BIOBANKS

Prospective study design and data analysis in UK Biobank

Naomi E. Allen[1,2]*, Ben Lacey[1,2], Deborah A. Lawlor[3,4], Jill P. Pell[5], John Gallacher[6,7], Liam Smeeth[8], Paul Elliott[9,10], Paul M. Matthews[12], Ronan A. Lyons[13], Anthony D. Whetton[14], Anneke Lucassen[15,16], Matthew E. Hurles[17], Michael Chapman[18], Andrew W. Roddam[19], Natalie K. Fitzpatrick[20], Anna L. Hansell[21], Rebecca Hardy[22], Riccardo E. Marioni[23], Valerie B. O'Donnell[24], Julie Williams[25], Cecilia M. Lindgren[26], Mark Effingham[1], Jonathan Sellors[1], John Danesh[27,28,29], Rory Collins[1,2]

Copyright © 2024 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works

Population-based prospective studies, such as UK Biobank, are valuable for generating and testing hypotheses about the potential causes of human disease. We describe how UK Biobank's study design, data access policies, and approaches to statistical analysis can help to minimize error and improve the interpretability of research findings, with implications for other population-based prospective studies being established worldwide.

INTRODUCTION

Population health research has come a long way in the past few decades, with major advances in our understanding of the causes of disease. In particular, prospective studies that were initiated in the 1950s, such as the British Doctors Study (*1*) and the Framingham Heart Study (*2*), have been invaluable for understanding the association between lifestyle factors and disease risk, because they overcome many of the biases inherent in case-control studies, most notably that risk factors for disease (exposures) are measured before disease onset. However, until recently, the conclusions that could be drawn from such studies were limited by small sample size, varying analytical approaches for defining various risk factors, and the relatively short duration of follow-up time to assess health outcomes. It was not until data from these different studies were integrated into large-scale individual-level meta-analyses that associations of exposures with disease risk were identified robustly. For example, it is now well established that circulating lipids and high blood pressure are causally related to vascular disease (*3*), adiposity with cardiovascular disease (*4*), menopausal hormone therapy use and alcohol consumption with breast cancer (*5*, *6*), and oral contraceptive use with a reduced risk of ovarian cancer (*7*).

More recently, there has been remarkable progress in research on the genetic determinants of disease. In the early 2000s, the literature was dominated by a plethora of genetic studies that focused on associations with particular conditions within specific "candidate" genes that were of a priori interest. Many of these studies involved small numbers of disease cases and yielded false-positive results that failed to replicate, often because of undue emphasis on post hoc selective reporting of the more extreme associations that were observed. Subsequently, improvements in assay technology led to genome-wide association studies that allowed hypothesis-free identification across the genome of variants associated with a particular phenotype. Much effort was typically spent on characterizing the phenotype under investigation precisely in the belief that outcome misclassification would have a substantial impact on the ability to detect associations. However, when meta-analyses of different studies were performed, which yielded much larger numbers of individuals with the outcome of interest (albeit differently defined), small-to-moderate associations between genetic variants and outcomes began to be identified reproducibly after stringent adjustment for multiple testing (*8*).

Even larger sample sizes—of the order of hundreds of thousands of participants—are needed to study gene-environment interactions, especially where the genetic variant or environmental exposure of interest is rare or has a small effect on disease risk (*9*). Consequently, there is a strategic need to establish large-scale, well-characterized, population-based prospective cohorts in which biological samples are collected and health outcomes are followed long-term to facilitate research into the determinants of disease.

UK Biobank combines scale, depth, duration, and accessibility

UK Biobank is a population-based prospective cohort of 500,000 men and women designed to enable research into the genetic, lifestyle,

[1]UK Biobank Ltd, Stockport, UK. [2]Nuffield Department of Population Health, University of Oxford, Oxford, UK. [3]Population Health Science, Bristol Medical School University of Bristol, Bristol, UK. [4]Medical Research Council Integrative Epidemiology Unit at the University of Bristol, Bristol, UK. [5]School of Health and Wellbeing, University of Glasgow, Scotland. [6]Department of Psychiatry, University of Oxford, Oxford, UK. [7]Dementias Platform UK, Department of Psychiatry, University of Oxford, Oxford, UK. [8]London School of Hygiene and Tropical Medicine, London, UK. [9]MRC Centre for Environment and Health, School of Public Health, Imperial College London, London, UK. [10]NIHR Health Protection Research Unit in Chemical Radiation Threats and Hazards, Imperial College London, UK. [11]Health Data Research UK, Imperial College London, London, UK. [12]UK Dementia Research Centre Institute and Department of Brain Sciences, Imperial College London, London, UK. [13]Population Data Science, Swansea University Medical School, Swansea, Wales. [14]Veterinary Health Innovation Engine, University of Surrey, Guildford, UK. [15]Nuffield Department of Medicine, University of Oxford, Oxford, UK. [16]Faculty of Medicine, Southampton University, Southampton, UK. [17]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge, UK. [18]NHS England, London, UK. [19]Our Future Health, London, UK. [20]Institute of Health Informatics, University College London, London, UK. [21]Centre for Environmental Health and Sustainability, University of Leicester, Leicester, UK. [22]School of Sport, Exercise and Health Sciences, Loughborough University, Loughborough, UK. [23]Centre for Genomic and Experimental Medicine, University of Edinburgh, Edinburgh, Scotland. [24]School of Medicine, Cardiff University, Cardiff, Wales. [25]UK Dementia Research Institute, Cardiff University, Cardiff, Wales. [26]Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK. [27]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [28]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK. [29]National Institute for Health Research Cambridge Biomedical Research Centre, University of Cambridge and Cambridge University Hospitals, Cambridge, UK.
*Corresponding author Email: naomi.allen@ndph.ox.ac.uk

Downloaded from https://www.science.org at University of Connecticut on January 10, 2024

and environmental determinants of a wide range of diseases of middle-to-old age (https://ukbiobank.ac.uk). It was established by the UK Medical Research Council (MRC) and Wellcome, which continue to fund it along with the British Heart Foundation, Cancer Research UK (CRUK), and the National Institute for Health and Care Research. The key design features are its easy accessibility, large-scale prospective nature, depth and range of risk factor data, and comprehensive linkage to health outcomes, which together enable academic and industry researchers worldwide to perform discovery science (table S1).

UK Biobank was designed to promote innovative science by maximizing access to the data in an equitable and transparent manner. All approved researchers (academic or commercial) can access all of the de-identified data to perform any type of health-related research that is in the public interest. This is the key criterion against which applications to access the data are considered, with restrictions only placed on their use for potentially contentious research (e.g., investigations that could lead to racial or sexual discrimination). Access to biological samples is currently largely restricted to assays that will be conducted on the whole, or large representative samples, of the cohort.

Ready access to such a large-scale, in-depth resource has encouraged researchers from many disciplines across academia and industry to collaborate to ensure that different types of complex data, e.g., whole-exome and whole-genome sequencing data, magnetic resonance imaging (MRI) scans, accelerometer waveform data, and electronic health records, are generated and analyzed appropriately. The ready accessibility of the data at low cost without requiring collaboration with, or peer review from, the UK Biobank study investigators has led to an exponential increase in research output. By the end of 2023, there were more than 38,000 registered researchers (84% from outside the United Kingdom) and about 10,000 published papers (attracting 2.5 million citations), with the number of publications increasing exponentially each year. In particular, the release to the worldwide research community of cohort-wide genome-wide genotyping and imputation data in 2017 has been hugely influential in advancing our understanding of the genetic determinants of disease.

The requirement that researchers publish their findings and make available any derived variables that have been generated as part of their research, together with the underlying code that generated the research output, enables the wider scientific community to critique, modify, and build upon the work of others in a transparent manner (*10*). For example, research groups with expertise in signal processing have created derived variables related to the intensity and duration of physical activity from the raw accelerometer data (*11*, *12*). Similarly, academic and commercial research groups with expertise in image analysis have made available variables derived from the MRI scans related to body fat distribution (*13*), fat and iron content of specific organs (*14*, *15*), as well as metrics of the structure and function of the brain (*16*) and heart (*17*). In this way, complex data that might otherwise only be of use to specialists in a narrow field of research are turned into well-curated derived variables that are integrated with other UK Biobank data and can be used extensively by nonspecialists to answer a range of research questions.

Easy access to such a wealth of data has led to new ways of presenting results. For example, summary statistics of all of the associations of individual genetic variants (*18*, *19*) and polygenic risk scores (*20*) with a wide range of phenotypes are now available via online browsers. This move toward the publication of all summary results rather than publication of particular results in traditional scientific journals (where cherry-picking the most "interesting" associations may introduce bias) is likely to accelerate scientific discovery and provide easier replication of associations across different studies. To help democratize access further, UK Biobank launched a cloud-based Research Analysis Platform in 2021 that allows streamlined access for researchers worldwide (particularly to the genome sequence data that are too large to transfer to researchers), as well as free computing and data storage for researchers from low- and middle-income countries and for early career researchers.

One consequence of researchers with different expertise accessing this wealth of data is the potential for unfamiliarity with various types of biases that are inherent in prospective studies that might influence results and with the complexities associated with data that are outside of their areas of expertise. All researchers accessing biomedical resources to study the determinants of disease need to be aware of small sample size (that may produce imprecise estimates due to random error), incomplete or inadequate measurement of risk factors (that may lead to systematic underestimation of disease associations), and health outcomes (that may lead to more imprecise estimates) and their potential confounding factors (that may obscure or lead to spurious associations between exposures and outcomes). Insufficient duration of follow-up may also lead to reverse causation bias, whereby the disease process influences potential risk factors (particularly nongenetic ones), especially for conditions with a long prodromal phase, such as Alzheimer's disease.

UK Biobank has been set up to help minimize random and systematic error so that it can support reliable research into the determinants of disease (table S1), although the general principles of careful study design and appropriate data analysis apply equally to all large-scale, prospective studies. There are a number of trade-offs that need to be considered when designing a cohort study, which relate to the size and heterogeneity of the study population, as well as to the methods used for its recruitment, data collection, and follow-up. UK Biobank has aimed to generate a large-scale, prospective biomedical resource that includes a wide range of exposure and health outcome measures collected as accurately as possible, with easy accessibility to the data. However, as with all prospective studies, it is important to consider, and if possible correct for, potential biases arising from the study design and collection of data.

The importance of a large-scale prospective design

UK Biobank recruited 502,000 volunteers aged 40 to 69 years between 2006 and 2010 from across England, Wales, and Scotland. This age group was selected to include individuals who were young enough that relatively few would have developed health conditions at the time of recruitment. As a prospective study, UK Biobank has many advantages for investigating the effects of genetic, lifestyle, and environmental factors on disease outcomes (*21*). In particular, information on exposures to potential risk factors can be assessed before disease develops, which avoids bias caused by differential recall of information about past exposures depending on an individual's outcome status (recall bias). The prospective design also allows investigation of factors that might be affected by disease processes or their treatments or by changes in an individual's behavior after the development of some condition (reverse causation bias). In addition, it can support studies of conditions that cannot readily be investigated retrospectively (e.g., fatal illnesses). Furthermore, by allowing

Downloaded from https://www.science.org at University of Connecticut on January 10, 2024

a wide range of different conditions to be studied within the same study population, the full effects of a particular exposure on all aspects of health can be better assessed (e.g., smoking on a wide range of different diseases). Likewise, the effects of many different exposures on a single disease can be determined, provided that sufficient numbers of cases have occurred to allow the separate and combined effects of exposures to be assessed reliably.

Prospective studies need to be large, because only a relatively small proportion of the participants will develop any given condition during follow-up. The rationale for recruiting 500,000 adults into UK Biobank was that it would enable large numbers of cases of the most common diseases to develop within a reasonable follow-up period (while also allowing detailed exposure information to be collected within funding and organizational constraints). For example, after a median follow-up of 12 years (i.e., by the end of 2020), linkage to electronic health care record data indicated that there had been at least 30,000 incident cases of diabetes, 25,000 cases of depression, 15,000 cases of myocardial infarction, and 9000 cases of breast cancer (Table 1). For the reliable detection of risk ratios of about 1.3 for the main effects of different exposures (ranging from those that are dichotomous variables to those that are continuous measures), about 5000 to 10,000 incident cases of a particular disease would be required (*22*). The need for a large sample size is even more evident when assessing combined effects. For example, when estimating the combined effect of blood pressure and age on the risk of coronary heart disease, the SEs of the estimates (and hence the 95% confidence intervals) are, on average, three times narrower with 500,000 versus 50,000 participants (*23*). As the UK Biobank participants age, the number of incident cases of different diseases is increasing substantially, allowing a wider range of outcomes to be investigated more completely. For example, by 2032, there will be more than 60,000 cases of diabetes and chronic obstructive pulmonary disease (COPD). The sheer size of the study also means that robust research into less common conditions will also be possible. For example, between 2020 and 2027, the number of cases of Alzheimer's disease, hip fracture, and Parkinson's disease is expected to more than double to about 17,000, 13,000, and 10,000, respectively (Table 1).

Comparing cohort characteristics with those of the wider population

In UK Biobank, the well-defined sampling frame means that it is possible to assess not only the overall participation rate but also the extent to which the study population differs from the wider population from which it was drawn. Postal invitations were sent to 9.2 million individuals aged 40 to 69 who were registered with the UK's National Health Service (NHS) and lived within a short traveling time (typically about 25 miles) of 1 of 22 dedicated assessment centers. The choice of their location was determined by population density, ease of access, and potential to reach certain types of participants (e.g., ethnic minority groups and those living in more socioeconomically deprived areas). During 2006 to 2010, 502,000 participants were recruited (5.5% of those invited). Although the participation rate was low and those who joined the study were somewhat healthier and wealthier than the UK population across the same age range (*24*), the cohort includes large numbers of individuals across a broad spectrum of potential risk factors that vary from low to high exposure.

It is this heterogeneity across different levels of exposure (e.g., genetic, lifestyle, sociodemographic, and environmental), and not the relatively low overall participation rate, that largely determines the generalizability of the findings to the broader UK population (*25*). For example, although individuals from more socioeconomic-deprived areas are underrepresented in UK Biobank (16% versus 33% in the UK population), there are sufficiently large numbers of this group (82,000) to enable reliable assessment of the association

Table 1. Cumulative numbers of observed (2020) and predicted incident cases of various health conditions.

Condition	Year of diagnosis		
	Observed*	Predicted	
	2020	2027	2032
Diabetes	31,000	54,000	70,000
Myocardial infarction	15,000	30,000	46,000
Stroke	12,000	25,000	37,000
COPD	25,000	47,000	65,000
Depression	25,000	39,000	47,000
Breast cancer	9,000	14,000	18,000
Colorectal cancer	5,000	8,000	11,000
Lung cancer	4,000	6,000	8,000
Prostate cancer	10,000	16,000	20,000
Hip fracture	5,000	13,000	22,000
Rheumatoid arthritis	4,000	6,000	8,000
Alzheimer's disease	5,000	17,000	37,000
Parkinson's disease	4,000	10,000	14,000

*Observed values are based on incident events identified from linkage to records of deaths, hospitalizations, cancers, and primary care in the cohort to the end of 2020.

Downloaded from https://www.science.org at University of Connecticut on January 10, 2024

of socioeconomic deprivation with disease risk. By contrast, although UK Biobank is reasonably representative of the national distribution for different ethnic groups, with 29,000 participants recruited from Black and other ethnic minority groups (which is about the same proportion, ~5%, as the rest of the UK population at that time) (*26*), it is insufficient to examine reliably the differences in exposure-disease associations by ethnicity. Although UK Biobank is currently the largest study in the world with whole-genome sequencing data on individuals of African and South Asian ancestry (*27*), the numbers are still relatively small (with about 10,000 participants in each ethnic group).

Researchers who wish to present simple summary statistics (e.g., means or proportions) using UK Biobank data that are representative of the underlying population could consider using sampling weights that reflect the population distribution of the variables under investigation, although such techniques have not been widely used. However, one research group found that standardization of the prevalence of lifestyle factors with those derived from national survey data did not substantially alter the magnitude or direction of the association of lifestyle factors with mortality from cardiovascular disease or cancer (*28*). The one notable exception was an attenuation of the apparent protective association of alcohol with cardiovascular disease, which has been shown to be likely affected by bias (*29*).

There are circumstances where lack of representativeness may introduce bias, particularly if the risk factors of interest are related to study selection (an example of collider bias) (*30*). For example, UK Biobank participants are more likely to be nonsmokers and to live in more affluent areas than the general population in the same age range. Given that area-level socioeconomic deprivation is moderately inversely correlated both with participation in UK Biobank and lung cancer, this nonrepresentativeness may attenuate the observed association of smoking with lung cancer if the effects of smoking and socioeconomic deprivation are not independent of each other (*31*). Likewise, UK Biobank participants are more likely to use vitamin supplements and to have lower incident disease rates than the general population (at least in the early years of follow-up), leading to an apparent inverse association between glucosamine supplement usage and mortality (*32*). Analyses involving genetic variants that cluster by place of birth also have the potential to yield biased associations if standard variables, such as assessment center and ancestry-based principal components, cannot completely correct for this latent structure (*33*). However, for most genetic analyses where confounding from other risk factors is likely low, selection bias would typically be expected to be modest.

Consequently, in the interpretation of all research findings—whether they arise from the UK Biobank study or other prospective studies—it is important to consider the extent to which they might be affected by selective participation (i.e., selection bias). Given that traditional methods of identifying and controlling for selection bias (and other types of bias) may not be adequate, graphical tools such as directed acyclic graphs may provide a useful visual representation of the underlying assumptions about the relationships among exposures, potential confounders, mediators, and outcomes, as well as how they relate to study participation (*34*). Sensitivity analyses that include factors correlated with participation (and ongoing engagement) as covariates in the exposure-disease model can be performed, and probability weighting, simulations, and multiple imputation can be used to explore the potential impact of missing values related to participation on effect estimates (*31*, *35*).

The general consistency of research findings in UK Biobank with those in other studies (*36*–*38*)—particularly studies considered to be representative of the underlying population—suggests that many of the exposure-disease associations found in UK Biobank are largely generalizable to other populations. For example, the direction and magnitude of associations of genetic variants with osteoarthritis in UK Biobank are consistent with the associations observed in deCODE, which recruited more than half of Iceland's adult population (*39*). Likewise, although the frequency of genetic variants may vary substantially in studies conducted in different populations (resulting in differing statistical power to detect associations), the direction and magnitude of genetic associations are typically similar across populations, e.g., the association of specific *GPR75* gene variants with obesity in United Kingdom, United States, and Mexico cohorts (*40*).

Nonetheless, there may be circumstances in which associations between an exposure and disease risk vary across different populations. For example, polygenic risk scores developed and tested in populations of European ancestry often perform less well when applied to African and South Asian populations, owing to differences in allele frequencies and linkage disequilibrium patterns between the ethnic groups (*41*). As such, other large population cohorts with biological samples are needed around the world to increase the heterogeneity of genetic and nongenetic risk factors for disease (*42*) (Table 2). For example, studies established in Mexico (150,000 participants) and China (500,000 participants) at about the same time as UK Biobank have enabled reliable investigation into the association between the risk of hypertension with body weight above and below the Western norm (*43*, *44*). Large-scale studies established across Europe and China have also taken advantage of the heterogeneity of dietary and other exposures across different populations (*45*, *46*). Genetic and other assays of stored samples in these studies are extending the range of genomic and other biological risk factors that can be investigated. New large-scale prospective studies are now established in the United States, e.g., All of Us (*47*) and the Million Veterans Program (*48*), and are also being established in Asia and parts of Africa, e.g., Noncommunicable Diseases Genetic Heritage Study in Nigeria (*49*, *50*). This will further increase the ability to assess associations with disease risk across a broad range of genetic (and nongenetic) factors as long as there is sufficient duration of follow-up.

Reliable assessment of a wide range of exposures

The inclusion of participants exposed to different levels of risk factors (e.g., ranging from low to high intake of different dietary factors, smoking, sun exposure, etc.) is of value in assessing the generalizability of findings, which is enhanced further by analyses across studies established in different populations. However, all observational studies face challenges of exposure measurement error, in which risk factors and their potential confounders are measured imperfectly or incompletely, thereby introducing both random error (when measurements fluctuate randomly around their true value) and systematic error (when measurements vary in the extent to which they are higher or lower than their true value).

As a result, UK Biobank has put substantial effort into collecting comprehensive, accurate, and high-quality data for many different types of exposures. Repeated measures have also been conducted in subsets of participants to address random error in exposure levels

Downloaded from https://www.science.org at University of Connecticut on January 10, 2024

Table 2. Sampling characteristics of selected general population prospective studies with at least 250,000 participants, containing genomic, behavioral, and health outcomes data. The International HundredK+ Cohorts Consortium (https://ihccglobal.org/) has details of other prospective studies of less than 250,000 participants.

Study name	Recruitment dates (range)	Location	Sample size	Sex; age at recruitment	Population from which the sample was recruited	Participation rate
23andMe (https://23andme.com)	2007–present	Global (mainly United States)	6.8 million	MF; 13+	Customers of a personal genetics company	Not known
45 and Up (*93*)	2006–2009	Australia	267,000	MF; 45+	New South Wales residents enrolled in Medicare, recruited through direct invitations	19%
All of Us (*47*)	2018–present	United States	Ongoing Aim: 1 million	MF; 18+	Varied approaches, many of which are targeted at under-represented groups via direct and indirect means	Not known
Canadian Partnership for Tomorrow's Health (CanPath) (*94*)	2008–present	Canada	330,000	MF; 30–74	Residents across Canada recruited into seven regional cohorts using varied approaches	Not known
China Kadoorie Biobank (*46*)	2004–2008	China	510,000	MF; 30–70	Residents of 10 geographically defined regions across China, recruited through direct invitations	30%
European Prospective Investigation into Cancer, Chronic Diseases, Nutrition and Lifestyle (EPIC) (*45*)	1992–2000	10 European countries	520,000	MF; 35–70	Residents from 23 centers located in 10 European countries recruited through direct invitations	Not known
Kaiser Permanente Research Bank (*95*)	2007–2013	United States	400,000	MF; 18+	Members of Kaiser Permanente health plan recruited through direct invitations, in-person communication, and via website	20–50% of each areas' insured population
Million Veterans Program (*48*)	2011–present	United States	Ongoing Aim: 1 million	MF; 18+	Members of the Veterans Health Administration System recruited through direct invitations and indirect (promotional materials) methods	14%
UK Biobank (*26*)	2006–2010	United Kingdom	500,000	MF; 40–69	Residents living close to 22 assessment centers in the United Kingdom, recruited via direct invitations	5.5%

and thereby be able to correct for regression-dilution bias. However, there is also real value in being able to perform cohort-wide repeat measures that would allow the relevance of individual changes in exposures over time to be assessed.

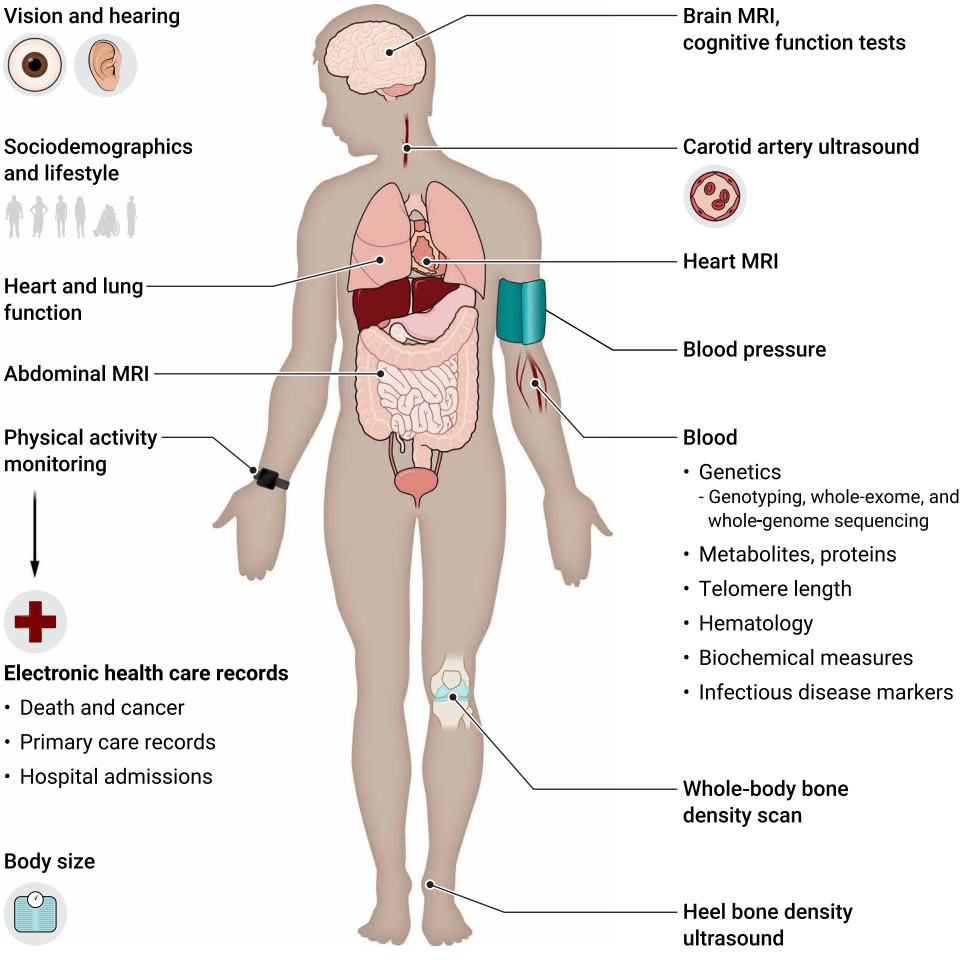
Depth and breadth of exposure measurements

In UK Biobank, a wide range of questionnaires and physical devices (e.g., spirometer to measure lung function, sphygmomanometer to measure blood pressure, bioimpedance device to measure body

Downloaded from https://www.science.org at University of Connecticut on January 10, 2024

composition, dynamometer to measure hand grip strength, etc.) have been used to collect data that are reliable, valid, and of high scientific value (Fig. 1) (*26*, *51*). Such data continue to be collected and extended. During recruitment, UK Biobank used touch-screen and computer-assisted personal interview direct data entry systems (instead of paper-based approaches that were routinely used at the time in such studies), as well as direct data transfer from measurement devices. This strategy enhanced data accuracy and completeness by supporting automated real-time consistency checks and data quality monitoring to identify and correct errors. Participants were also asked to bring certain information (e.g., medications, operations, family history, and birth details) to reduce errors associated with memory recall. However, perhaps the greatest benefit of using a touch-screen data entry model was that it reduced the time taken to collect data and thereby enabled a greater range of potential risk factors for disease to be collected. For example, data on sociodemographic factors (income, education, and occupation), ethnicity, family history, lifestyle (diet, alcohol consumption, smoking history, physical activity, sleep, sun exposure, and sexual history), early life factors, psychosocial factors, medical history, and cognition and environmental exposures were all collected via the touch-screen questionnaire in about 15 min.

A wide range of physical measurements were also taken for all 500,000 participants, comprising blood pressure, anthropometry (sitting and standing height, weight, waist and hip circumference, and bioimpedance measures), hand grip strength, measures of vision, and lung function. Blood and urine samples were also collected for long-term storage (Fig. 1). A proportion of the cohort also underwent a heel ultrasound for bone density, pulse wave velocity of arterial stiffness, a hearing test (180,000 participants), an eye examination (including refractive index), intraocular pressure measurements, a retinal photograph and optical coherence tomography (120,000 participants), a cardiorespiratory fitness test with a four-lead electrocardiogram (78,000 participants), and collection of a saliva sample (~85,000 participants). Since the baseline assessment, UK Biobank continues to collect additional data from large subsets of the cohort. This has included data on physical activity using a 7-day accelerometer (in 100,000 participants, with 2500 undergoing a repeat assessment), a multimodal imaging assessment (in up to 100,000 participants, with 60,000 undergoing a repeat assessment over the next few years), and a series of web-based questionnaires that cover specific exposures in more depth (e.g., diet, cognition, and occupational history).



CREDIT: A. MASTIN/*SCIENCE TRANSLATIONAL MEDICINE*

Fig. 1. Types of data in UK Biobank. Shown are the types of data collected in UK Biobank, including data collected at in-person assessments such as lifestyle factors, medical history, blood pressure and other physical measures, and imaging scans. Other data include information from online questionnaires, data generated from biological samples, and data derived from electronic health care records.

Rigorous approaches have also been taken to sample collection, processing, retrieval, and assay measurement. Before the start of the UK Biobank, a series of pilot studies were conducted to determine the optimal method for sample collection and processing (*52*), followed by the development of (at that time) a state-of-the-art robotic system and sample tracking software to ensure consistency of sample processing. Currently, genomic data (genome-wide genotyping and imputation, whole-exome and whole-genome sequence data, and telomere length) as well as hematological and biochemical data are available for the whole cohort (Fig. 1). UK Biobank's general policy of performing cohort-wide assays supports research into a wide number of conditions and helps to avoid measurement errors that would otherwise occur with different assay methods, reagents, and equipment in different laboratories used in different subsets of the cohort at different times. To facilitate quality control, algorithms were developed to retrieve sample aliquots in a sequence that avoided clustering by recruitment location, date, or time of the day (*53*). Consequently, when assaying samples from participants in this quasi-random order, the mean biomarker concentration across batches and analyzers should be constant, which allows correction for variation caused by laboratory drift. Throughout the assay process,

Downloaded from https://www.science.org at University of Connecticut on January 10, 2024

the data are reviewed to identify issues and either address them in real time (e.g., if specific batches require remeasurement) or make any adjustments retrospectively. For example, after assay measurements of blood biochemistry markers, these data were corrected for systematic error caused by unexpected dilution that occurred in some aliquots during sample processing (*53*). Moreover, highly efficient assay methods minimize sample depletion, with currently less than 10% of the baseline blood samples used so far. This will allow other types of assays (e.g., epigenetics, transcriptomics, and proteomics) to be conducted on a cohort-wide basis when technological advances make this possible.

The collection of different types of data that describe the same (or highly related) exposures can also contribute to accuracy. In particular, a more precise assessment performed in a subset of participants could be used to correct for any random and systematic error inherent in the less precise baseline measures conducted in the full cohort (*54*). For example, data from an accelerometer device worn by 100,000 UK Biobank participants were used to calibrate self-reported physical activity estimates provided by all 500,000 participants at recruitment (*55*). Similarly, data on body fat composition available from dual-energy x-ray absorptiometry scans (*56*), which are being collected in up to 100,000 participants attending an imaging assessment, can be used to calibrate the bio-impedance measures available from the full cohort. Detailed dietary data from web-based questionnaires collected from more than 200,000 participants can also be used to predict food and nutrient intake for the entire cohort, as demonstrated in other studies (*54*).

The collection of data on a wide range of measures enables researchers to allow not only for more complete and accurate measurement of exposures but also for potential confounders (extraneous factors that are associated with the exposure and outcome) and mediators (factors that are on the causal pathway between the exposure and the outcome). This is important, because random error in exposure measures can cause systematic attenuation of any true association, whereas random measurement error of confounders can result in an apparent exposure-disease association where none really exists. For example, the observed inverse association of fruit and vegetable intake with cardiovascular disease risk in UK Biobank is likely to be due largely to residual confounding by socioeconomic factors, which are difficult to assess and therefore subject to measurement error (*57*). The ability of UK Biobank to obtain more detailed information in the future about socioeconomic factors (such as education, occupa web-based questionnaires) would enable more precise characterization and, hence, even better adjustment for these important factors.

Because all epidemiological studies suffer, to a greater or lesser extent, from imperfect measurement of exposures and their potential confounders, various analytical methods have been developed to quantify and control for this. Perhaps the simplest approach is the comparison of likelihood ratio statistics associated with the exposure of interest in the models before and after adjustment for covariates. Generally speaking, a large proportional reduction in the likelihood ratio chi-square (LRχ^2) test after addition to the model of covariates indicates that the association likely remains affected by residual confounding, given that adjustment for confounders that are perfectly measured would be expected to reduce the χ^2 even further (*6*). An increasingly popular approach to distinguish the likely causal effect of an exposure (from that of extraneous confounders) is the use of Mendelian randomization—facilitated in UK Biobank by the extensive genetic information available on all participants—whereby specific genetic variants are used as proxies for exposures of interest or their confounders. For example, this approach has provided strong support for a causal role of body fat mass and interleukin-6 in the development of cardiovascular conditions (*58*, *59*). Conversely, Mendelian randomization has not provided support for a protective effect of vitamin D against coronavirus disease (COVID-19) (*60*), cancer, or cardiovascular outcomes (*61*), although it should be noted that Mendelian randomization analyses may also be affected by bias in some circumstances (*62*). When associations of genetic variants with the relevant nongenetic risk factors are weak (such that Mendelian randomization may not be effective), the likely impact of residual confounding due to imprecision in measured variables included in the model can be assessed using other analytical approaches, such as probabilistic or multiple-bias analysis (*34*, *63*). The use of different analytical strategies to triangulate evidence (e.g., comparing results from models that include traditional observational variables with those that use genetic instrumental variables) will enable researchers to assess different biases and their potential impact on causal inference in a more robust manner.

Repeated exposure measurements

Random errors in the measurement of risk factors can lead to substantial underestimation of the strength of their associations with subsequent health outcomes (regression dilution bias) (*64*, *65*) and to substantial residual confounding when measurement error is present in confounders (*66*). These biases may be increased further through random error in risk factor measurements that occur during prolonged follow-up in prospective cohorts. For example, the true association of blood pressure and cholesterol with cardiovascular disease risk may be underestimated by about one-third in the first decade of follow-up and up to two-thirds in the third decade (*64*). However, despite regression dilution being one of the most important biases in exposure-disease associations, it is often overlooked in analyses of prospective studies, including UK Biobank (with some exceptions) (*67*–*70*). It is possible to correct for regression dilution bias by using repeat measures from a relatively small subset of the cohort. UK Biobank performed a repeat assessment on 20,000 participants in 2012–2013 to allow researchers to address this issue specifically. Remeasures collected during the imaging assessment of up to 100,000 UK Biobank participants during 2014–2024 and a repeat assessment of up to 60,000 during 2019–2029 can be used to make appropriate time-dependent corrections for the effects of regression dilution bias.

In addition to addressing error caused (largely) by random error in baseline risk factors, repeated measures would also enable correction for systematic intra-individual changes in exposures over time, which may lead to either overestimation or underestimation of associations depending on the nature and magnitude of misclassification. For example, secular trends in the reduction of smoking or exposure to environmental pollutants may lead to an underestimation of their association with disease risk if solely based on baseline measures. To help address this issue, UK Biobank is exploring opportunities to collect information on longitudinal changes in environmental exposures (e.g., from existing data on changes in participants' residential location or future data collection using smartphone GPS tracking) to enable more accurate inferences to be made about how changes in environmental exposures affect health in the long term. It is also intended to repeat previous web-based questionnaires to

Downloaded from https://www.science.org at University of Connecticut on January 10, 2024

capture longitudinal changes in specific lifestyle factors, such as diet and sleep.

Whereas repeated measures of the baseline assessment are being captured during the imaging assessments in a subset of the UK Biobank cohort, it would be desirable to perform a future repeat assessment of a wide range of exposures in as many of the participants as possible. This would allow investigation of how lifestyle, physical, and biochemical changes over time influence disease risk and progression, thereby helping to determine the temporality of associations and their underlying mechanisms. Data collection for as many surviving participants as possible would also reduce systematic error caused by differential participation rates that are related to the exposures and outcomes under investigation. UK Biobank generally has excellent participant engagement with an ongoing series of repeated web-based questionnaires (with response rates of >50%), physical activity monitoring (45% for the first assessment, of whom 63% also performed a repeat assessment), and imaging assessments (24% for the first assessment and 65% for a repeat assessment). However, researchers should be aware that participants who engage in ongoing data collection activities (including repeat assessments) might not be representative of the cohort as a whole. For example, genetic variants associated with completing UK Biobank online questionnaires and activity monitoring are correlated with several metrics of better health (*31*). Attrition bias has been documented in other prospective studies (*71*–*73*), suggesting that similar factors affect ongoing participant engagement in many cohorts, regardless of their design, recruitment strategy, or study population.

Comprehensive ascertainment of health outcomes

To minimize bias in exposure-disease associations, it is important that health outcomes are identified in a comprehensive manner and in as much detail as possible. Linkage to routine electronic health records, supplemented with information from self-reported questionnaires and other remote methods, and in-person assessments focused on specific outcomes (such as dementia) will help to deeply characterize health outcomes that are of high priority. The ability to combine these data from disparate sources to generate "off-the-shelf" outcomes that can be easily interpreted by nonspecialists will further increase the usability and reproducibility of research using these data.

All cohort studies need a comprehensive and efficient way of following participants' health over the long term to identify a wide range of disease outcomes. Unlike many countries (including the United States and most low- to middle-income countries), the UK's NHS collates and stores electronic health administrative records for clinical care. However, the data content, format, and governance requirements may differ for England, Wales, and Scotland. To identify a wide range of health outcomes over a prolonged period, UK Biobank has linked to these health administrative records for all participants. This has the advantage of minimizing ascertainment bias and reducing loss to follow-up or attrition bias by providing cohort-wide follow-up information without the need for active participant recontact, which may be incomplete. Moreover, the low rate of UK Biobank participants requesting that all of their data and samples be withdrawn from the study (0.2%, most of which occurred soon after recruitment) also minimizes systematic bias associated with loss to follow-up from nonrandom subgroups of the cohort.

To date, UK Biobank has linked NHS health care data from centralized national cancer and death registries and hospital inpatient admissions for all participants. In contrast, primary care data are not centralized but instead are held by commercial electronic system suppliers under the control of individual general practices, so it has been more challenging to obtain the agreements necessary to obtain these data for all participants. Primary care data are currently available for 45% of the UK Biobank cohort for general research purposes (which represents complete coverage from one primary care system supplier, up to 2016/2017) and for 80% of the cohort for COVID-19 research (complete coverage from two system suppliers in England, up to mid-2021, enabled by emergency legislation to facilitate COVID-19 research). Whereas both subsets are broadly representative of the cohort with respect to the distribution of potential exposures, researchers should be encouraged to check these underlying assumptions before analysis. Incorporation of primary care data for all 500,000 participants for all types of health-related research would be of enormous value because it will increase substantially the number of health outcomes that can be detected (thereby increasing statistical power) and their diagnostic accuracy (thereby increasing specificity). For example, although addition of primary care data would increase the numbers of myocardial infarction cases identified by less than 5%, the numbers of cases identified of diabetes and COPD would increase by about 40% (Fig. 2). Primary care data are also important for investigating risk factors associated with disease severity, where associations may differ between milder disease subtypes generally captured in primary care records and more severe disease captured in hospital admission records.

Whereas linkage to health records ensures comprehensive coverage, there is the possibility of "collider bias" if health outcomes are differentially ascertained based on participant characteristics (e.g., ethnicity), as reported by some researchers in the context of COVID-19 studies (*74*). However, there are a range of analytical approaches that can be used to investigate this type of bias (*74*–*76*), and the ascertainment of most health outcomes is not so strongly influenced by these characteristics.

Specificity of health outcomes

Given that the prospective nature of cohort studies facilitates research into many diseases, the challenge is not only how to identify probable cases of disease but also how to increase the precision and specificity of those diagnoses. The aim is to avoid a situation where insufficient data on health outcomes lead to misclassification of cases and noncases, thereby reducing statistical power to detect an association. As such, UK Biobank's aim is to ascertain as many cases as possible (i.e., to achieve adequate sensitivity) while minimizing the number of false-positive cases (i.e., achieving a high positive predictive value). It is worth recognizing that it is not necessary to identify all cases of a disease because false negatives will be diluted by the much larger number of "true" controls and so will have limited impact. To help identify as many cases as possible, UK Biobank administers various web-based questionnaires, developed in close collaboration with relevant experts, to collect data on health outcomes that are incompletely recorded in health records, such as depression and anxiety (*77*), as well as neurodevelopmental and gastrointestinal conditions.

It is also important to characterize disease subtypes because low biological specificity can limit interpretation of results. To address this, UK Biobank (*78*–*80*) and other open-access resources (*81*) have developed a number of algorithmically defined health outcomes based on interoperable code lists from electronic health care

Downloaded from https://www.science.org at University of Connecticut on January 10, 2024

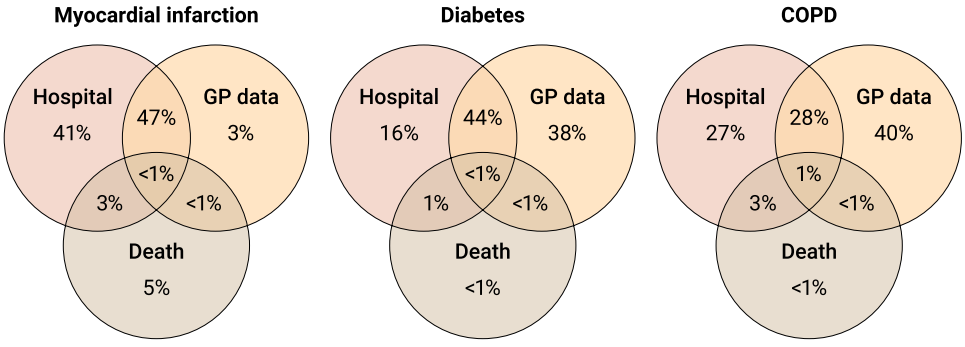


Fig. 2. Incident disease cases since recruitment. Shown is the proportion of incident cases for three common diseases (myocardial infarction, diabetes, and COPD) ascertained since recruitment of these participants into UK Biobank. Disease cases were identified through hospital inpatient admissions, primary care (i.e., data from general practitioners) data, and death data available in electronic health care records.

records. Diagnostic codes contained in these records have also been mapped to a common standard (ICD-10) to facilitate broad-brush research. Whereas these coded health outcomes may be sufficient for most research purposes, they may lack specificity to identify disease subtypes. This could lead to materially biased estimates of associations if the determinants of these apparently similar, but etiologically different, disease subtypes differ. For example, whereas blood pressure is strongly positively associated with the risk of both ischemic and hemorrhagic stroke (*82*), the association of cholesterol and certain genetic variants with stroke differs substantially by subtype (*83*, *84*), providing clues to the underlying etiology of this heterogeneous condition. To increase the specificity of health outcomes beyond the available coded data, UK Biobank intends to collect detailed data on disease subtypes over the next few years. For example, this could include disease-specific registers, such as the National Diabetes Audit, that collects data on diabetes subtypes, clinical scans to identify stroke subtypes, digitized histopathology slides to determine tumor morphological subtypes, and in-person assessments to characterize dementia subtypes.

It is possible to identify some disease subtypes using other data already available in the UK Biobank resource. For example, biochemistry measures have been used to ascertain chronic kidney disease (*85*), MRI scans collected in up to 100,000 participants have been used to define dilated cardiomyopathy (*86*) and nonalcoholic fatty liver disease (*87*), and genetic data have been used to distinguish diabetes subtypes (*88*). However, researchers should be aware of the potential for generating misleading associations where the exposure of interest (e.g., genetic variants or biochemistry measures) has, in part, been used to define the outcome.

Long duration of follow-up

For any prospective study, a long duration of follow-up (i.e., decades or more) is needed for sufficiently large numbers of health outcomes to accrue for reliable investigation. It also allows for the identification of recurring events and factors associated with disease progression. Although the incidence of common health outcomes during the early years of follow-up in UK Biobank was somewhat lower than in the general population due to the "healthy volunteer" effect, which is typical of such studies (*89*), its impact is now reduced because the cohort has aged. With prolonged follow-up, large numbers of incident cases of a wide range of conditions have already occurred. Over the next 5 to 10 years, there will be thousands of incident cases of common outcomes (Table 1), enabling reliable investigation of their genetic, lifestyle, and environmental determinants.

The rationale for recruiting middle-aged participants was to collect risk factor data many years before the development of any given condition, thereby minimizing reverse causation bias. However, conditions that have a long prodromal phase (e.g., dementia or diabetes) or that can exist for years before a clinical diagnosis is made (such as prostate cancer) may affect the levels of risk factors measured at recruitment and create spurious associations. For example, associations observed between high insulin-like growth factor-I (IGF-I) concentrations and increased risks of cataract and diabetes were substantially attenuated after excluding the first 5 years of follow-up in UK Biobank (*90*), suggesting that baseline IGF-I concentrations may be altered as a result of early pathophysiological processes. Other large-scale longitudinal studies have also shown that apparent inverse associations between lifestyle factors and dementia risk are also likely to be due to reverse causation bias during the first 10 to 15 years of follow-up (*91*). Consequently, researchers should consider the impact of exclusion of participants with prevalent disease before analysis and perform sensitivity analyses to assess exposure-disease associations across different periods of follow-up to determine whether the first years of follow-up should be excluded (*92*).

CONCLUSIONS

The success of UK Biobank has been due, in large part, to not only the altruism of the 500,000 volunteers but also the global research community who have been—and continue to be—involved in expanding the range of exposures and outcomes available for research. Such enhancements (e.g., sample assays, linkage to specific health care datasets and environmental sources, etc.) help to ensure that this resource fulfills the needs of researchers and remains at the forefront of scientific discovery.

UK Biobank's large-scale prospective design and easy access to a wealth of genetic, phenotypic, and health data provide a powerful resource to help address previously unanswerable questions of the determinants of incident disease and enabling research into risk prediction and identification of early biomarkers of disease. Whereas the UK Biobank study has attempted to minimize random and systematic errors in the measurement of exposures and outcomes with good study design, researchers need to use the data in ways that best answer the questions posed and to be aware of and, where necessary, to use analytical methods to take account of potential biases when interpreting research findings.

Easy accessibility of UK Biobank data and research results (including the underlying analytical code) is enabling the community to directly peer review research by undertaking replication analyses or to apply different methods to the same research question to confirm or refute the findings of others. In particular, investigation of approaches used to identify and quantify the uncertainty of the results based on sensitivity analyses that examine systematic bias will

CREDIT: A. MASTIN/SCIENCE TRANSLATIONAL MEDICINE

Downloaded from https://www.science.org at University of Connecticut on January 10, 2024

provide transparency in the interpretation of findings that have, until now, generally been underreported.

Whereas UK Biobank is well suited to address a wide range of health-related research questions, similar studies in other populations with different ranges of exposures and outcomes are needed. Together, they will enable a greater range of risk factors and diseases to be analyzed and allow for replication of associations, which is essential before determining the extent to which any specific research findings are generalizable to different populations. Scientific discoveries benefit from the availability of data from diverse populations that cover a wide range of the many different genetic, ancestral, ethnic, lifestyle, and environmental factors that may influence risk of a broad range of diseases.

Supplementary Materials

This PDF file includes:

Table S1

REFERENCES AND NOTES

1. R. Doll, A. B. Hill, Lung cancer and other causes of death in relation to smoking; a second report on the mortality of British doctors. *BMJ* **2**, 1071–1081 (1956).
2. J. Truett, J. Cornfield, W. Kannel, A multivariate analysis of the risk of coronary heart disease in Framingham. *J. Chronic Dis.* **20**, 511–524 (1967).
3. Emerging Risk Factors Collaboration, S. Erqou, S. Kaptoge, P. L. Perry, E. D. Angelantonio, A. Thompson, I. R. White, S. M. Marcovina, R. Collins, S. G. Thompson, J. Danesh, Lipoprotein(a) concentration and the risk of coronary heart disease, stroke, and nonvascular mortality. *JAMA* **302**, 412–423 (2009).
4. Emerging Risk Factors Collaboration, D. Wormser, S. Kaptoge, E. D. Angelantonio, A. M. Wood, L. Pennells, A. Thompson, N. Sarwar, J. R. Kizer, D. A. Lawlor, B. G. Nordestgaard, P. Ridker, V. Salomaa, J. Stevens, M. Woodward, N. Sattar, R. Collins, S. G. Thompson, G. Whitlock, J. Danesh, Separate and combined associations of body-mass index and abdominal adiposity with cardiovascular disease: Collaborative analysis of 58 prospective studies. *Lancet* **377**, 1085–1095 (2011).
5. Collaborative Group on Hormonal Factors in Breast Cancer, Breast cancer and hormone replacement therapy: Collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer. *Lancet* **350**, 1047–1059 (1997).
6. Collaborative Group on Hormonal Factors in Breast Cancer, Alcohol, tobacco and breast cancer—Collaborative reanalysis of individual data from 53 epidemiological studies, including 58,515 women with breast cancer and 95,067 women without the disease. *Br. J. Cancer* **87**, 1234–1245 (2002).
7. Collaborative Group on Hormonal Factors in Ovarian Cancer, V. Beral, R. Doll, C. Hermon, R. Peto, G. Reeves, Ovarian cancer and oral contraceptives: Collaborative reanalysis of data from 45 epidemiological studies including 23,257 women with ovarian cancer and 87,303 controls. *Lancet* **371**, 303–314 (2008).
8. E. Uffelmann, Q. Q. Huang, N. S. Munung, J. de Vries, Y. Okada, A. R. Martin, C. M. Martin, T. Lappalainen, D. Posthuma, Genome-wide association studies. *Nat. Rev. Methods Primers* **1**, 59 (2021).
9. J. A. Luan, M. Y. Wong, N. E. Day, N. J. Wareham, Sample size determination for studies of gene-environment interaction. *Int. J. Epidemiol.* **30**, 1035–1040 (2001).
10. M. Conroy, J. Sellors, M. Effingham, T. J. Littlejohns, C. Boultwood, L. Gillions, C. L. M. Sudlow, R. Collins, N. E. Allen, The advantages of UK Biobank's open-access strategy for health research. *J. Intern. Med.* **286**, 389–397 (2019).
11. S. Cassidy, H. Fuller, J. Chau, M. Catt, A. Bauman, M. I. Trenell, Accelerometer-derived physical activity in those with cardio-metabolic disease compared to healthy adults: A UK Biobank study of 52,556 participants. *Acta Diabetol.* **55**, 975–979 (2018).
12. A. Doherty, D. Jackson, N. Hammerla, T. Plötz, P. Olivier, M. H. Granat, T. White, V. T. van Hees, M. I. Trenell, C. G. Owen, S. J. Preece, R. Gillions, S. Sheard, T. Peakman, S. Brage, N. J. Wareham, Large scale population assessment of physical activity using wrist worn accelerometers: The UK biobank study. *PLOS One* **12**, e0169649 (2017).
13. M. Borga, J. West, J. D. Bell, N. C. Harvey, T. Romu, S. B. Heymsfield, O. D. Leinhard, Advanced body composition assessment: From body mass index to body composition profiling. *J. Invest. Med.* **66**, 1–9 (2018).
14. Y. Liu, N. Basty, B. Whitcher, J. D. Bell, E. P. Sorokin, N. van Bruggen, E. L. Thomas, M. Cule, Genetic architecture of 11 organ traits derived from abdominal MRI using deep learning. *eLife* **10**, e65554 (2021).
15. A. McKay, H. R. Wilman, A. Dennis, M. Kelly, M. L. Gyngell, S. Neubauer, J. D. Bell, R. Banerjee, E. L. Thomas, Measurement of liver iron by magnetic resonance imaging in the UK Biobank population. *PLOS One* **13**, e0209340 (2018).
16. F. Alfaro-Almagro, M. Jenkinson, N. K. Bangerter, J. L. R. Andersson, L. Griffanti, G. Douaud, S. N. Sotiropoulos, S. Jbabdi, M. Hernandez-Fernandez, E. Vallee, D. Vidaurre, M. Webster, P. M. Carthy, C. Rorden, A. Daducci, D. C. Alexander, H. Zhang, I. Dragonu, P. M. Matthews, K. L. Miller, S. M. Smith, Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage* **166**, 400–424 (2018).
17. W. Bai, H. Suzuki, J. Huang, C. Francis, S. Wang, G. Tarroni, F. Guitton, N. Aung, K. Fung, S. E. Petersen, S. K. Piechnik, S. Neubauer, E. Evangelou, A. Dehghan, D. P. O'Regan, M. R. Wilkins, Y. Guo, P. M. Matthews, D. Rueckert, A population-based phenome-wide association study of cardiac and aortic structure and function. *Nat. Med.* **26**, 1654–1662 (2020).
18. O. Canela-Xandri, K. Rawlik, A. Tenesa, An atlas of genetic associations in UK Biobank. *Nat. Genet.* **50**, 1593–1599 (2018).
19. G. McInnes, Y. Tanigawa, C. DeBoever, A. Lavertu, J. E. Olivieri, M. Aguirre, M. A. Rivas, Global Biobank Engine: Enabling genotype-phenotype browsing for biobank summary statistics. *Bioinformatics* **35**, 2495–2497 (2019).
20. T. G. Richardson, S. Harrison, G. Hemani, G. Davey Smith, An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *eLife* **8**, e43657 (2019).
21. D. A. Grimes, K. F. Schulz, Cohort studies: Marching towards outcomes. *Lancet* **359**, 341–345 (2002).
22. P. R. Burton, A. L. Hansell, I. Fortier, T. A. Manolio, M. J. Khoury, J. Little, P. Elliott, Size matters: Just how big is BIG?: Quantifying realistic sample size requirements for human genome epidemiology. *Int. J. Epidemiol.* **38**, 263–273 (2009).
23. S. Lewington, personal correspodence (2022).
24. A. Fry, T. J. Littlejohns, C. Sudlow, N. Doherty, L. Adamska, T. Sprosen, R. Collins, N. E. Allen, Comparison of sociodemographic and health-related characteristics of UK biobank participants with those of the general population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).
25. K. J. Rothman, J. E. Gallacher, E. E. Hatch, Why representativeness should be avoided. *Int. J. Epidemiol.* **42**, 1012–1014 (2013).
26. C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, R. Collins, UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Med.* **12**, e1001779 (2015).
27. B. V. Halldorsson, H. P. Eggertsson, K. H. S. Moore, H. Hauswedell, O. Eiriksson, M. O. Ulfarsson, G. Palsson, M. T. Hardarson, A. Oddsson, B. O. Jensson, S. Kristmundsdottir, B. D. Sigurpalsdottir, O. A. Stefansson, D. Beyter, G. Holley, V. Tragante, A. Gylfason, P. I. Olason, F. Zink, M. Asgeirsdottir, S. T. Sverrisson, B. Sigurdsson, S. A. Gudjonsson, G. T. Sigurdsson, G. H. Halldorsson, G. Sveinbjornsson, K. Norland, U. Styrkarsdottir, D. N. Magnusdottir, S. Snorradottir, K. Kristinsson, E. Sobech, H. Jonsson, A. J. Geirsson, I. Olafsson, P. Jonsson, O. B. Pedersen, C. Erikstrup, S. Brunak, S. R. Ostrowski, DBDS Genetic Consortium, G. Thorleifsson, F. Jonsson, P. Melsted, I. Jonsdottir, T. Rafnar, H. Holm, H. Stefansson, J. Saemundsdottir, D. F. Gudbjartsson, O. T. Magnusson, G. Masson, U. Thorsteinsdottir, A. Helgason, H. Jonsson, P. Sulem, K. Stefansson, The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
28. E. Stamatakis, K. B. Owen, L. Shepherd, B. Drayton, M. Hamer, A. E. Bauman, Is Cohort Representativeness Passé? Poststratified Associations of Lifestyle Risk Factors with Mortality in the UK Biobank. *Epidemiol* **32**, 179–188 (2021).
29. J. R. Emberson, D. A. Bennett, Effect of alcohol on risk of coronary heart disease and stroke: Causality, bias, or a bit of both? *Vasc. Health Risk Manag.* **2**, 239–249 (2006).
30. S. Ebrahim, G. Davey Smith, Commentary: Should we always deliberately be non-representative? *Int. J. Epidemiol.* **42**, 1022–1026 (2013).
31. J. Tyrrell, J. Zheng, R. Beaumont, K. Hinton, T. G. Richardson, A. R. Wood, G. Davey Smith, T. M. Frayling, K. Tilling, Genetic predictors of participation in optional components of UK Biobank. *Nat. Commun.* **12**, 886 (2021).
32. K. Suissa, M. Hudson, S. Suissa, Glucosamine and lower mortality and cancer incidence: Selection bias in the observational studies. *Pharmacoepidemiology Drug Saf.* **31**, 1272–1279 (2022).
33. S. Haworth, R. Mitchell, L. Corbin, K. H. Wade, T. Dudding, A. Budu-Aggrey, D. Carslake, G. Hemani, L. Paternoster, G. D. Smith, N. Davies, D. J. Lawson, N. J. Timpson, Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat. Commun.* **10**, 333 (2019).
34. T. L. Lash, M. P. Fox, R. F. MacLehose, G. Maldonado, L. C. McCandless, S. Greenland, Good practices for quantitative bias analysis. *Int. J. Epidemiol.* **43**, 1969–1985 (2014).
35. M. R. Munafò, K. Tilling, A. E. Taylor, D. M. Evans, G. D. Smith, Collider scope: When selection bias can substantially influence observed associations. *Int. J. Epidemiol.* **47**, 226–235 (2018).
36. Emerging Risk Factors Collaboration, E. D. Angelantonio, S. Kaptoge, D. Wormser, P. Willeit, A. S. Butterworth, N. Bansal, L. M. O'Keeffe, P. Gao, A. M. Wood, S. Burgess,

Downloaded from https://www.science.org at University of Connecticut on January 10, 2024

D. F. Freitag, L. Pennells, S. A. Peters, C. L. Hart, L. L. Håheim, R. F. Gillum, B. G. Nordestgaard, B. M. Psaty, B. B. Yeap, M. W. Knuiman, P. J. Nietert, J. Kauhanen, J. T. Salonen, L. H. Kuller, L. A. Simons, Y. T. van der Schouw, E. Barrett-Connor, R. Selmer, C. J. Crespo, B. Rodriguez, W. M. M. Verschuren, V. Salomaa, K. Svärdsudd, P. van der Harst, C. Björkelund, L. Wilhelmsen, R. B. Wallace, H. Brenner, P. Amouyel, E. L. M. Barr, H. Iso, A. Onat, M. Trevisan, R. B. D'Agostino Sr., C. Cooper, M. Kavousi, L. Welin, R. Roussel, F. B. Hu, S. Sato, K. W. Davidson, B. V. Howard, M. J. G. Leening, M. Leening, A. Rosengren, M. Dörr, D. J. H. Deeg, S. Kiechl, C. D. A. Stehouwer, A. Nissinen, S. Giampaoli, C. Donfrancesco, D. Kromhout, J. F. Price, A. Peters, T. W. Meade, E. Casiglia, D. A. Lawlor, J. Gallacher, D. Nagel, O. H. Franco, G. Assmann, G. R. Dagenais, J. W. Jukema, J. Sundström, M. Woodward, E. J. Brunner, K.-T. Khaw, N. J. Wareham, E. A. Whitsel, I. Njølstad, B. Hedblad, S. Wassertheil-Smoller, G. Engström, W. D. Rosamond, E. Selvin, N. Sattar, S. G. Thompson, J. Danesh, Association of Cardiometabolic Multimorbidity With Mortality. *JAMA* **314**, 52–60 (2015).

37. H. S. Dashti, S. E. Jones, A. R. Wood, J. M. Lane, V. T. van Hees, H. Wang, J. A. Rhodes, Y. Song, K. Patel, S. G. Anderson, R. N. Beaumont, D. A. Bechtold, J. Bowden, B. E. Cade, M. Garaulet, S. D. Kyle, M. A. Little, A. S. Loudon, A. I. Luik, F. A. J. L. Scheer, K. Spiegelhalder, J. Tyrrell, D. J. Gottlieb, H. Tiemeier, D. W. Ray, S. M. Purcell, T. M. Frayling, S. Redline, D. A. Lawlor, M. K. Rutter, M. N. Weedon, R. Saxena, Genome-wide association study identifies genetic loci for self-reported habitual sleep duration supported by accelerometer-derived estimates. *Nat. Commun.* **10**, 1100 (2019).
38. J. Deelen, D. S. Evans, D. E. Arking, N. Tesi, M. Nygaard, X. Liu, M. K. Wojczynski, M. L. Biggs, A. van der Spek, G. Atzmon, E. B. Ware, C. Sarnowski, A. V. Smith, I. Seppälä, H. J. Cordell, J. Dose, N. Amin, A. M. Arnold, K. L. Ayers, N. Barzilai, E. J. Becker, M. Beekman, H. Blanché, K. Christensen, L. Christiansen, J. C. Collerton, S. Cubaynes, S. R. Cummings, K. Davies, B. Debrabant, J.-F. Deleuze, R. Duncan, J. D. Faul, C. Franceschi, P. Galan, V. Gudnason, T. B. Harris, M. Huisman, M. A. Hurme, C. Jagger, I. Jansen, M. Jylhä, M. Kähönen, D. Karasik, S. L. R. Kardia, A. Kingston, T. B. L. Kirkwood, L. J. Launer, T. Lehtimäki, W. Lieb, L.-P. Lyytikäinen, C. Martin-Ruiz, J. Min, A. Nebel, A. B. Newman, C. Nie, E. A. Nohr, E. S. Orwoll, T. T. Perls, M. A. Province, B. M. Psaty, O. T. Raitakari, M. J. T. Reinders, J.-M. Robine, J. I. Rotter, P. Sebastiani, J. Smith, T. I. A. Sørensen, K. D. Taylor, A. G. Uitterlinden, W. van der Flier, S. J. van der Lee, C. M. van Duijn, D. van Heemst, J. W. Vaupel, D. Weir, K. Ye, Y. Zeng, W. Zheng, H. Holstege, D. P. Kiel, K. L. Lunetta, P. E. Slagboom, J. M. Murabito, A meta-analysis of genome-wide association studies identifies multiple longevity genes. *Nat. Commun.* **10**, 3669 (2019).
39. U. Styrkarsdottir, S. H. Lund, G. Thorleifsson, F. Zink, O. A. Stefansson, J. K. Sigurdsson, K. Juliusson, K. Bjarnadottir, S. Sigurbjornsdottir, S. Jonsson, K. Norland, L. Stefansdottir, A. Sigurdsson, G. Sveinbjornsson, A. Oddsson, G. Bjornsdottir, R. L. Gudmundsson, G. H. Halldorsson, T. Rafnar, I. Jonsdottir, E. Steingrimsson, G. L. Norddahl, G. Masson, P. Sulem, H. Jonsson, T. Ingvarsson, D. F. Gudbjartsson, U. Thorsteinsdottir, K. Stefansson, Meta-analysis of Icelandic and UK data sets identifies missense variants in SMO, IL11, COL11A1 and 13 more new loci associated with osteoarthritis. *Nat. Genet.* **50**, 1681–1687 (2018).
40. P. Akbari, A. Gilani, O. Sosina, J. A. Kosmicki, L. Khrimian, Y.-Y. Fang, T. Persaud, V. Garcia, D. Sun, A. Li, J. Mbatchou, A. E. Locke, C. Benner, N. Verweij, N. Lin, S. Hossain, K. Agostinucci, J. V. Pascale, E. Dirice, M. Dunn, R. G. Center, Discov EHR Collaboration, W. E. Kraus, S. H. Shah, Y.-D. I. Chen, J. I. Rotter, D. J. Rader, O. Melander, C. D. Still, T. Mirshahi, D. J. Carey, J. Berumen-Campos, P. Kuri-Morales, J. Alegre-Díaz, J. M. Torres, J. R. Emberson, R. Collins, S. Balasubramanian, A. Hawes, M. Jones, B. Zambrowicz, A. J. Murphy, C. Paulding, G. Coppola, J. D. Overton, J. G. Reid, A. R. Shuldiner, M. Cantor, H. M. Kang, G. R. Abecasis, K. Karalis, A. N. Economides, J. Marchini, G. D. Yancopoulos, M. W. Sleeman, J. Altarejos, G. D. Gatta, R. Tapia-Conyer, M. L. Schwartzman, A. Baras, M. A. R. Ferreira, L. A. Lotta, Sequencing of 640,000 exomes identifies GPR75 variants associated with protection from obesity. *Science* **373**, eabf8683 (2021).
41. L. Duncan, H. Shen, B. Gelaye, J. Meijsen, K. Ressler, M. Feldman, R. Peterson, B. Domingue, Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).
42. R. Collins, M. K. Balaconis, S. Brunak, Z. Chen, M. De Silva, J. M. Gaziano, G. S. Ginsburg, P. Jha, P. Kuri, A. Metspalu, N. Mulder, N. Risch, Global priorities for large-scale biomarker-based prospective cohorts. *Cell Genom.* **2**, 100141 (2022).
43. Z. Chen, M. Smith, H. Du, Y. Guo, R. Clarke, Z. Bian, R. Collins, J. Chen, Y. Qian, X. Wang, X. Chen, X. Tian, X. Wang, R. Peto, L. Li, Blood pressure in relation to general and central adiposity among 500 000 adult Chinese men and women. *Int. J. Epidemiol.* **44**, 1305–1319 (2015).
44. L. Gnatiuc, J. Alegre-Díaz, J. Halsey, W. G. Herrington, M. López-Cervantes, S. Lewington, R. Collins, R. Tapia-Conyer, R. Peto, J. R. Emberson, P. Kuri-Morales, Adiposity and Blood Pressure in 110 000 Mexican Adults. *Hypertension* **69**, 608–614 (2017).
45. E. Riboli, R. Kaaks, The EPIC Project: Rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int. J. Epidemiol.* **26** (Suppl 1), S6–S14 (1997).
46. Z. Chen, J. Chen, R. Collins, Y. Guo, R. Peto, F. Wu, L. Li, China Kadoorie Biobank of 0.5 million people: Survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.* **40**, 1652–1666 (2011).
47. J. C. Denny, J. L. Rutter, D. B. Goldstein, A. Philippakis, J. W. Smoller, G. Jenkins, E. Dishman, The "All of Us" Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
48. K. M. Harrington, X.-M. T. Nguyen, R. J. Song, K. Hannagan, R. Quaden, D. R. Gagnon, K. Cho, J. E. Deen, S. Muralidhar, T. J. O'Leary, J. M. Gaziano, S. B. Whitbourne; VA Million Veteran Program, Gender differences in demographic and health characteristics of the Million Veteran Program cohort. *Womens Health Issues* **29** (Suppl 1), S56–S66 (2019).
49. T. Chikowore, A. B. Kamiza, O. H. Oduaran, T. Machipisa, S. Fatumo, Non-communicable diseases pandemic and precision medicine: Is Africa ready? *EBioMedicine* **65**, 103260 (2021).
50. P. Song, A. Gupta, I. Y. Goon, M. Hasan, S. Mahmood, R. Pradeepa, S. Siddiqui, G. S. Frost, D. Kusuma, M. Miraldo, F. Sassi, N. J. Wareham, S. Ahmed, R. M. Anjana, S. Brage, N. G. Forouhi, S. Jha, A. Kasturiratne, P. Katulanda, K. I. Khawaja, M. Loh, M. K. Mridha, A. R. Wickremasinghe, J. S. Kooner, J. C. Chambers, Data resource profile: Understanding the patterns and determinants of health in South Asians – The South Asia Biobank. *Int. J. Epidemiol.* **50**, 717–718e (2021).
51. T. J. Littlejohns, J. Holliday, L. M. Gibson, S. Garratt, N. Oesingmann, F. Alfaro-Almagro, J. D. Bell, C. Boultwood, R. Collins, M. C. Conroy, N. Crabtree, N. Doherty, A. F. Frangi, N. C. Harvey, P. Leeson, K. L. Miller, S. Neubauer, S. E. Petersen, J. Sellors, S. Sheard, S. M. Smith, C. L. M. Sudlow, P. M. Matthews, N. E. Allen, The UK Biobank imaging enhancement of 100,000 participants: Rationale, data collection, management and future directions. *Nat. Commun.* **11**, 2624 (2020).
52. P. Elliott, T. C. Peakman, The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int. J. Epidemiol.* **37**, 234–244 (2008).
53. N. E. Allen, M. Arnold, S. Parish, M. Hill, S. Sheard, H. Callen, D. Fry, S. Moffat, M. Gordon, S. Welsh, P. Elliott, R. Collins, 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.* **5**, 222 (2021).
54. R. Kaaks, E. Riboli, Validation and calibration of dietary intake measurements in the EPIC project: Methodological considerations. European Prospective Investigation into Cancer and Nutrition. *Int. J. Epidemiol.* **26** (Suppl. 1), S15–S25 (1997).
55. M. Pearce, T. Strain, Y. Kim, S. J. Sharp, K. Westgate, K. Wijndaele, T. Gonzales, N. J. Wareham, S. Brage, Estimating physical activity from self-reported behaviours in large-scale population studies using network harmonisation: Findings from UK Biobank and associations with disease outcomes. *Int. J. Behav. Nutr. Phys. Act.* **17**, 40 (2020).
56. D. Malden, B. Lacey, J. Emberson, F. Karpe, N. Allen, D. Bennett, S. Lewington, Body fat distribution and systolic blood pressure in 10,000 adults with whole-body imaging: UK Biobank and Oxford BioBank. *Obesity* **27**, 1200–1206 (2019).
57. Q. Feng, J. H. Kim, W. Omiyale, J. Bešević, M. Conroy, M. May, Z. Yang, S. Y. Wong, K. K. Tsoi, N. Allen, B. Lacey, Raw and cooked vegetable consumption and risk of cardiovascular disease: A study of 400,000 adults in UK Biobank. *Front. Nutr.* **9**, 831470 (2022).
58. M. K. Georgakis, R. Malik, D. Gill, N. Franceschini, C. L. M. Sudlow, M. Dichgans, Interleukin-6 Signaling Effects on Ischemic Stroke and Other Cardiovascular Outcomes: A Mendelian Randomization Study. *Circ. Genom. Precis. Med.* **13**, e002872 (2020).
59. S. C. Larsson, M. Bäck, J. M. B. Rees, A. M. Mason, S. Burgess, Body mass index and body composition in relation to 14 cardiovascular conditions in UK Biobank: A Mendelian randomization study. *Eur. Heart J.* **41**, 221–226 (2020).
60. G. Butler-Laporte, T. Nakanishi, V. Mooser, D. R. Morrison, T. Abdullah, O. Adeleye, N. Mamlouk, N. Kimchi, Z. Afrasiabi, N. Rezk, A. Giliberti, A. Renieri, Y. Chen, S. Zhou, V. Forgetta, J. B. Richards, Vitamin D and COVID-19 susceptibility and severity in the COVID-19 Host Genetics Initiative: A Mendelian randomization study. *PLOS Med.* **18**, e1003605 (2021).
61. X. Meng, X. Li, M. N. Timofeeva, Y. He, A. Spiliopoulou, W. Q. Wei, A. Gifford, H. Wu, T. Varley, P. Joshi, J. C. Denny, S. M. Farrington, L. Zgaga, M. G. Dunlop, P. McKeigue, H. Campbell, E. Theodoratou, Phenome-wide mendelian-randomization study of genetically determined vitamin D on multiple health outcomes using the UK Biobank study. *Int. J. Epidemiol.* **48**, 1425–1434 (2019).
62. G. D. Smith, Mendelian randomisation and vitamin D: The importance of model assumptions. *Lancet Diabetes Endocrinol.* **11**, 14 (2023).
63. S. Greenland, Multiple-bias modelling for analysis of observational data. *J. Royal. Stat. Soc. Ser. A* **168**, 267–306 (2005).
64. R. Clarke, M. Shipley, S. Lewington, L. Youngman, R. Collins, M. Marmot, R. Peto, Underestimation of risk associations due to regression dilution in long-term follow-up of prospective studies. *Am. J. Epidemiol.* **150**, 341–353 (1999).
65. S. MacMahon, R. Peto, J. Cutler, R. Collins, P. Sorlie, J. Neaton, R. Abbott, J. Godwin, A. Dyer, J. Stamler, Blood pressure, stroke, and coronary heart disease. Part 1, Prolonged differences in blood pressure: Prospective observational studies corrected for the regression dilution bias. *Lancet* **335**, 765–774 (1990).
66. A. N. Phillips, G. D. Smith, How independent are "independent" effects? Relative risk estimation when correlated exposures are measured imprecisely. *J. Clin. Epidemiol.* **44**, 1223–1231 (1991).

Downloaded from https://www.science.org at University of Connecticut on January 10, 2024

67. V. Codd, Q. Wang, E. Allara, C. Musicha, S. Kaptoge, S. Stoma, T. Jiang, S. E. Hamby, P. S. Braund, V. Bountziouka, C. A. Budgeon, M. Denniff, C. Swinfield, M. Papakonstantinou, S. Sheth, D. E. Nanus, S. C. Warner, M. Wang, A. V. Khera, J. Eales, W. H. Ouwehand, J. R. Thompson, E. D. Angelantonio, A. M. Wood, A. S. Butterworth, J. N. Danesh, C. P. Nelson, N. J. Samani, Polygenic basis and biomedical consequences of telomere length variation. *Nat. Genet.* **53**, 1425–1433 (2021).
68. C. E. Rutter, L. A. C. Millard, M. C. Borges, D. A. Lawlor, Exploring regression dilution bias using repeat measurements of 2858 variables in up to 49,000 UK Biobank participants. *Int. J. Epidemiol.* **52**, 1545–1556 (2022).
69. S. Tin Tin, G. K. Reeves, T. J. Key, Endogenous hormones and risk of invasive breast cancer in pre- and post-menopausal women: Findings from the UK Biobank. *Br. J. Cancer* **125**, 126–134 (2021).
70. K. A. Wartolowska, A. J. S. Webb, Midlife blood pressure is associated with the severity of white matter hyperintensities: Analysis of the UK Biobank cohort study. *Eur. Heart J.* **42**, 750–757 (2021).
71. M. J. Adams, W. D. Hill, D. M. Howard, H. S. Dashti, K. A. S. Davis, A. Campbell, T.-K. Clarke, I. J. Deary, C. Hayward, D. Porteous, M. Hotopf, A. M. M. Intosh, Factors associated with sharing e-mail information and mental health survey participation in large population cohorts. *Int. J. Epidemiol.* **49**, 410–421 (2020).
72. J. Beller, S. Geyer, J. Epping, Health and study dropout: Health aspects differentially predict attrition. *BMC Med. Res. Methodol.* **22**, 31 (2022).
73. A. E. Taylor, H. J. Jones, H. Sallis, J. Euesden, E. Stergiakouli, N. M. Davies, S. Zammit, D. A. Lawlor, M. R. Munafò, G. Davey Smith, K. Tilling, Exploring the association of genetic factors with participation in the Avon Longitudinal Study of Parents and Children. *Int. J. Epidemiol.* **47**, 1207–1216 (2018).
74. G. J. Griffith, T. T. Morris, M. J. Tudball, A. Herbert, G. Mancano, L. Pike, G. C. Sharp, J. Sterne, T. M. Palmer, G. D. Smith, K. Tilling, L. Zuccolo, N. M. Davies, G. Hemani, Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat. Commun.* **11**, 5749 (2020).
75. M. Chadeau-Hyam, B. Bodinier, J. Elliott, M. D. Whitaker, I. Tzoulaki, R. Vermeulen, M. Kelly-Irving, C. Delpierre, P. Elliott, Risk factors for positive and negative COVID-19 tests: A cautious and in-depth analysis of UK Biobank data. *Int. J. Epidemiol.* **49**, 1454–1467 (2020).
76. L. A. C. Millard, A. Fernández-Sanlés, A. R. Carter, R. A. Hughes, K. Tilling, T. P. Morris, D. Major-Smith, G. J. Griffith, G. L. Clayton, E. Kawabata, G. D. Smith, D. A. Lawlor, M. C. Borges, Exploring the impact of selection bias in observational studies of COVID-19: A simulation study. *Int. J. Epidemiol.* **52**, 44–57 (2023).
77. K. A. S. Davis, J. R. I. Coleman, M. Adams, N. Allen, G. Breen, B. Cullen, C. Dickens, E. Fox, N. Graham, J. Holliday, L. M. Howard, A. John, W. Lee, R. McCabe, A. McIntosh, R. Pearsall, D. J. Smith, C. Sudlow, J. Ward, S. Zammit, M. Hotopf, Mental health in UK Biobank - development, implementation and results from an online questionnaire completed by 157,366 participants: A reanalysis. *BJPsych Open* **6**, e18 (2020).
78. K. Rannikmäe, K. Ngoh, K. Bush, R. Al-Shahi Salman, F. Doubal, R. Flaig, D. E. Henshall, A. Hutchison, J. Nolan, S. Osborne, N. Samarasekera, C. Schnier, W. Whiteley, T. Wilkinson, K. Wilson, R. Woodfield, Q. Zhang, N. Allen, C. L. M. Sudlow, Accuracy of identifying incident stroke cases from linked health care data in UK Biobank. *Neurology* **95**, e697–e707 (2020).
79. B. Rubbo, N. K. Fitzpatrick, S. Denaxas, M. Daskalopoulou, N. Yu, R. S. Patel, H. Hemingway, Use of electronic health records to ascertain, validate and phenotype acute myocardial infarction: A systematic review and recommendations. *Int. J. Cardiol.* **187**, 705–711 (2015).
80. T. Wilkinson, C. Schnier, K. Bush, K. Rannikmae, D. E. Henshall, C. Lerpiniere, N. E. Allen, R. Flaig, T. C. Russ, D. Bathgate, S. Pal, J. T. O'Brien, C. L. M. Sudlow, Identifying dementia outcomes in UK Biobank: A validation study of primary care, hospital admissions and mortality data. *Eur. J. Epidemiol.* **34**, 557–565 (2019).
81. V. Kuan, S. Denaxas, A. Gonzalez-Izquierdo, K. Direk, O. Bhatti, S. Husain, S. Sutaria, M. Hingorani, D. Nitsch, C. A. Parisinos, R. T. Lumbers, R. Mathur, R. Sofat, J. P. Casas, I. C. K. Wong, H. Hemingway, A. D. Hingorani, A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *Lancet* **1**, e63–e77 (2019).
82. S. Lewington, R. Clarke, N. Qizilbash, R. Peto, R. Collins, Age-specific relevance of usual blood pressure to vascular mortality: A meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet* **360**, 1903–1913 (2002).
83. S. Lewington, G. Whitlock, R. Clarke, P. Sherliker, J. Emberson, J. Halsey, N. Qizilbash, R. Peto, R. Collins, Blood cholesterol and vascular mortality by age, sex, and blood pressure: A meta-analysis of individual data from 61 prospective studies with 55,000 vascular deaths. *Lancet* **370**, 1829–1839 (2007).
84. Neurology Working Group of the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, Stroke Genetics Network (SiGN), International Stroke Genetics Consortium (ISGC), Identification of additional risk loci for stroke and small vessel disease: A meta-analysis of genome-wide association studies. *Lancet Neurol.* **15**, 695–707 (2016).
85. W. Luo, L. Gong, X. Chen, R. Gao, B. Peng, Y. Wang, T. Luo, Y. Yang, B. Kang, C. Peng, L. Ma, M. Mei, Z. Liu, Q. Li, S. Yang, Z. Wang, J. Hu, Lifestyle and chronic kidney disease: A machine learning modeling study. *Front. Nutr.* **9**, 918576 (2022).
86. R. A. Shah, B. Asatryan, G. S. Dabbagh, N. Aung, M. Y. Khanji, L. R. Lopes, S. van Duijvenboden, A. Holmes, D. Muser, A. P. Landstrom, A. M. Lee, P. Arora, C. Semsarian, V. K. Somers, A. T. Owens, P. B. Munroe, S. E. Petersen, C. A. A. Chahal, Genotype-First Approach Investigators, Frequency, penetrance, and variable expressivity of dilated cardiomyopathy-associated putative pathogenic gene variants in UK Biobank participants. *Circulation* **146**, 110–124 (2022).
87. D. Chahal, D. Sharma, S. Keshavarzi, F. A. Q. Arisar, K. Patel, W. Xu, M. Bhat, Distinctive clinical and genetic features of lean vs overweight fatty liver disease using the UK Biobank. *Hepatol. Int.* **16**, 325–336 (2022).
88. N. J. Thomas, S. E. Jones, M. N. Weedon, B. M. Shields, R. A. Oram, A. T. Hattersley, Frequency and phenotype of type 1 diabetes in the first six decades of life: A cross-sectional, genetically stratified survival analysis from UK Biobank. *Lancet Diabetes Endocrinol.* **6**, 122–129 (2018).
89. P. R. Burton, A. L. Hansell, UK Biobank: The expected distribution of incident and prevalent cases of chronic disease and the statistical power of nested case-control studies (UK Biobank Technical Reports, 2005).
90. K. Papier, A. Knuppel, A. Perez-Cornago, E. L. Watts, T. Y. N. Tong, J. A. Schmidt, N. Allen, T. J. Key, R. C. Travis, Circulating insulin-like growth factor-I and risk of 25 common conditions: Outcome-wide analyses in the UK Biobank study. *Eur. J. Epidemiol.* **37**, 25–34 (2022).
91. S. Floud, R. F. Simpson, A. Balkwill, A. Brown, A. Goodill, J. Gallacher, C. Sudlow, P. Harris, A. Hofman, S. Parish, G. K. Reeves, J. Green, R. Peto, V. Beral, Body mass index, diet, physical inactivity, and the incidence of dementia in 1 million UK women. *Neurology* **94**, e123–e132 (2020).
92. T. Strain, K. Wijndaele, S. J. Sharp, P. C. Dempsey, N. Wareham, S. Brage, Impact of follow-up time and analytical approaches to account for reverse causality on the association between physical activity and health outcomes in UK Biobank. *Int. J. Epidemiol.* **49**, 162–172 (2020).
93. K. Bleicher, R. Summerhayes, S. Baynes, M. Swarbrick, T. Navin Cristina, H. Luc, G. Dawson, A. Cowle, X. Dolja-Gore, M. McNamara, Cohort Profile Update: The 45 and Up Study. *Int. J. Epidemiol.* **52**, e92–e101 (2023).
94. T. J. B. Dummer, P. Awadalla, C. Boileau, C. Craig, I. Fortier, V. Goel, J. M. T. Hicks, S. Jacquemont, B. M. Knoppers, N. Le, T. M. Donald, J. M. Laughlin, A.-M. Mes-Masson, A.-M. Nuyt, L. J. Palmer, L. Parker, M. Purdue, P. J. Robson, J. J. Spinelli, D. Thompson, J. Vena, M. Zawati, CPTP Regional Cohort Consortium, The Canadian Partnership for Tomorrow Project: A pan-Canadian platform for research on chronic disease prevention. *CMAJ* **190**, E710–E717 (2018).
95. H. S. Feigelson, C. L. Clarke, S. K. Van Den Eeden, S. Weinmann, A. N. Burnett-Hartman, S. Rowell, S. G. Scott, L. L. White, M. Ter-Minassian, S. A. A. Honda, D. R. Young, A. Kamineni, T. Chinn, A. Lituev, A. Bauck, E. A. M. Glynn, The Kaiser Permanente Research Bank Cancer Cohort: A collaborative resource to improve cancer care and survivorship. *BMC Cancer* **22**, 209 (2022).

Acknowledgments: We thank J. Mills and A. Motley for constructing the figures and G. D. Smith for helpful suggestions. Additional thanks to UK Biobank Access, the epidemiological and data analyst team, for the tireless work to help researchers access the data. We would particularly like to thank the 500,000 participants in the UK Biobank study for their enormous generosity and altruism and their continued interest, support, and involvement. **Funding:** UK Biobank has core funding from the MRC, Wellcome, British Heart Foundation, Cancer Research UK, and the National Institute for Health Research. **Competing interests:** All authors are past or present members of the UK Biobank Strategic Oversight Committee or the UK Biobank Senior Team. J.D. serves on scientific advisory boards for AstraZeneca and Novartis. R.C. is named on US patent no. 9957563B2 regarding a statin-related myopathy genetic test, but any share in royalty and other payments has been waived in favor of the Nuffield Department of Population Health, University of Oxford. R.E.M. is a scientific advisor to Optima Partners and the Epigenetic Clock Development Foundation and has received a speaker fee from Illumina. P.M.M. has received consultancy or speaker fees from Roche, Merck, Biogen, Rejuveron, Sangamo, Nodthera, Novartis, and Biogen. P.M.M. has received research or educational funds from Biogen, Novartis, Merck, and GlaxoSmithKline.

Submitted 24 October 2022
Accepted 13 December 2023
Published 10 January 2024
10.1126/scitranslmed.adf4428

Downloaded from https://www.science.org at University of Connecticut on January 10, 2024

Science Translational Medicine

Prospective study design and data analysis in UK Biobank

Naomi E. Allen, Ben Lacey, Deborah A. Lawlor, Jill P. Pell, John Gallacher, Liam Smeeth, Paul Elliott, Paul M. Matthews, Ronan A. Lyons, Anthony D. Whetton, Anneke Lucassen, Matthew E. Hurles, Michael Chapman, Andrew W. Roddam, Natalie K. Fitzpatrick, Anna L. Hansell, Rebecca Hardy, Riccardo E. Marioni, Valerie B. O'Donnell, Julie Williams, Cecilia M. Lindgren, Mark Effingham, Jonathan Sellors, John Danesh, and Rory Collins

Sci. Transl. Med. **16** (729), eadf4428. DOI: 10.1126/scitranslmed.adf4428

View the article online
https://www.science.org/doi/10.1126/scitranslmed.adf4428
Permissions
https://www.science.org/help/reprints-and-permissions

Downloaded from https://www.science.org at University of Connecticut on January 10, 2024

Use of this article is subject to the Terms of service

Science Translational Medicine (ISSN 1946-6242) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Translational Medicine* is a registered trademark of AAAS.

Copyright © 2024 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works