JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY
© 2021 PUBLISHED BY ELSEVIER ON BEHALF OF THE
AMERICAN COLLEGE OF CARDIOLOGY FOUNDATION

VOL. 78, NO. 1, 2021

JACC FOCUS SEMINAR: THE BEST OF POPULATION RESEARCH STUDIES

JACC FOCUS SEMINAR

United Kingdom Biobank (UK Biobank)

JACC Focus Seminar 6/8

Rishi Caleyachetty, MBBS, PhD, Thomas Littlejohns, PhD, Ben Lacey, MBChB, DPhil, Jelena Bešević, PhD, Megan Conroy, MSc, Rory Collins, FRS, FMedSci, Naomi Allen, DPhil

ABSTRACT

An increasing number of people are now living with cardiovascular disease (CVD), with concomitant CVD-related hospitalizations, operations, and prescriptions. To ultimately deliver optimal cardiovascular care, access to population-based biobanks with data on multiomics, phenotypes, and lifestyle risk factors are crucial. UK Biobank is a cohort study that incorporated data between 2006 and 2010 from over half a million individuals (40 to 69 years of age) at recruitment from across the United Kingdom. As one of the most accessible, largest, and in-depth cohort studies in the world, UK Biobank continues to enhance the resource with the addition of data from various omics platforms (eg, genomics, metabolomics, proteomics), multimodal imaging, self-reported risk factors and health outcomes, and linkage to electronic health records. The vision of UK Biobank is to allow as many researchers as possible to apply their expertise and imagination to undertake research to prevent, diagnose, and treat a wide range of chronic conditions, including CVD.
(J Am Coll Cardiol 2021;78:56–65) © 2021 Published by Elsevier on behalf of the American College of Cardiology Foundation.

Although there have been declines in cardiovascular disease (CVD) mortality, CVD remains a leading cause of premature death globally, a major contributor to disability, and a significant economic burden (1). The prevalence of several modifiable risk factors (such as obesity and type 2 diabetes) is increasing (2), and more people are now living with CVD, often with a high burden of medications and with concomitant CVD-related hospitalizations (3).

Since the last half of the 20th century, there has been great effort to define, identify, and modify CVD risk factors (eg, obesity, type 2 diabetes, hypertension, dyslipidemia, tobacco smoking, and physical inactivity) and develop therapies for those at risk of CVD. As a result, declines in age-adjusted cardiovascular mortality rates have been largely attributed to reductions in major risk factors and cardiological treatments (4,5). Despite this success, delivering optimal cardiovascular care to individuals is challenging. Individuals can have very heterogeneous CVD presentations and disease progression, and may respond very differently to interventions. Understanding the individual differences in CVD presentation, progression, and intervention response could further our diagnostic and prognostic capabilities, and support more effective targeting of current and future preventative and therapeutic options for CVD.

Largescale population-based prospective cohort studies can be used to study the etiology and progression of complex diseases such as CVD over an individual's lifetime. CVD is typically caused by a constellation of exposures that might each have small-to-moderate effects and interact with each other in complex ways. To examine a wide range of exposures, extensive information needs to be collected through questionnaires and physical measurements, as well as by measuring biomarkers

Listen to this manuscript's audio summary by Editor-in-Chief Dr. Valentin Fuster on JACC.org.

From the Big Data Institute, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom.
The authors attest they are in compliance with human studies committees and animal welfare regulations of the authors' institutions and Food and Drug Administration guidelines, including patient consent where appropriate. For more information, visit the Author Center.

Manuscript received February 17, 2021; accepted March 21, 2021.

ISSN 0735-1097/$36.00

https://doi.org/10.1016/j.jacc.2021.03.342

HIGHLIGHTS

- Continued advances in epidemiology are needed to improve the prevention and treatment of cardiovascular diseases.
- The UK Biobank prospective cohort is a resource with largescale and in-depth phenotyping and genomic data.
- Ongoing data collection includes whole-exome and whole-genome sequencing, metabolome and proteome measurements, MRI imaging, and health record linkages.
- The UK Biobank is easily accessible to bona fide researchers around the world.

from stored biological samples (eg, genetic or biochemical markers). Thousands of cases of a CVD (eg, myocardial infarction, ischemic stroke) may be required to study reliably the effects of different exposures with precision and power (6).

THE ORIGINAL GOALS OF THE STUDY GROUP

Funded primarily by the Medical Research Council and Wellcome Trust, UK Biobank is a large, prospective cohort study that was established to examine genetic and lifestyle risk factors for a variety of chronic diseases (including CVD). Between 2006 and 2010, approximately 9.2 million individuals 40 to 69 years of age who were living in England, Wales, and Scotland were invited to join the study, and 5.5% participated in the baseline assessment (7).

A broad range of phenotypic and biological data relevant to CVD epidemiology has been collected from UK Biobank participants (**Central Illustration**). Data collected at recruitment included self-reported lifestyle and medical information, physical measures (eg, blood pressure, anthropometry, spirometry) and biological samples (blood, urine, and saliva) (**Table 1**) (7). All of the data (including variable names, missingness, and summary statistics) can be viewed on UK Biobank's online Data Showcase. An R package (R Foundation for Statistical Computing, Vienna, Austria) to manage, query, and visualize UK Biobank data, as well as retrieve disease diagnoses and explore genetic metadata, has been developed by researchers and is freely available to those wishing to explore the extensive dataset in detail (8). UK Biobank continues to expand the resource by collecting extensive data directly from participants (**Table 2**). These include a series of web-based questionnaires sent to all participants with an e-mail address (N = 330,000) about particular exposures (eg, diet, occupation) and conditions that are not easily ascertained through linkage to medical records (eg, cognition, mental health, pain). UK Biobank has also collected objective physical activity data (on 100,000 participants), and is undertaking assessments of multimodal imaging (target of 100,000 participants) and cardiac monitoring (target of >20,000 participants). The size of the imaging enhancement study has necessitated the development of automated largescale image quality control, analytics, and image-based phenotype extraction pipelines (9). At least 10,000 of the imaged participants are in the process of being reinvited to undergo a complete repeat of the imaging enhancement in order to assess changes in imaging phenotypes over time, which is of particular interest for those interested in cognitive decline and other outcomes strongly associated with ageing. In addition, UK Biobank continues to increase its value by converting the information contained in the biological samples, which are limited and depletable, into data that can be widely shared. To date, this has included the measurement of a range of blood and urine biomarkers of interest for research into common conditions (including CVD) for all 500,000 participants (10), genotyping (11), and whole-exome and whole-genome sequencing, as well as the collection of other omics data, such as nuclear magnetic resonance metabolomics and proteomics. In 2017, UK Biobank released genome-wide genotyping array data (that measured 805,000 single-nucleotide polymorphisms) and which was imputed to more than 90 million variants for about 487,000 individuals, 94% of whom self-identified as being of White ethnicity. As a result, most analyses of genomic data from the resource have been restricted to those of White ethnicity (12).

As a longitudinal resource, 1 of UK Biobank's main aims is to follow the health of all participants through linkage to electronic health records, which are crucial to precision medicine in cardiology (13). Data from primary care records include a detailed record of activities relating to symptoms, diagnosis, prescriptions, investigations, and referrals. Such linkage has the potential to enable a much better understanding of the presentation of CVD when patients first visit their primary physician and of CVD's long-term management, because UK Biobank captures information that is not well-characterized in hospital records or other disease registries. However, linkage to primary care data is not without challenges. In 2019, UK Biobank made available primary care data

ABBREVIATIONS AND ACRONYMS

BMI = body mass index

CMR = cardiovascular magnetic resonance

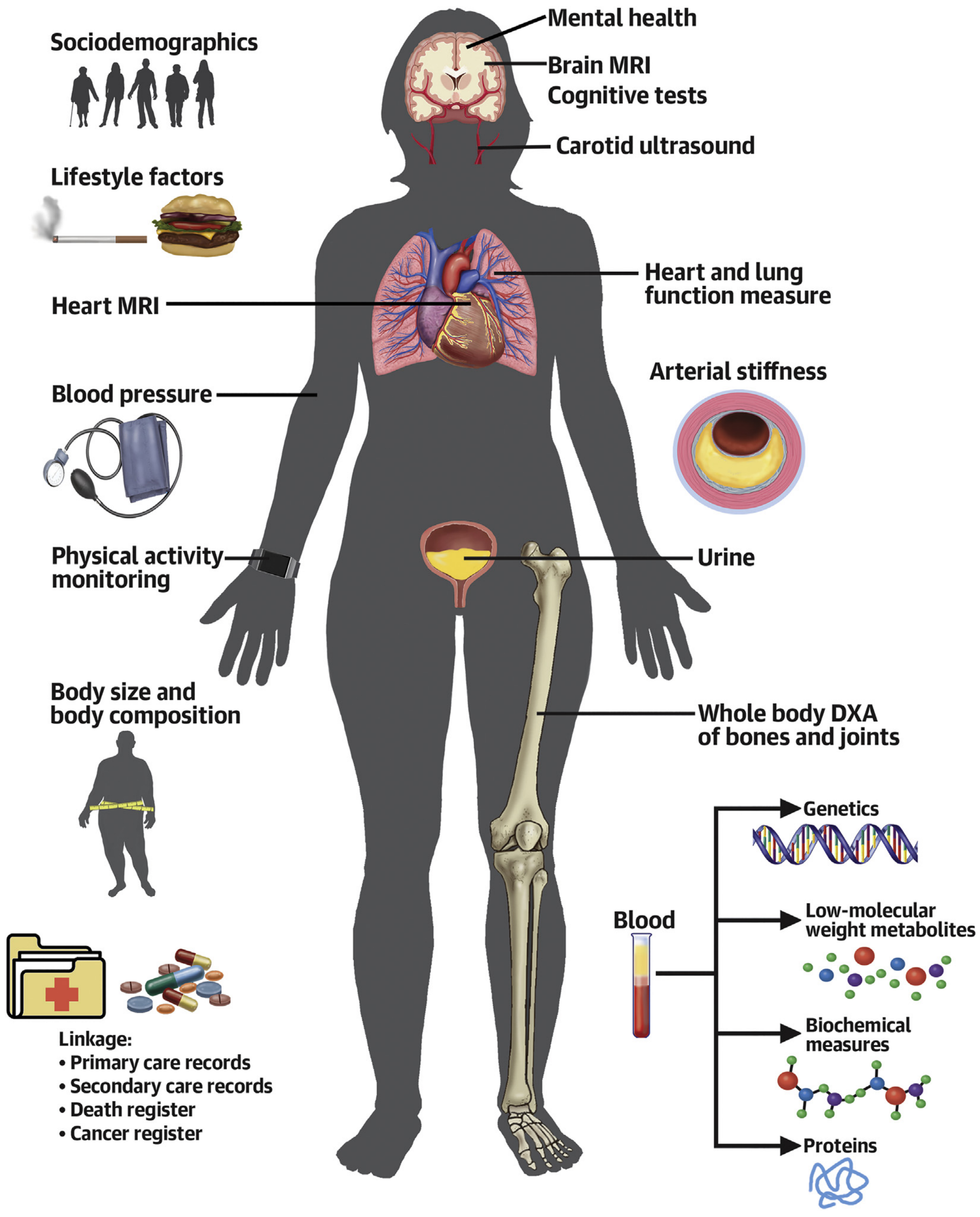
CVD = cardiovascular disease

GWAS = genome-wide association study

MRI = magnetic resonance imaging

PRS = polygenic risk score(s)

CENTRAL ILLUSTRATION UK Biobank Data Types



Caleyachetty, R. et al. J Am Coll Cardiol. 2021;78(1):56–65.

At recruitment, participants answered questions on sociodemographic, lifestyle, and health-related factors, and completed a range of physical measures. Blood, urine, and saliva samples were provided by participants, allowing many different types of assays to be performed (genetic, proteomic, and metabolomic analyses). Further enhancements were introduced, including brain, cardiac, and abdominal magnetic resonance imaging (MRI), dual-energy x-ray absorptiometry (DXA), and carotid ultrasound. Electronic health records linkage into the UK Biobank resource was used to expand the information available for clinical and public health research. Cardiovascular diseases are complex states, and therefore, understanding cardiovascular disease requires a collection of a broad range of phenotypic and biological data.

(received directly from the primary care electronic health record system suppliers) for approximately 260,000 participants (45% of the cohort), although further efforts to secure these data have been hampered by governance issues. Nonetheless, linkage of UK Biobank participants to hospital inpatient admissions to identify CVD outcomes has been shown to be both an accurate and comprehensive method, because it does not rely on participant feedback (7). To quickly support researchers in using the health outcome data, UK Biobank has started to generate research-ready health outcomes (including myocardial infarction and stroke) using rule-based phenotype algorithms. Researchers can also benefit from the progress that has been made by the HDR UK (Health Data Research UK) Phenomics platform, such as those generated by the open-access CALIBER (ClinicAl disease research using LInked Bespoke studies and Electronic health Records) Portal that has developed and validated a wide range of phenotypes (including those related to CVD) from national primary and secondary care records, and disease and mortality registries (14,15).

Although the UK Biobank is arguably 1 of the world's leading biomedical resource for research, the cohort sample is not representative of the U.K. population with regard to a number of sociodemographic, physical, lifestyle, and health-related characteristics (16). In particular, participants are typically less socioeconomically deprived and more health-conscious than the general population, and thus the data cannot be used to estimate population prevalence or incidence rates. A lack of representativeness (or selection bias) may limit the external validity (or generalizability) of some measures of association if the outcome of interest is linked to selection into the study (17). However, as long as there is broad heterogeneity across almost all exposures of interest (eg, socioeconomic deprivation, diet, physical activity), most findings will be broadly applicable to the population as a whole (18). This is particularly the case for genetic associations because genetic variants are unlikely to be associated with self-selection and are not generally associated with other lifestyle factors (19). The implications of selection bias for CVD epidemiology has been investigated in detail, whereby effect estimates for the associations of CVD risk factors with CVD mortality derived from UK Biobank data were compared with those derived from nationally sampled (ie, representative) cohort studies from England and Scotland. Reassuringly, they showed a high degree of concordance for CVD risk factor associations across both studies (20). Researchers may also find the use of statistical techniques such as weighting (21) can be helpful (under careful assumptions) to transpose UK Biobank effect estimates to match the target population.

TABLE 1 Data Collected at the Baseline Assessment

Measures	Details
Touchscreen questionnaire and computer-assisted verbal interview	
Sociodemographic	Ethnicity, education, employment, household information, Townsend deprivation index (socioeconomic status)
Lifestyle	Smoking; alcohol consumption; physical activity; diet; sleep
Environmental factors	Current address; current (or last) occupation; domestic heating and cooking fuel; housing; means of travel; shift work; mobile phone use; sun exposure
Early life factors	Birthplace, birth weight, breastfed, childhood body size and height, maternal smoking, handedness, adopted, and part of multiple birth
Family history	Illnesses of father/mother/siblings, age of parents, age parents died, and number of siblings
Psychosocial factors	Social support, bipolar/major depression, anxiety, nerves, psychological traits, and mood
Health and medical history	Medical conditions, medications, operations, cancer screening, pain, oral health, eyesight, hearing, and general health
Sex-specific factors	Male specific—first facial hair, age voice broke, hair/balding pattern, children fathered; female specific—hormone replacement therapy, contraception, pregnancy, menstruation, menopause, and cervical test
Cognitive function	Pairs matching; reaction time; prospective memory[a]; fluid intelligence[a]; numeric memory[b]
Hearing tests	Speech reception threshold[a]
Physical measures	
Blood pressure and heart rate	Two automated measures taken 1 min apart
Arterial stiffness[c]	Pulse wave velocity using infrared sensor at the finger
Grip strength	Right- and left-hand isometric grip strength
Anthropometrics	Standing/sitting height, waist/hip circumference, weight body mass index, and whole-body bioimpedance measures
Spirometry	Up to 3 measures within a 6-min period
Bone mineral density[d]	Calcaneal ultrasound
Eye measures[e]	Refractive index, intraocular pressure; acuity; retinal photograph; optical coherence tomography
Fitness test[e]	Cycle ergometry with electrocardiogram heart rate monitoring
Sample collection	
Blood	45 ml divided into 6 tubes, includes whole blood, serum, plasma, red cells, buffy coat
Urine	9 ml in 1 tube
Saliva[e]	2.5 ml in 1 tube

[a]Assessed in last 200,000 participants. [b]Assessed in 50,000 participants. [c]Measured in 170,000 participants. [d]Measured in 1 heel for 170,000 participants and in both heels for 320,000 participants. [e]Performed in the last 100,000 to 150,000 participants.

MOST IMPORTANT UK BIOBANK FINDINGS TO DATE

The availability of a very large and deeply characterized dataset that is readily accessible to bona fide researchers globally is already starting to make important contributions to health research.

TABLE 2 Current and Planned Future Data Available in the UK Biobank

Data Type	Details	Date of Data Acquisition	Data First Available
Genetic			
Genotype	Genome-wide genotyping was performed on all UK Biobank participants using the UK Biobank Axiom Array. Approximately 850,000 variants were directly measured, with >90 million variants imputed using the Haplotype Reference Consortium and UK10K + 1000 Genomes reference panels.	2013-2015	Q3 2017
Whole-exome sequencing	Exome sequencing for 50,000 participants was undertaken by Regeneron and GlaxoSmithKline. A further consortium (comprising Regeneron, AbbVie, Alnylam Pharmaceuticals, AstraZeneca, Biogen, Pfizer, Takeda and Bristol Myers Squibb) are undertaking exome sequencing on the remaining 450,000 participants.	2017-2021	Q1 2019 (50,000 participants) Q4 2020 (200,000 participants)
Whole-genome sequencing	The Medical Research Council provided funding for a pilot project (the Vanguard) to perform whole-genome sequencing on 50,000 participants, undertaken by the Wellcome Sanger Institute, Cambridge. A consortium of government (UK Research and Innovation [UKRI]), industry (Amgen, AstraZeneca, GlaxoSmithKline, and Johnson & Johnson) and charity (The Wellcome Trust) have funded whole-genome sequencing of the remaining 450,000 participants.	2020-	Expected Q3 2021 (200,000 participants)
Biomarkers			
Telomere length	Leucocyte telomere length measured in all 500,000 participants.	2015-2020	Expected Q1 2021
Biochemical measures	34 biomarkers assayed in the plasma, serum, red blood cells, and urine samples. Chosen based on their scientific relevance for studying a wide range of diseases, and included established risk factors for disease (eg, lipids for vascular disease, sex hormones for cancer), diagnostic measures (eg, HbA_{1c} for diabetes and rheumatoid factor for arthritis) or markers of phenotypes that were not otherwise well assessed (eg, renal and liver function).	2006-2010 2012-2013	Urinary biomarkers Q4 2016 Blood biomarkers Q1 2019
Plasma metabolites	Nightingale Health: NMR-metabolomics assay from blood samples collected at baseline assessment and at the first repeat assessment visit for all 500,000 participants. The platform measures over 200 metabolites, which will provide detailed data on circulating lipids, lipoprotein subclasses, fatty acid composition and various other low-molecular metabolites.	2020-	Expected Q1 2021 (120,000 participants)
Plasma proteins	Measurement of 1,500 plasma proteins using Olink's assay in 50,000 participants. Study funded by an industrial consortium including Amgen, AstraZeneca, Biogen, Bristol Myers Squibb, Genentech (a member of the Roche Group), GlaxoSmithKline (GSK), the Janssen Pharmaceutical Companies of Johnson & Johnson, Pfizer Inc, Regeneron and Takeda Pharmaceutical Co. Ltd.	2021-	Pending

Continued on the next page

In order to apply for access to data from UK Biobank, each applicant must demonstrate that they are a bona fide researcher (ie, they must register from, and be affiliated with, an approved research institute), and the application must involve health-related research that is in the public interest. All applicants are treated the same—whether academic, governmental, charitable, or commercial, or whether from domestic or international organizations—and all applications are assessed according to the same consistent criteria (22). As of September 2020, there have been approximately 1,500 publications, with at least 200 (16%) focused on CVD.

The UK Biobank is a uniquely powerful resource and offers unprecedented opportunities for new scientific discoveries in CVD. However, because the participants were recruited between 2006 and 2010, and a prospective cohort requires long-term follow-up of health outcomes, the most important findings from UK Biobank will start to emerge over the next 5 to 10 years. Nonetheless, the following examples highlight recent CVD research using UK Biobank that are likely to change clinical practice in the near future, and serve to illustrate the growing potential of this important resource.

Largescale cohort-wide genotyping and genotype imputation through the UK Biobank Axiom Array from Affymetrix (Thermo Fisher Scientific, Waltham, Massachusetts) has led to a noticeable shift in the use of genetic data in clinical research. Some genetic variants, mostly single nucleotide polymorphisms, are associated with an increased risk of coronary heart disease. These variants typically have small effects and correspond to a small proportion of variants that are causally associated with risk, meaning that they have limited predictive power (23). Alternative approaches that UK Biobank research has demonstrated include the utilization of data from a large number of variants across the genome to predict an individual's risk. There are several methods for combining risk variants, such as weighted polygenic risk scores (PRS), which are the summed effects of all the risk alleles for a trait in an individual (24).

Screening for common genetic variants could be a potentially useful tool for CVD disorders in the clinical setting. Khera et al. (25) developed a PRS based on a large number of genetic variants shown to be previously associated with coronary artery disease through recent genome-wide association studies (GWAS). When applied to the UK Biobank population,

TABLE 2 Continued

Data Type	Details	Date of Data Acquisition	Data First Available
Web-based questionnaires	Participants with e-mail addresses (330,000) are sent web-based questionnaires to collect detailed information on exposures or health outcomes that are difficult to capture through linkage to electronic health records.		
24-h dietary recall	Includes information on consumption of over 200 food and drink items over the last 24 h and was used to generate estimated daily nutrient intakes. The questionnaire was sent on 4 occasions over a 16-month period. 176 012 participants completed the questionnaire at least once (53% response rate) and 27 535 completed it 4 times (16%).	2011	Q3 2012
Cognitive function	Includes a series of cognitive tests, of which 4 were repeated from the baseline assessment (fluid intelligence, reaction time, numeric memory, pairs test) in addition to 2 further tests (trail making, symbol digit substitution). 120,800 participants completed (36% response rate).	2014	Q4 2015
Occupational history	Included information on lifetime employment history, occupational exposures and related medical information. ~121,300 participants completed (35% response rate).	2015	Q3 2015
Mental health	Included information on lifetime mental health events (including depression, bipolar affective disorder, and generalized anxiety disorder), alcohol and cannabis use, unusual and psychotic experiences, traumatic events, self-harm behaviors and subjective wellbeing. 157,400 participants completed (47% response rate).	2017	Q3 2017
Gastrointestinal health	Included information on gastrointestinal symptoms, and their effects on participants. Available for ~174,800 participants (53% response rate).	2017	Q3 2018
Food preferences	Included information on various food (and other) preferences for ~182,200 participants (55% response rate).	2019	Q1 2020
Pain	Included information on the causes of pain, severity and duration of chronic pain among participants for ~167,200 participants (50% response rate).	2019	Q1 2021
Health record linkage			
Death registrations	ICD-coded cause-specific mortality	2006–	Q1 2013 (England and Wales) Q4 2013 (Scotland)
Cancer registrations	ICD-coded cancer diagnoses	England 1971– Scotland 1957– Wales 1971–	Q1 2013 (England and Wales) Q4 2013 (Scotland)
Hospital inpatient episodes	ICD-coded diagnoses, OPCS-coded procedures. Critical care data for participants in England is available from April 2011 (a small number of records exist before this date).	England 1997– Scotland 1977– Wales 1999–	2013 2020
Primary care	Read-coded information including diagnoses, measurements, referrals, prescriptions. Available for a subset of the cohort (~230,000 participants).	England 1938– Scotland 1939– Wales 1948–	Q3 2019
Repeat of baseline assessment	20,000 participants repeated all baseline assessment measures at one assessment center, Stockport, United Kingdom.	2012–2013	Q4 2013
Multimodal imaging	MRI of brain, heart and body, carotid ultrasound and whole-body DXA scan of bones and joints (50,000 out of 100,000 collected). Repeat imaging among 10,000 participants funded by Dementia Platform UK.	2014– Repeat imaging visit 2019–	2014–
Cardiac monitor	A subset of participants will be asked to wear a cardiac monitor for 2 weeks to investigate the impact of heart rhythm disturbances.	Ongoing	Pending
Accelerometry	100,000 participants wore an Axivity AX3 triaxial wrist accelerometer for a 7-day period. Derived summary data on duration and intensity of activity are available. Repeat measurements in ~2,500 participants 4 times over a year.	2013–2016 Repeat quarterly assessment 2018	Q1 2016 2019

DXA = dual-energy x-ray absorptiometry; HbA_{1c} = glycosylated hemoglobin; ICD = International Classification of Diseases; MRI = magnetic resonance imaging; NMR = nuclear magnetic resonance; OPCS = Office of Population Censuses and Surveys Classification of Interventions and Procedures; Q = quarter.

they reported that the PRS could identify 8% of people of European ancestry that had at least a 3-fold increased likelihood for prevalent coronary artery disease. The usefulness for identifying those at increased risk in their current form is modest, and this result poses interesting questions about the utility of this information in screening those at highest risk for early interventions or to optimize existing screening programs. Although incorporating genetic risk scores into clinical practice may aid risk stratification, several scientific and ethical challenges remain in defining the role of PRS in health care. For example, the use of a PRS for coronary artery disease compared with existing pooled cohort equations was only associated with modest improvement in the predictive accuracy for incident coronary artery disease and was shown to improve risk stratification for only a small proportion of individuals (26). An important ethical concern is that, so far, PRS have largely been derived from European DNA sequences, and hence, polygenic risk prediction in non-European populations may have reduced prediction accuracy.

Individuals carrying certain genetic variants for CVD may benefit from the knowledge that lifestyle choices can influence their future risk of disease (27). The combination of detailed genomic and phenotypic data at scale makes UK Biobank unique in being able to examine such questions. Tikkanen et al. (28) reported that among UK Biobank participants with an increased genetic risk of CVD, physical activity as measured either with an objective activity monitor (Axivity AX3 wrist-worn triaxial accelerometer, Newcastle upon Tyne, United Kingdom) or via a subjective questionnaire (short-form IPAQ [International Physical Activity Questionnaire] questionnaire) was associated with a lower risk of coronary heart disease and atrial fibrillation across strata for CVD genetic risk. However, the associations for questionnaire-based physical activity were notably more modest than those obtained using objectively measured physical activity. Overall, the study findings indicated that increased genetic risk of coronary heart disease and atrial fibrillation could be mitigated to some extent by increased physical activity (28). In another example, Rutten-Jacobs et al. (29) evaluated the associations of a PRS and healthy lifestyle (no current smoking, healthy diet, body mass index [BMI] <30 kg/m^2, and moderate physical activity 2 or more times weekly) with incident stroke and found that an unfavorable lifestyle was associated with an increased risk of stroke regardless of their genetic risk score (29).

Genetic variants can also be used as instruments for strengthening causal inference in observational cardiovascular epidemiology studies (30). Considering their fixed nature and Mendel's first and second laws of inheritance, the technique of using genetic variants for appraising causality (Mendelian randomization) minimizes the susceptibility of reverse causation bias and confounding (31). This approach has been used to determine whether BMI is causally associated with a range of cardiovascular conditions. This remains an important question because of the debate about an apparent "obesity paradox," in which a higher BMI has been shown, in some studies, to be associated with improved survival for those with CVD (32). For example, Larsson et al. (33) found that genetically predicted higher BMI (and particularly fat mass index) was significantly positively associated with 8 cardiovascular conditions (ie, aortic valve stenosis, heart failure, deep vein thrombosis, arterial hypertension, peripheral artery disease, coronary artery disease, atrial fibrillation, and pulmonary embolism). These findings therefore provide strong support that higher body fat is likely to be associated with increased risk of most cardiovascular outcomes. Although Mendelian randomization is a useful tool to assess the likely causality of risk factors on health outcomes in observational epidemiology, the approach is subject to various assumptions and has some limitations that researchers need to be aware of when interpreting the likely causal nature of the relationship (31).

With the increasing availability of high-density genotypic information in the UK Biobank, understanding genotype-phenotype associations will become more dependent on the availability of high-quality phenotypic information. In this context, Aragam et al. (34), using UK Biobank's extensive phenotypic and genotypic data, conducted a GWAS of heart failure and further examined whether refined phenotypic classification of specific heart failure subpopulations would support detection of novel genetic loci that reflect distinct etiologic heart failure subtypes. Heart failure phenotypes in UK Biobank were defined using a combination of self-reported questionnaire data (confirmed by a trained health care professional) and linked hospital admission and death registry data. Their GWAS of heart failure ($N = 7{,}382$) yielded multiple known loci for known heart failure risk factors (such as coronary artery disease and atrial fibrillation). However, when the analysis was restricted to the subset of those with "nonischemic cardiomyopathy" ($n = 2{,}038$), they found strong genetic signals at loci associated with dilated cardiomyopathy that were independent of known heart failure risk factors and associated with intermediate traits of left ventricular structure and function in individuals without clinical heart failure. Such phenotypic refinement can support the discovery of novel genetic signals that reflect distinct etiologic heart failure subtypes, creating a unique opportunity possibly to improve heart failure care.

Cardiovascular magnetic resonance (CMR) imaging is becoming increasingly important in cardiovascular medicine (35), and is a key component of an ongoing multimodality imaging study in the UK Biobank, the largest in the world that aims to collect imaging scan data on 100,000 participants. This includes CMR imaging data, as well as brain and body magnetic resonance imaging (MRI), dual-energy X-ray absorptiometry, and carotid ultrasound (36). Repeat imaging on at least 10,000 participants began in 2019, offering the opportunity for researchers to examine possible changes in imaging phenotypes over time. Complementary to this, a repeat imaging study to assess specifically the effects of SARS-CoV-2 on internal organs, including its possible effects on the cardiovascular system, is also underway.

The analysis and interpretation of cardiac structural and functional indices from the MRI scans can help reveal insights into subclinical cardiovascular mechanisms related to a wide range of CVDs (36,37). The cardiac MRI scan is performed using a clinical wide-bore 1.5-T scanner (MAGNETOM Aera, Syngo Platform VD13A, Siemens Healthcare, Erlangen, Germany). Only a limited range of features are automatically extracted (such as left ventricular ejection fraction and cardiac output); however, efforts are underway to develop automated processing tools that can extract a broader range of cardiac phenotypes (38). The UK Biobank's CMR dataset has already become a well-established reference dataset and is being used in clinical research (39,40). The combination of the world's largest CMR imaging dataset with extensive phenotypic and genetic data offers an unprecedented resource for cardiovascular research. For example, using CMR imaging data on 36,000 UK Biobank participants, Pirruccello et al. (41) identified 45 novel genetic loci associated with left ventricular structure and function, including many known genes associated with Mendelian cardiomyopathies. A polygenic score of MRI-derived left ventricular end-systolic volume was strongly associated with incident dilated cardiomyopathy in UK Biobank participants. These results further implicate common genetic polymorphisms in the pathogenesis of dilated cardiomyopathy (41).

It is well-established that genetic variation in patients can influence their response to drug therapy (42). As such, identifying which patients will respond best to certain drugs before treatment starts will lead to better treatment effects and reduced adverse events. In the largest pharmacogenetics study to date, McInnes et al. (43) used genotype data from 487,409 participants in UK Biobank to analyze pharmacogenetic variation in 14 clinically important genes at a population scale. They found that all of participants have at least 1 clinically relevant pharmacogenetic variant, with an average of 4 actionable pharmacogenetic variants, leading to an average of 12 drugs that require an alternate drug or dosage according to CPIC (Clinical Pharmacogenetics Implementation Consortium) guidelines.

FUTURE RESEARCH DIRECTIONS

UK Biobank is 1 of the largest and most comprehensive population cohort studies globally. With easy access to deep phenotypic, genomic, imaging, and health outcomes over years to come, UK Biobank provides unique opportunities for cardiovascular research. Encouraging further high-quality research in CVD necessitates that the UK Biobank resource is periodically enhanced if it is to continue to drive innovative cardiovascular research that will impact clinical practice and public health more directly (44,45). Given this background, the UK Biobank has identified promising future research directions.

The need for clinicians to diagnose and treat individuals with familial forms of CVD is important and, in some areas of CVD, the underlying genes remain largely unknown (46). Although array data—and the subsequent imputed data—capture the spectrum of common genetic variants, rare variation that is more likely to modify protein sequences and have large phenotypic consequences is less well captured through this approach (46). Using next-generation sequencing, a research consortium (**Table 2**) is undertaking a project to sequence the exomes of all 500,000 UK Biobank participants, with data for 200,000 participants made available for other researchers to use made in October 2020 (and the remainder being made available in 2021). Exons, the parts of our DNA that code for protein, are collectively known as the exome. Whole-exome sequencing allows the direct assessment of protein-altering variants, and the potential for a more granular understanding of the genetic architecture of inherited CVDs (47). Fahed et al. (48) performed whole-exome sequencing data on UK Biobank participants with and without coronary artery disease, and showed that PRS modified the penetrance of monogenic risk variants for familial hypercholesterolemia, with the probability of disease by age 75 years ranging from 17% to 78% for coronary artery disease. Insight from this study could inform decisions about the timing and intensity of lipid-lowering therapy for individuals with familial hypercholesterolemia (48). Whole-exome sequencing has also been reported to be a valuable screening tool in establishing the clinical diagnosis of poorly defined cases of sudden cardiac death (46). In addition to whole-exome sequencing, UK Biobank is also undertaking whole-genome sequencing for the entire cohort (**Table 2**). Metabolomics has emerged as a means for measuring metabolites (chemical intermediates), which not only includes molecules from multiple metabolic pathway, but also provides a functional integration of upstream genetic, transcriptomic, and proteomic variation, as well as environmental exposures, thereby reflecting molecular processes more proximal to CVD (49). Nightingale Health, a Finnish biomedical science company, is using their nuclear magnetic resonance-based metabolomics platform to analyze the plasma samples of all 500,000 UK Biobank participants (including 20,000 participants who attended a repeat assessment 4 to 5 years after their baseline samples was taken) for over 200 metabolic biomarkers related to CVD, type 2 diabetes, and other common chronic diseases. In a subsample, these metabolite measurements will be repeated. This detailed metabolic profiling of UK Biobank participants can provide an enhanced view of the complex molecular mechanisms underpinning the onset and progression of CVD. Integration of UK Biobank genomics and metabolomics data may also help assess the contribution of genetic variation to circulating plasma metabolite concentrations.

CONCLUSIONS

Our understanding of cardiovascular disease is being transformed as a result of the availability of large amounts of in-depth genetic (as well as other omics) and health information from half a million UK participants. UK Biobank is globally accessible to approved researchers undertaking vital health research and provides an unparalleled opportunity to enable scientific discoveries that improve the prevention, diagnosis, and treatment of CVD. Of course, several challenges exist. Largescale epidemiological studies, such as UK Biobank, continually need to robustly manage, document, and make available large, complex data in an accessible manner. Maximizing the use of these rich and complex data requires advances in data visualization and analytical methods, and a desire to engage in collaborative science. These challenges are not trivial, but as 1 of the most accessible, in-depth, largest, and adaptable biomedical data resources in the world, UK Biobank is well positioned to leverage advances in scientific tools and technologies to enable scientific discoveries that benefit human health.

ACKNOWLEDGMENTS The authors acknowledge the members of the UK Biobank Steering Committee. Additional thanks to the UK Biobank Access team for their tireless work on research registrations, applications and output and providing this information for the article. The authors are very grateful to the participants of the UK Biobank for their continuing interest and involvement in the study.

FUNDING SUPPORT AND AUTHOR DISCLOSURES

The authors have reported that they have no relationships relevant to the contents of this paper to disclose.

ADDRESS FOR CORRESPONDENCE: Dr. Ben Lacey, UK Biobank Epidemiology Group, Big Data Institute, Nuffield Department of Medicine, University of Oxford, Old Road Campus, Oxford OX3 7LF, United Kingdom. E-mail: ben.lacey@ndph.ox.ac.uk.

REFERENCES

1. Roth GA, Mensah GA, Johnson CO, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the GBD 2019 study. J Am Coll Cardiol 2020;76:2982–3021.

2. Timmis A, Townsend N, Gale CP, et al. European Society of Cardiology: cardiovascular disease statistics 2019. Eur Heart J 2020;41:12–85.

3. Bhatnagar P, Wickramasinghe K, Wilkins E, Townsend N. Trends in the epidemiology of cardiovascular disease in the UK. Heart 2016;102:1945–52.

4. Unal B, Critchley JA, Capewell S. Explaining the decline in coronary heart disease mortality in England and Wales between 1981 and 2000. Circulation 2004;109:1101–7.

5. Capewell S, Morrison CE, McMurray JJ. Contribution of modern cardiovascular treatment and risk factor changes to the decline in coronary heart disease mortality in Scotland between 1975 and 1994. Heart 1999;81:380.

6. Collins R. What makes UK Biobank special. Lancet 2012;379:1173–4.

7. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 2015;12:e1001779.

8. Hanscombe KB, Coleman JRI, Traylor M, Lewis CM. ukbtools: an R package to manage and query UK Biobank data. PLoS One 2019;14:e0214311.

9. Littlejohns TJ, Sudlow C, Allen NE, Collins R. UK Biobank: opportunities for cardiovascular research. Eur Heart J 2019;40:1158–66.

10. Allen NE, Arnold M, Parish S, et al. Approaches to minimising the epidemiological impact of sources of systematic and random variation that may affect biochemistry assay data in UK Biobank. Wellcome Open Res 2020;5:222.

11. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 2018;562:203–9.

12. Mendis S. The contribution of the Framingham Heart Study to the prevention of cardiovascular disease: a global perspective. Prog Cardiovasc Dis 2010;53:10–4.

13. Antman EM, Loscalzo J. Precision medicine in cardiology. Nat Rev Cardiol 2016;13:591–602.

14. Denaxas S, Gonzalez-Izquierdo A, Direk K, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J Am Med Inform Assoc 2019;26:1545–59.

15. Kuan V, Denaxas S, Gonzalez-Izquierdo A, et al. A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. Lancet Digit Health 2019;1:e63–77.

16. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. Am J Epidemiol 2017;186:1026–34.

17. Munafò MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. Int J Epidemiol 2018;47:226–35.

18. Rothman K, Hatch E, Gallacher J. Representativeness is not helpful in studying heterogeneity of effects across subgroups. Int J Epidemiol 2014;43:633–4.

19. Ebrahim S, Davey Smith G. Commentary: should we always deliberately be non-representative. Int J Epidemiol 2013;42:1022–6.

20. Batty GD, Gale CR, Kivimäki M, Deary IJ, Bell S. Comparison of risk factor associations in UK Biobank against representative, general population based studies with conventional response rates: prospective cohort study and individual participant meta-analysis. BMJ 2020;368:m131.

21. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. Am J Epidemiol 2017;186:1010–4.

22. Conroy M, Sellors J, Effingham M, et al. The advantages of UK Biobank's open-access strategy for health research. J Intern Med 2019;286:389–97.

23. Choi SW, Mak TS-H, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc 2020;15:2759–72.

24. Rotter JI, Lin HJ. An outbreak of polygenic scores for coronary artery disease. J Am Coll Cardiol 2020;75:2781.

25. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet 2018;50:1219–24.

26. Elliott J, Bodinier B, Bond TA, et al. Predictive accuracy of a polygenic risk score-enhanced prediction model vs a clinical risk score for coronary artery disease. JAMA 2020;323:636–45.

27. Horne J, Madill J, O'Connor C, Shelley J, Gilliland J. A systematic review of genetic testing and lifestyle behaviour change: are we using high-quality genetic interventions and considering behaviour change theory. Lifestyle Genom 2018;11:49–63.

28. Tikkanen E, Gustafsson S, Ingelsson E. Associations of fitness, physical activity, strength, and genetic risk with cardiovascular disease: longitudinal analyses in the UK Biobank study. Circulation 2018;137:2583–91.

29. Rutten-Jacobs LC, Larsson SC, Malik R, et al. Genetic risk, incident stroke, and the benefits of adhering to a healthy lifestyle: cohort study of 306 473 UK Biobank participants. BMJ 2018;363:k4168.

30. Bennett DA, Holmes MV. Mendelian randomisation in cardiovascular research: an introduction for clinicians. Heart 2017;103:1400–7.

31. Haycock PC, Burgess S, Wade KH, Bowden J, Relton C, Davey Smith G. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. Am J Clin Nutr 2016;103:965–78.

32. Curtis JP, Selter JG, Wang Y, et al. The obesity paradox: body mass index and outcomes in patients with heart failure. Arch Intern Med 2005;165:55–61.

33. Larsson SC, Bäck M, Rees JMB, Mason AM, Burgess S. Body mass index and body composition in relation to 14 cardiovascular conditions in UK Biobank: a Mendelian randomization study. Eur Heart J 2020;41:221–6.

34. 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.

35. von Knobelsdorff-Brenkenhoff F, Schulz-Menger J. Role of cardiovascular magnetic resonance in the guidelines of the European Society of Cardiology. J Cardiovasc Magn Reson 2016;18:6.

36. Petersen SE, Matthews PM, Bamberg F, et al. Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK Biobank - rationale, challenges and approaches. J Cardiovasc Magn Reson 2013;15:46.

37. Tarroni G, Bai W, Oktay O, et al. Large-scale quality control of cardiac imaging in population studies: application to UK Biobank. Sci Rep 2020;10:2408.

38. Littlejohns TJ, Holliday J, Gibson LM, et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nat Commun 2020;11:2624.

39. Aung N, Sanghvi MM, Zemrak F, et al. Association between ambient air pollution and cardiac morpho-functional phenotypes: insights from the UK Biobank population imaging study. Circulation 2018;138:2175–86.

40. Petersen SE, Sanghvi MM, Aung N, et al. The impact of cardiovascular risk factors on cardiac structure and function: Insights from the UK Biobank imaging enhancement study. PLoS One 2017;12:e0185114.

41. Pirruccello JP, Bick A, Wang M, et al. Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy. Nat Commun 2020;11:2254.

42. Blakey JD, Hall IP. Current progress in pharmacogenetics. Br J Clin Pharmacol 2011;71:824–31.

43. McInnes G, Lavertu A, Sangkuhl K, Klein TE, Whirl-Carrillo M, Altman RB. Pharmacogenetics at scale: an analysis of the UK Biobank. Clin Pharmacol Ther 2021;109:1528–37.

44. Khoury MJ, Gwinn M, Ioannidis JP. The emergence of translational epidemiology: from scientific discovery to population health impact. Am J Epidemiol 2010;172:517–24.

45. Keyes K, Galea S. What matters most: quantifying an epidemiology of consequence. Ann Epidemiol 2015;25:305–11.

46. Seidelmann SB, Smith E, Subrahmanyan L, et al. Application of whole exome sequencing in the clinical diagnosis and management of inherited cardiovascular diseases in adults. Circ Cardiovasc Genet 2017;10:e001573.

47. Van Hout CV, Tachmazidou I, Backman JD, et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature 2020;586:749–56.

48. Fahed AC, Wang M, Homburger JR, et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. Nat Commun 2020;11:3635.

49. McGarrah RW, Crown SB, Zhang GF, Shah SH, Newgard CB. Cardiovascular metabolomics. Circ Res 2018;122:1238–58.

KEY WORDS biobanks, cardiovascular diseases, cohort