UK Biobank prospective cohort design and analytical considerations

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]

[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

[8] London School of Hygiene and Tropical Medicine, London, UK

[9] MRC Centre for Environment and Health, School of Public Health, School of Public Health, Imperial College London, London, UK

[10] NIHR Biomedical Research Centre, 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, 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 Digital, 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 Emails:

naomi.allen@ndph.ox.ac.uk

Overline: Biobanks

One sentence summary: This article describes approaches to study design, resource access and data analysis in UK Biobank to facilitate health-related research

Abstract

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

Introduction

Population health research has come a long way in the last 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 as they overcome many of the biases inherent in case-control studies (most notably that exposures (i.e. risk factors for disease) are measured prior to disease onset). However, until recently, the conclusions that could be drawn from such studies were limited by small sample size, varying analytical approaches taken to define 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 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 (GWAS) that allowed hypothesis-free identification across the genome of variants associated with a particular phenotype. Much effort was typically spent on characterising 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 that 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-characterised, 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 and environmental determinants of a wide range of diseases of middle-to-old age (www.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 (BHF), Cancer Research UK (CR-UK) and National Institute for Health and Care Research (NIHR). 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 (Supplementary Table 1).

UK Biobank was designed to promote innovative science by maximising access to the data in an equitable and transparent manner. All approved researchers (academic or commercial) can access all of the de-identified data in order 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 (for example, 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 cohort 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 wave-form data, and electronic health records) are generated and analysed 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 30,000 registered researchers (80% from outside the UK) and about 9,000 publications (attracting 270,000 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) and 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 non-specialists 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 towards 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 democratise access further, UK Biobank launched a cloud-based Research Analysis Platform in 2021 that allows streamlined access for researchers worldwide (in particular 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, as well as 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 under-estimation 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 (in particular, non-genetic ones), especially for conditions with a long prodromal phase, such as Alzheimer's disease.

UK Biobank has been set up to help minimise random and systematic error so that it can support reliable research into the determinants of disease (Supplementary Table 1), 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, and 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 40-69 years at recruitment 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 treatment, or by changes in an individual's behavior following 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 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, as 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 organisational constraints). For example, after a median follow-up of 12 years (i.e. by end-2020), linkage to electronic healthcare 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 10,000 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 5,000-10,000 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 joint effect of blood pressure and age on the risk of coronary heart disease, the standard error 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 over 50,000 cases of diabetes and chronic obstructive pulmonary disease. 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 that of the wider population

In UK Biobank, the well-defined sampling frame means that it is possible to assess not just 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 40–69, who were registered with the UK's National Health Service (NHS) and lived within a short travelling time (typically about 25 miles) of one 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 socio-economically deprived areas). During 2006-2010, 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 risk factors (i.e. that vary from low to high exposure levels of a wide range of potential risk factors).

It is this heterogeneity across different levels of exposures (e.g., genetic, lifestyle, sociodemographic and environmental exposures) - and not the relatively low overall participation rate - that largely determines the generalisability of the findings to the broader UK population (25). For example, although individuals from more socio-economic deprived areas are under-represented 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 of socio-economic deprivation with disease risk. By contrast, although UK Biobank is reasonably representative of the distribution for different ethnic groups, with 29,000 participants recruited from Black and other ethnic minority groups (which was about the same proportion, ~5%, as the rest of the UK population at the time) (26), it is insufficient to examine reliably the differences in exposure-disease associations by ethnicity. Even though 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 (for example, 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 used widely. However, one research group found that standardisation 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 non-smokers and to live in more affluent areas than the general population in the same age range. Given that area-level socio-economic deprivation is moderately inversely correlated both with participation in UK Biobank and lung cancer, this non-representativeness may attenuate the observed association of smoking with lung cancer if the effects of smoking and socio-economic deprivation are not independent or synergistic (31). Likewise, UK Biobank participants were more likely to use 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 centre 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 between exposures, potential confounders, mediators, and outcomes, and 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; 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) – in particular, studies considered to be representative of the underlying population – suggest 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 UK, US and Mexico cohorts (40).

Nonetheless, there may be circumstances in which associations between an exposure and disease risk varies 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 non-genetic 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 risk factors that can now be investigated. New large-scale prospective studies are now established in the US 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., Non-communicable 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 non-genetic) 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 generalisability 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 significant 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 and thereby be able to correct for regression-dilution bias. However, there is 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 measurement

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 composition, dynamometer to measure hand grip strength, etc.) have been used (Fig. 1) to collect data that are reliable, valid and of high scientific value (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, occupation), ethnicity, family history, lifestyle (diet, alcohol consumption, smoking history, physical activity, sleep, sun exposure, sexual history), early life factors, psychosocial factors, medical history, cognition and environmental exposures were all collected via the touch-screen questionnaire in about fifty minutes.

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, 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 cardio-respiratory fitness test with a 4-lead electrocardiogram (ECG) (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 2,500 undergoing a repeat assessment), a multi-modal 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, occupational history).

Rigorous approaches have also been taken to sample collection, processing, retrieval and assay measurement. Prior to the start of 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 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, 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 day (53). Consequently, when assaying samples from participants in this quasi-random order, the mean biomarker concentration across batches and analysers should be constant, which allows correction for variation caused by laboratory drift. Throughout the assay process, the

data are reviewed to identify issues and either address them in real time (e.g., if specific batches require re-measurement) or make any adjustments retrospectively. For example, following 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, the use of highly efficient assay methods minimises sample depletion (with currently less than 10% of the baseline blood sample used so far), which will allow other types of assays (e.g., epigenetics, transcriptomics amd 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 was 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 in over 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, as 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 socio-economic 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 socio-economic factors (such as education, occupation and income via linkage to administrative datasets or specific web-based questionnaires) would enable more precise characterisation 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 ($LR\chi^2$) test after the addition to the model of covariates is indicative that the association likely remains affected by residual confounding, as 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 Randomisation – facilitated in analyses of UK Biobank by the extensive genetic information available on all of the 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 development of cardiovascular conditions (58, 59). Conversely, Mendelian Randomisation has not provided support for a protective effect of vitamin D against COVID-19 (60), cancer or cardiovascular outcomes (61), although it should be noted that Mendelian Randomisation analyses may also be affected by bias in some circumstances (62). When associations of genetic variants with the relevant non-genetic risk factors are weak (such that Mendelian Randomisation 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 (for example, 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), as well as 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. Re-measures 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 over-estimation or under-estimation 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 in order to 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, and 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.

Reliable assessment of a wide range of health outcomes

To minimise 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 characterise 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 non-specialists will further increase the usability and reproducibility of research using these data.

Comprehensive ascertainment of health outcomes

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 US and most low-to-middle income countries), the UK's National Health Service (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 minimising ascertainment bias and reducing loss-to-follow-up or attrition bias by providing cohort-wide follow-up information without the need for active participant re-contact, 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 minimises systematic bias associated with loss to follow-up from non-random subgroups of the cohort.

To date, UK Biobank has linked NHS healthcare data from centralised national cancer and death registries and hospital inpatient admissions for all participants. In contrast, primary care data are not centralised 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 prior to analysis. Incorporation of primary care data for all 500,000 participants for all types of health-related research would be of enormous value as 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, whereas 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 chronic obstructive pulmonary disease (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 data.

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 research (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 are 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 to increase the precision and specificity of those diagnoses. The aim is to avoid a situation where insufficient data on health outcomes leads to misclassification of cases and non-cases, 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 minimising the number of false-positive cases (i.e., achieving a high positive predictive value). It is worth recognising that it is not necessary to identify all cases of a disease as false negatives will be diluted by the much larger number of 'true' controls (and so 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), and neurodevelopmental and gastrointestinal conditions.

It is also important to characterise disease subtypes as 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 inter-operable code lists from electronic healthcare 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, while blood pressure is strongly

positively associated with the risk of both ischaemic and haemorrhagic stroke (82), the association of cholesterol and certain genetic variants with stroke differ substantially by subtype (83, 84) providing clues to the underlying aetiology 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 sub-types 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 sub-types, digitised histopathology slides to determine tumour morphological subtypes, and in-person assessments to characterise dementia subtypes.

It is possible to identify some disease sub-types 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 non-alcoholic 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. While 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 as 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 five to ten 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 minimising 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 five 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 10-15 years of follow-up (91). Consequently, researchers should consider the impact of exclusion of participants with prevalent disease prior to 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 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 healthcare datasets and environmental sources, etc.) help to ensure that the resource fulfils 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 provides a powerful resource to help address previously unanswerable questions of the determinants of incident disease, as well as enabling research into risk prediction and identification of early biomarkers of disease. Whereas the UK Biobank study has attempted to minimise 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, 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 provide a level of transparency in the interpretation of findings that has, until now, generally been under-reported.

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. Taken together, they will enable a greater range of risk factors and diseases to be analysed 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.

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, Lipoprotein(a) concentration and the risk of coronary heart disease, stroke, and nonvascular mortality, *JAMA*, **302**, 412-423 (2009).

(4) Emerging Risk Factors Collaboration, 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, 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 Diabetologica*, **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. Dahlqvist 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. McCarthy, 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. *Nature Med*, **26,** 1654-1662 (2020).

(18) O. Canela-Xandri, K. Rawlik, A. Tenesa, An atlas of genetic associations in UK Biobank. *Nature 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 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, 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).

766 (31) J. Tyrrell, J. Zheng, R. Beaumont, K. Hinton, T.G. Richardson, A.R. Wood, G. Davey
767 Smith, T.M. Frayling, K. Tilling, Genetic predictors of participation in optional components of
768 UK Biobank. *Nature Comms*, **12,** 886 (2021).
769
770 (32) K. Suissa, M. Hudson, S. Suissa, Glucosamine and lower mortality and cancer incidence:
771 Selection bias in the observational studies, Pharmacoepidemiology Drug Saf, **31,** 1272-1279
772 (2022).
773
774 (33) S. Haworth, R. Mitchell, L. Corbin, K.H. Wade, T. Dudding, A. Budu-Aggrey, D. Carslake,
775 G. Hemani, L. Paternoster, G.D. Smith, N. Davies, D.J. Lawson, J.T. N, Apparent latent
776 structure within the UK Biobank sample has implications for epidemiological analysis. *Nature*
777 *Comms*, **10**, 333 (2019).
778
779 (34) T.L. Lash, M.P. Fox, R.F. MacLehose, G. Maldonado, L.C. McCandless, S. Greenland,
780 Good practices for quantitative bias analysis. *Int J Epidemiol,* **43,** 1969-1985 (2014).
781
782 (35) M.R. Munafò, K. Tilling, A.E. Taylor, D.M. Evans, G. Davey Smith, Collider scope: when
783 selection bias can substantially influence observed associations. *Int J Epidemiol*, **47**, 226-235
784 (2018).
785
786 (36) Emerging Risk Factors Collaboration, Association of Cardiometabolic Multimorbidity With
787 Mortality, *JAMA*, **314**, 52-60 (2015).
788
789 (37) H.S. Dashti, S.E. Jones, A.R. Wood, J.M. Lane, V.T. van Hees, H. Wang, J.A. Rhodes,
790 Y. Song, K. Patel, S.G. Anderson, R.N. Beaumont, D.A. Bechtold, J. Bowden, B.E. Cade, M.
791 Garaulet, S.D. Kyle, M.A. Little, A.S. Loudon, A.I. Luik, F. Scheer, K. Spiegelhalder, J. Tyrrell,
792 D.J. Gottlieb, H. Tiemeier, D.W. Ray, S.M. Purcell, T.M. Frayling, S. Redline, D.A. Lawlor,
793 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, *Nature Comms*, **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. *Nature Comm*, **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. *Nature 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.

822 Agostinucci, J.V. Pascale, E. Dirice, M. Dunn, W.E. Kraus, S.H. Shah, Y.I. Chen, J.I. Rotter,
823 D.J. Rader, O. Melander, C.D. Still, T. Mirshahi, D.J. Carey, J. Berumen-Campos, P. Kuri-
824 Morales, J. Alegre-Díaz, J.M. Torres, J.R. Emberson, R. Collins, S. Balasubramanian, A.
825 Hawes, M. Jones, B. Zambrowicz, A.J. Murphy, C. Paulding, G. Coppola, J.D. Overton, J.G.
826 Reid, A.R. Shuldiner, M. Cantor, H.M. Kang, G.R. Abecasis, K. Karalis, A.N. Economides, J.
827 Marchini, G.D. Yancopoulos, M.W. Sleeman, J. Altarejos, G. Della Gatta, R. Tapia-Conyer,
828 M.L. Schwartzman, A. Baras, M.A.R. Ferreira, L.A. Lotta, Sequencing of 640,000 exomes
829 identifies GPR75 variants associated with protection from obesity. *Science,* **373**, eabf8683
830 (2021).
831
832 (41) L. Duncan, H. Shen, B. Gelaye, J. Meijsen, K. Ressler, M. Feldman, R. Peterson, B.
833 Domingue, Analysis of polygenic risk score usage and performance in diverse human
834 populations, *Nature Comms*, **10**, 3328 (2019).
835
836 (42) R. Collins, M.K. Balaconis, S. Brunak, Z. Chen, M. De Silva, J.M. Gaziano, G.S. Ginsburg,
837 P. Jha, P. Kuri, A. Metspalu, N. Mulder, N. Risch, Global priorities for large-scale biomarker-
838 based prospective cohorts, *Cell Genomics*, **2**, 100141 (2022).
839
840 (43) Z. Chen, M. Smith, H. Du, Y. Guo, R. Clarke, Z. Bian, R. Collins, J. Chen, Y. Qian, X.
841 Wang, X. Chen, X. Tian, X. Wang, R. Peto, L. Li, Blood pressure in relation to general and
842 central adiposity among 500 000 adult Chinese men and women. *Int J Epidemiol*, **44**, 1305-
843 1319 (2015).
844
845 (44) L. Gnatiuc, J. Alegre-Díaz, J. Halsey, W.G. Herrington, M. López-Cervantes, S.
846 Lewington, R. Collins, R. Tapia-Conyer, R. Peto, J.R. Emberson, P. Kuri-Morales, Adiposity
847 and Blood Pressure in 110 000 Mexican Adults. *Hypertension,* **69,** 608-614 (2017).
848

(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, S6-14 (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. *New Engl J Med*, **381,** 668-676 (2019).

(48) K.M. Harrington, X.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, Gender Differences in Demographic and Health Characteristics of the Million Veteran Program Cohort. *Womens Health Issues,* **29** Suppl 1, S56-66 (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.

877 Harvey, P. Leeson, K.L. Miller, S. Neubauer, S.E. Petersen, J. Sellors, S. Sheard, S.M. Smith,
878 C.L.M. Sudlow, P.M. Matthews, N.E. Allen, The UK Biobank imaging enhancement of 100,000
879 participants: rationale, data collection, management and future directions. *Nature Comms*, **11**,
880 2624 (2020).
881
882 (52) P. Elliott, T.C. Peakman, The UK Biobank sample handling and storage protocol for the
883 collection, processing and archiving of human blood and urine. *Int J Epidemiol*, **37,** 234-244
884 (2008).
885
886 (53) N.E. Allen, M. Arnold, S. Parish, M. Hill, S. Sheard, H. Callen, D. Fry, S. Moffat, M.
887 Gordon, S. Welsh, P. Elliott, R. Collins, Approaches to minimising the epidemiological impact
888 of sources of systematic and random variation that may affect biochemistry assay data in UK
889 Biobank. *Wellcome Open Res*, **5,** 222 (2021).
890
891 (54) R. Kaaks, E. Riboli, Validation and calibration of dietary intake measurements in the EPIC
892 project: methodological considerations. European Prospective Investigation into Cancer and
893 Nutrition, *Int J Epidemiol*, **26**, S15-25 (1997).
894
895 (55) M. Pearce, T. Strain, Y. Kim, S.J. Sharp, K. Westgate, K. Wijndaele, T. Gonzales, N.J.
896 Wareham, S. Brage, Estimating physical activity from self-reported behaviours in large-scale
897 population studies using network harmonisation: findings from UK Biobank and associations
898 with disease outcomes. *Int J Behav Nutr Phys Act*, **17**, 40 (2020).
899
900 (56) D. Malden, B. Lacey, J. Emberson, F. Karpe, N. Allen, D. Bennett, S. Lewington, Body
901 Fat Distribution and Systolic Blood Pressure in 10,000 Adults with Whole-Body Imaging: UK
902 Biobank and Oxford BioBank. *Obesity*, **27**, 1200-1206 (2019).
903

(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. *Frontiers in 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. *Europ 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: Series 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).

(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. Di 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. *Nature 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. *Europ 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. McIntosh, 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. Davey Smith, K. Tilling, L. Zuccolo, N.M. Davies, G. Hemani, Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nature Comms*, **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. Davey 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**, 705-11 (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) CHARGE and ISGC Consortium, 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, *Frontiers Nutr*, **9,** 918576 (2022).

(86) R.A. Shah, B. Asatryan, G. Sharaf 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, 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. *Europ 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 Epiemiol*, **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. McDonald, J. McLaughlin, 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, The Canadian Partnership for Tomorrow Project: a pan-Canadian platform for research on chronic disease prevention. *CMAJ*, **190,** e710-717 (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. McGlynn, The Kaiser Permanente Research Bank Cancer Cohort: a collaborative resource to improve cancer care and survivorship. *BMC Cancer*, **22**, 209 (2022).

Acknowledgements

We thank Jenny Mills and Alicia Motley for constructing the figures and George Davey Smith for helpful suggestions. Additional thanks to the UK Biobank Access team for their tireless

work on research registrations, applications and output. The authors would 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 Medical Research Council, Wellcome, British Heart Foundation, Cancer Research UK and 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 and consults; British Heart Foundation Centre of Research Excellence, University of Cambridge; the National Institute for Health and Care Research Blood and Transplant Research Unit in Donor Health and Behaviour, University of Cambridge; Health Data Research UK; Wellcome Genome Campus and University of Cambridge; Department of Human Genetics, Wellcome Sanger Institute, Hinxton, UK. R.C. is named on US patent #9957563B2 regarding a statin-related myopathy genetic test but any share in royalty and other payments has been waived in favour of the Nuffield Department of Population Health, University of Oxford. Financial Relationships: UK Biobank Consortium Funding and Enhancements (Novartis, Regeneron Pharmaceuticals, Merck, AstraZeneca). 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 has received consultancy or speaker fees from Roche, Merck, Biogen, Rejuveron, Sangamo, Nodthera, Novartis and Biogen. P.M. has received research or educational funds from Biogen, Novartis, Merck and GlaxoSmithKline.

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

Condition	Year of diagnosis		
	Observed[1]	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

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

Table 2. Sampling characteristics of selected general population prospective studies with at least 250,000 participants, containing genomic, behavioural and health outcome data[1]

Study name	Recruitment dates (range)	Location	Sample size	Sex; Age at recruitment	Population from which the sample was recruited	Participation rate
23andMe (www.23andme.com)	2007 - present	Global (but mainly USA)	6.8M	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	USA	Ongoing. Aim: 1M	MF; 18+	Varied approaches, many of which are targeted at underrepresented 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 7 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 centres located in 10 European countries recruited through direct invitations	not known
Kaiser Permanente Research Bank (95)	2007 – 2013	USA	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	USA	Ongoing. Aim: 1M	MF; 18+	Members of the Veterans Health Administration System recruited through direct invitations and indirect (promotional materials) methods	14%
UK Biobank (26)	2006 - 2010	UK	500,000	MF: 40-69	Residents living close to 22 assessment centres in the UK, recruited via direct invitations	5.5%

[1] The IHCC (https://ihccglobal.org/) has details of other prospective studies of less than 250,000 participants

Fig. 1. Illustration of the types of data collected in UK Biobank, which includes data collected at in-person assessments (e.g. lifestyle factors, medical history, blood pressure and other physical measures, imaging scans), data from online questionnaires, data generated from biological samples and data from electronic healthcare records over time

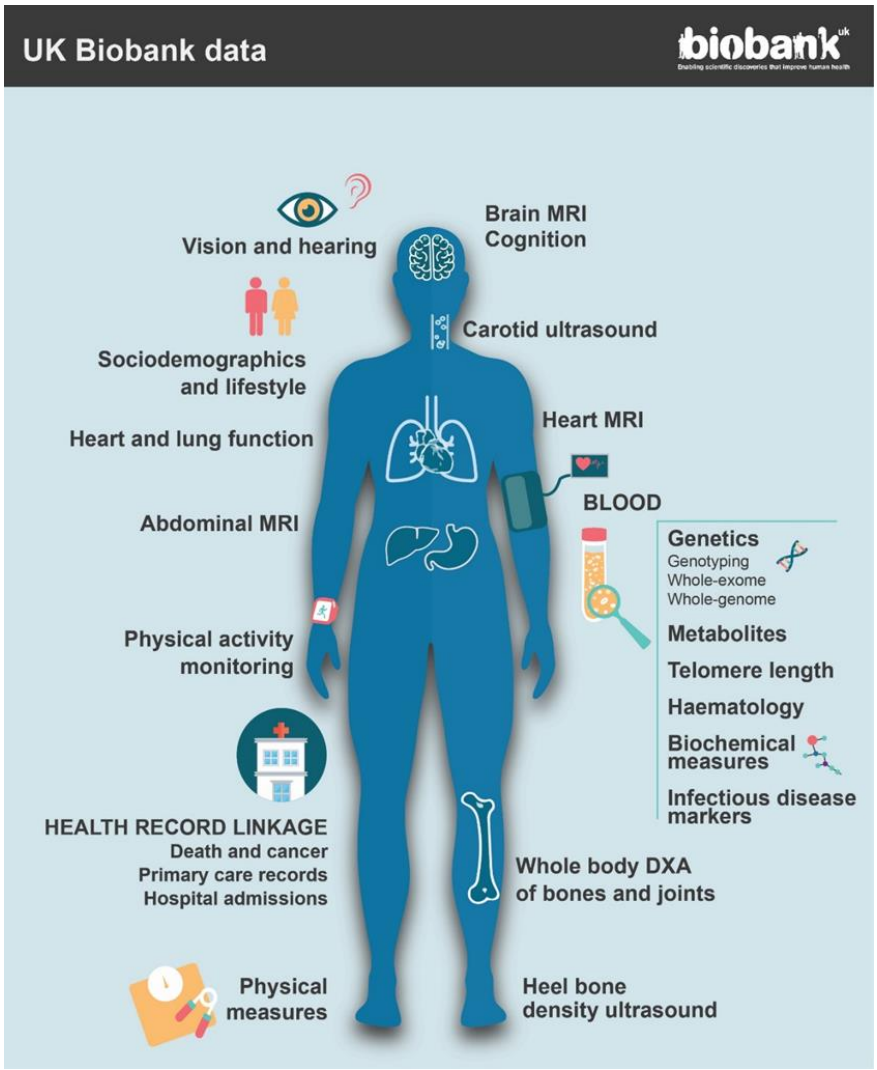
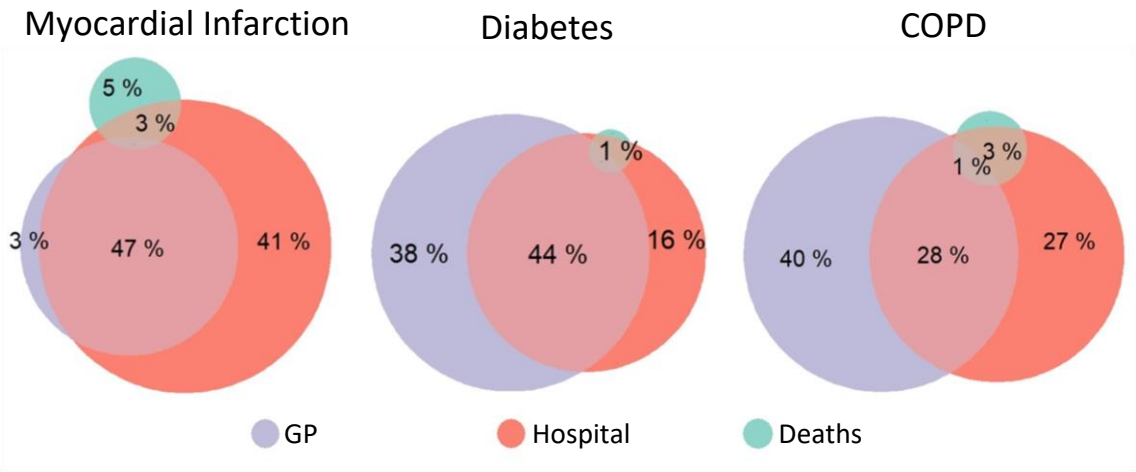


Fig. 2. The proportion of incident cases (i.e. ascertained since recruitment into the study) identified through hospital inpatient admissions, primary care and death data for some common exemplar conditions (myocardial infarction, diabetes and chronic obstructive pulmonary disease)



Supplementary Table 1. UK Biobank's design approach and potential analytical approaches to reduce error in exposure-disease associations

Design characteristic	Rationale and error to be addressed	UK Biobank's design approach	Potential analytical approaches
Large-scale prospective design	Exposure data collected prior to disease (i.e. to reduce recall bias and reverse causation bias)	Exposures measured at recruitment and participants' health followed up over time via linkage to longitudinal healthcare records	Perform longitudinal analyses to determine exposure-disease associations
	Large numbers of participants are needed to provide sufficient statistical power for reliable assessment of risk factors with health outcomes (i.e. to reduce random error)	Recruitment of 500,000 participants	Pool data and /or results with other studies to increase sample size
Participation rate and comparing cohort characteristics with that of the wider population	Comparison of results from UK Biobank with studies in other populations needed to determine generalizability of research findings	Postal invites sent to all 9.2M individuals age 40-69 years and living within travelling distance of an assessment centre	Control for factors associated with study participation and retention Use Directed Acyclic Graphs (DAGs) to investigate the underlying

	The study population should be sufficiently heterogeneous to include a wide range of risk factors under investigation to enable generalizable assessments of exposure-disease associations High participant engagement needed to obtain high response rates for continued data collection activities (i.e. to reduce systematic error due to responder bias)	Recruited participants with widely varying risk factor levels[1]	assumptions and to identify potential sources of bias of the association(s) Perform sensitivity analyses to assess impact of missing data resulting from non-random participation Compare results with studies from different populations, including meta-analytical approaches that assess the impact of UK Biobank data on the overall findings
Reliable assessment of a wide range of exposures:			
Depth and breadth of exposure measurement	Comprehensive characterisation of participants' behavioural, environment and germline genome are needed to identify independent risk factors for disease (i.e. to reduce confounding)	Comprehensive (i.e. cohort-wide) assessment of exposures (incl. genomics and other biomarkers) to reduce missing values of variables	Use of DAGs to clarify the presence and direction of potential confounders and mediators Adjust for multiple relevant factors

	Data on exposures need to be complete and accurate to improve precision (i.e. reduce random and systematic error).	Standardised data collection protocol used to ensure data were collected accurately and in a consistent manner Sample assays performed in the full cohort at the same time facilitate quality control Supplemented crude measures with detailed objective assessments (e.g. accelerometer to assess physical activity)	Use of genetic causal inference models Use different analytical approaches to triangulate evidence Use of simulations (e.g. probabilistic bias analysis) to assess likely impact of measurement error of the exposure and confounder(s) on the risk estimate Calibrate variables for the full cohort based on more precise measures performed in a subset
Repeated exposure measures	Repeated exposure measures are needed to enable accurate assessment of long-term average exposure (i.e. to reduce regression dilution bias).	Repeat assessments performed in subsets of the cohorts	Correct for regression dilution bias using repeated measures
Reliable assessment of a wide range of health outcomes:			

Comprehensive ascertainment of health outcomes	Passive cohort-wide collection of health outcomes is needed to minimise ascertainment bias and reduce loss-to-follow-up (or attrition) bias.	Cohort-wide linkage to routine administrative health records	Control for factors associated with differential ascertainment of health outcomes that are based on participant characteristics
Specificity of health outcomes	Accurate ascertainment of outcomes (and their subtypes) needed to increase their specificity and positive predictive value (i.e. reduce random error associated with false-positives)	Detailed ascertainment using diagnostic codes and other biochemical, imaging data etc.	Perform subgroup analyses by disease sub-type, where these data are available Develop research agenda to implement novel approaches for accurate disease classification
Long duration of follow-up	Long-term follow-up of participants' health needed to enable assessment of temporality of associations (i.e. reduce reverse causation bias) and to accrue large numbers of incident disease (i.e. reduce random error)	Linkage to routine administrative health records since recruitment: ~15 years complete follow-up	Consider impact of exclusion/inclusion of prevalent disease cases on analysis Perform sensitivity analyses that exclude initial follow-up periods

[1] UK Biobank does not cover the full range of racial/ethnic diversity (given the study is based in the UK) or age (the study recruited individuals aged 40-69 years).