolaemia
90	I632	649	Cerebral infarction due to unspecified occlusion or stenosis of precerebral arteries
91	I472A	627	Ventricular tachycardia
92	I612	623	Intracerebral haemorrhage in hemisphere, unspecified
93	I501B	615	Left ventricular failure, Oedema of lung
94	I470	598	Re-entry ventricular arrhythmia
95	I352	589	Aortic (valve) stenosis with insufficiency
96	N183	585	Chronic kidney disease, stage 3
97	I471R	577	Supraventricular tachycardia, atrioventricular [AV]: re-entrant (nodal) [AVNRT]
98	I728	567	Aneurysm and dissection of other specified arteries
99	N184	566	Chronic kidney disease, stage 4
100	N178	561	Other acute renal failure

Supplementary Table 6: Overview of causes of death in CHB-CVDC according to the Danish Registry of Causes of Death

Cause of death	***n***	***% of dead***
Cancer	*13715*	*35.8*
Heart disease	*6478*	*16.9*
Respiratory disease	*3937*	*10.3*
Other diseases in the circulatory system	*3097*	*8.1*
Diseases of the digestive system	*2019*	*5.3*
Endocrine, nutritional, and metabolic diseases	*1427*	*3.7*
Unknown medical information	*1185*	*3.1*
Mental and behavioural disorders	*1102*	*2.9*
Diseases of the nervous system	*946*	*2.5*
Accidents	*917*	*2.4*
Certain infectious and parasitic diseases	*885*	*2.3*
Diseases of the genitourinary system	*799*	*2.1*
Symptoms, signs, and abnormal clinical and laboratory findings, not elsewhere classified	*644*	*1.7*
Neoplasms	*288*	*0.8*
Diseases of the musculoskeletal system and connective tissue	*262*	*0.7*
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	*212*	*0.6*
Suicide	*175*	*0.5*
Diseases of the skin and subcutaneous tissue	*69*	*0.2*
Congenital malformations, deformations, and chromosomal abnormalities	*65*	*0.2*
Other causes	*27*	*0.1*

Supplementary Table 7: Overview of databases and variables

Databases	**Variables**
The National Patient Registry	ICD10/ICD8, SKS codes. Patient type, hospital code, department code, admission date, admission time, discharge date, discharge time, days of treatment, mode of admission, reason for contact, mode of discharge, speciality, action diagnosis, accident code, accident code counterpart, accident code case, accident code activity, accident code place, accident code mechanism, accident code traffic, outpatient date, diagnosis, type of diagnosis, additional diagnosis, diagnosis modification, SKS procedure (operation), SKS procedure classification (operation), additional code (operation), procedure hospital (operation), procedure department (operation), procedure day (operation), procedure hour (operation), procedure minute (operation), SKS procedure (diagnosis and treatment), SKS procedure classification (diagnosis and treatment), additional code (diagnosis and treatment), procedure hospital (diagnosis and treatment), procedure department (diagnosis and treatment), procedure day (diagnosis and treatment), procedure hour (diagnosis and treatment), procedure minute (diagnosis and treatment), and other similar variables that are needed for these classifications.
The Danish Registry of Causes of Death	Date of death, cause of death, place of death.
Central Person Registry (CPR)	Status (dead, emigrated or living) and also relatedness between cases.
The Danish National Prescription Registry	Medicine bought in pharmacies with prescription
BigTempHealth	Validate diagnoses and drug prescriptions. Include data from electronic health records (EHRs) and laboratory test results in the phenotypic characterization.
Danish Agency of Labour Market and Recruitment	Working history and occupation
Health Care Statistics Registry	Speciality, service code, year, month, number of services to access risk factors for heart diseases. This also includes data on dental diseases and treatments to access dental risk factors for heart disease.
The Danish Laboratory Database	Patient CPR, sampling date, sampling time, analysis code, Laboratory ID-CODE, value, unit, result type, reference interval upper limit, reference interval lower limit, and NPU code.
The Copenhagen GP Laboratory database	Electrocardiography data
The Regions imaging data, including data in IntelliSpace, PACS and in Xeroviewer	Echocardiography, computer tomography (CT), Magnetic Resonans (MR), imaging, coronary arteriography, and nuclear imaging.

The Regions ECG and Holter database; Kardia	
National Clinical Registries (RKKB):	
Danish Anesthesia Database	Information on blood pressure, height, weight, smoking, co-morbidity, ID, date, department, hospital, region, indication, operation information, complications, ASA score, and other related variables
Danish Stroke Registry	Disease classification, information on smoking history, cardiac disease, comorbidity, ID, date, department, hospital, region, alcohol, AK treatment, diabetes, bleeding or infraction, hypertension, treatment, outcome, height, weight, and other related variables
Danish Heart Registry	Disease classification, indication, operation type, left ventricular ejection fraction (LEVF), coronary artery pathology, height, weight, smoking, diabetes history, package, treatment, ID, date, department, hospital, region, complications, EURO score, lung disease, cerebrovascular disease, previous cardiac surgery, creatinine, endocarditis, pulmonal hypertension, angina, left ventricular dysfunction, and other related variables
Danish Heart Failure Registry	Disease classification, disease severity measures, co-morbidity, life style factors, treatments, ID, date, department, hospital, region alcohol, BMI, diabetes, hypertension, tobacco, ejection fraction, NYHA group, COPD, and other related variables
Danish Heart Disease Rehabilitation Database	Disease classification, cardiovascular disease history, co-morbidity, LVEF, lifestyle factors, cardiac risk factors, current disease measures including CSS and NYHA classification, waist circumference, current treatment, current laboratory values, HRQoL score, ID, date, department, hospital, region, treatment, liver values, CK, BMI, depression, GFR, diabetes values, lung function, exercise test, blood pressure, blood sugar values incl. HbA1c, lipids, HRQL, ejection fraction, height, weight, and other related variables.
Danarrest and the Danish Registry of Cardiac Arrest	Cardiac cause of death, ID, date, department, hospital, region and similar data as for the other databases
Atrial Fibrillation Database	Information on ID, date, department, hospital, region and similar data as for the other databases
Danish Pacemaker and ICD register	Information on ID, date, department, hospital, region, devices, indication, and similar data as for the other databases
Danish Ablation Database	Information on ID, date, department, hospital, region, indication, conduction disorder, operation, and similar data as for the other databases

The PATS Database	Includes data on patients that has undergone invasive cardiac procedures until 2016 – we will include data on ID, date, department, hospital, region, indication, operation procedures, risk factors, laboratory values, co-morbidity and organ functions, and similar data as for the other databases
The Database for Familial Hypercholesterolemia	Information on ID, date, department, hospital, region, indication, dyslipidemia disorder, genetics, and similar data as for the other databases
Progeny	The database holds data on clinical and genetic findings in families with inherited cardiac diseases including pedigrees. Information on ID, date, department, hospital, region, indication for assessment, findings, and similar data as for the other databases
Danish Obesity Surgery Database	Information on ID, date, department, hospital, region, diabetes, height, weight, diabetes type, hypertension, lipids, depression, asthma, COPD, PCO, joint complaints, tobacco, alcohol, HRQL, treatment, and other related variables
Danish Registry for Chronic Obstructive Lung Diseases	Information on ID, date, department, hospital, region, BMI, co-morbidity, lung function measures, diagnosis, tobacco, tobacco intervention, and other related variables
Sleep Apnea Database	Information on ID, date, department, hospital, region, diagnosis, treatment
KARBASE	Information on ID, date, department, hospital, region, diagnosis, intervention, death, date of death, date of operation, co-morbidity, BMI, tobacco, alcohol, and other related variables

Supplementary Table 8: P-values and confidence intervals of the effect sizes regression lines.

	Entire cohort with DBDS		**CHB-CVDC as controls**		**DBDS as controls**	
Phenotype	p-value	2.5%:97.5%	p-value	2.5%:97.5%	p-value	2.5%:97.5%
CAD	<2.2*10-16	0.76:0.88	<2.2*10-16	0.60:0.68	<2.2*10-16	0.73:0.89
AF	<2.2*10-16	0.65:0.83	<2.2*10-16	0.56:0.72	<2.2*10-16	0.52:0.69
HF	0.01079	0.14:0.83	0.0062	0.17:0.80	0.0161	0.13:1.01
HDL	<2.2*10-16	0.81:0.95				
LDL	<2.2*10-16	0.61:0.70				
TC	<2.2*10-16	0.66:0.75				
TG	<2.2*10-16	0.88:1.1				
SBP	<2.2*10-16	0.09:0.12				
DBP	<2.2*10-16	0.13:0.18				

Supplementary Table 9: Comparison of different control groups

Atrial Fibrillation

	Variants with concordant direction of effect	Replicated/Total	Replicated/Power to Replicate	Cases	Controls
CHB as cases, DBDS and CHB as controls. As presented in Table 2	137/140 (98%)	96/140 (69%)	96/109 (88%)	30 229	157 669
CHB as cases and controls	139/140 (99%)	83/140 (59%)	83/102 (81%)	30 152	65 870
CHB as cases, DBDS as controls	136/140 (97%)	30/140 (21%)	30/54 (56%)	30 152	91 280

Coronary Artery Disease

	Variants with concordant direction of effect	Replicated/Total	Replicated/Power to Replicate	Cases	Controls
CHB as cases, DBDS and CHB as controls. As presented in Table 2	236/241 (98%)	90/241 (37%)	90/137 (66%)	33 746	154 311
CHB as cases and controls	233/241 (97%)	71/241 (29%)	71/113 (63%)	37 862	69 824
CHB as cases, DBDS as controls	232/241 (96%)	35/241 (15%)	35/72 (49%)	38 561	104 011

Heart Failure

	Variants with concordant direction of effect	Replicated/Total	Replicated/Power to Replicate	Cases	Controls
CHB as cases, DBDS and CHB as controls. As presented in Table 2	14/15 (93%)	9/15 (60%)	9/10 (90%)	21 443	167 068
CHB as cases and controls	14/15 (93%)	7/15 (47%)	7/9 (78%)	21 421	74601
CHB as cases, DBDS as controls	15/15 (100%)	5/15 (33%)	5/6 (83%)	21 421	91892

Supplementary Table 10: Residual confounding

Phenotype	Design	Intercept	Intercept SE	Ratio
AF	Entire cohort with DBDS	1,0957	0,0087	0,3366
AF	CHB-CVDC	1,0356	0,0079	0,1566
AF	DBDS control	1,1030	0,0088	0,5366
CAD	Entire cohort with DBDS	1,1163	0,0092	0,3995
CAD	CHB-CVDC	1,0303	0,0085	0,1887
CAD	DBDS control	1,1638	0,0099	0,5091
HF	Entire cohort with DBDS	1,0705	0,0073	0,5313
HF	CHB-CVDC	1,0203	0,0065	0,3414
HF	DBDS control	1,0977	0,0079	0,5709

Supplementary Table 11: Genetic Correlation

Phenotype	Design	Comparison study	rg	se	p
AF	Entire cohort with DBDS	Nielsen 2018	1,0778	0,0417	2,76E-147
AF	CHB-CVDC	Nielsen 2018	0,9647	0,035	1,83E-167
AF	CHB-CVDC case / DBDS control	Nielsen 2018	NA	NA	NA
CAD	Entire cohort with DBDS	Cardiogramplusc4d	1,0452	0,0675	4,32E-54
CAD	CHB-CVDC	Cardiogramplusc4d	0,9249	0,0667	1,12E-43
CAD	CHB-CVDC case / DBDS control	Cardiogramplusc4d	NA	NA	NA
HF	Entire cohort with DBDS	Shah 2020	0,9808	0,1027	1,26E-21
HF	CHB-CVDC	Shah 2020	-0,8248	0,1289	1,57E-10
HF	CHB-CVDC case / DBDS control	Shah 2020	NA	NA	NA

Supplementary references

1 Gasparini A. comorbidity: An R package for computing comorbidity scores. *JOSS* 2018;**3**:648. doi:10.21105/joss.00648

2 Quan H, Sundararajan V, Halfon P, *et al.* Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care* 2005;**43**:1130–9. doi:10.1097/01.mlr.0000182534.19832.83

3 Schmidt M, Schmidt SAJ, Sandegaard JL, *et al.* The Danish National Patient Registry: a review of content, data quality, and research potential. *Clin Epidemiol* 2015;**7**:449–90. doi:10.2147/CLEP.S91125

4 Kildemoes HW, Sørensen HT, Hallas J. The Danish National Prescription Registry. *Scand J Public Health* 2011;**39**:38–41. doi:10.1177/1403494810394717

5 Hansen TF, Banasik K, Erikstrup C, *et al.* DBDS Genomic Cohort, a prospective and comprehensive resource for integrative and temporal analysis of genetic, environmental and lifestyle factors affecting health of blood donors. *BMJ Open* 2019;**9**:e028401. doi:10.1136/bmjopen-2018-028401

6 Aragam KG, Chaffin M, Levinson RT, *et al.* Phenotypic Refinement of Heart Failure in a National Biobank Facilitates Genetic Discovery. *Circulation* 2019;**139**:489–501. doi:10.1161/CIRCULATIONAHA.118.035774

7 Krogager ML, Torp-Pedersen C, Mortensen RN, *et al.* Short-term mortality risk of serum potassium levels in hypertension: a retrospective analysis of nationwide registry data. *Eur Heart J* 2017;**38**:104–12. doi:10.1093/eurheartj/ehw129

8 Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* 2015;**31**:3555–7. doi:10.1093/bioinformatics/btv402

9 R Core Team. R: A Language and Environment for Statistical Computing. 2020.https://www.R-project.org/

10 Purcell S, Neale B, Todd-Brown K, *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;**81**:559–75. doi:10.1086/519795

11 Chang CC, Chow CC, Tellier LC, *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;**4**. doi:10.1186/s13742-015-0047-8

12 Zhou W, Nielsen JB, Fritsche LG, *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* 2018;**50**:1335–41. doi:10.1038/s41588-018-0184-y

13 Loh P-R, Tucker G, Bulik-Sullivan BK, *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* 2015;**47**:284–90. doi:10.1038/ng.3190

14 Abraham G, Qiu Y, Inouye M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* 2017;**33**:2776–8. doi:10.1093/bioinformatics/btx299

15 RStudio Team. RStudio: Integrated Development for R. 2020.https://www.rstudio.com/

16 Bulik-Sullivan BK, Loh P-R, Finucane H, *et al.* LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies. *Nat Genet* 2015;**47**:291–5. doi:10.1038/ng.3211

17 Buniello A, MacArthur JAL, Cerezo M, *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 2019;**47**:D1005–12. doi:10.1093/nar/gky1120