UK Biobank prospective cohort design and analytical considerations

Naomi E. Allen1,2, Ben Lacey1,2, Deborah A. Lawlor3,4, Jill P. Pell5, John Gallacher6,7, Liam Smeeth8, Paul Elliott9,10, Paul M. Matthews12, Ronan A. Lyons13, Anthony D. Whetton14, Anneke Lucassen15,16, Matthew E. Hurles17, Michael Chapman18, Andrew W. Roddam19, Natalie K. Fitzpatrick20, Anna L. Hansell21, Rebecca Hardy22, Riccardo E. Marioni23, Valerie B. O'Donnell24, Julie Williams25, Cecilia M. Lindgren26, Mark Effingham1, Jonathan Sellors1, John Danesh27,28,29, Rory Collins1,2

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, 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.
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](http://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
costentious 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 and proteomics) to be conducted on a cohort-wide basis when technological advances make this possible.

The collection of different types of data that describe the same (or highly related) exposures can also contribute to accuracy. In particular, a more precise assessment performed in a subset of participants could be used to correct for any random and systematic error inherent in the less precise baseline measures conducted in the full cohort (54). For example, data from an accelerometer device worn by 100,000 UK Biobank participants 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 ($\text{LR}\chi^2$) 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 $\chi^2$ 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


5. Collaborative Group on Hormonal Factors in Breast Cancer, Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of


(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

(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

(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,

(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

(16) F. Alfaro-Almagro, M. Jenkinson, N.K. Bangerter, J.L.R. Andersson, L. Griffanti, G.
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

(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
(2020).

(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 

(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. 

(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 

(23) S. Lewington, personal correspondence (2022).

Allen, Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank 

(25) K.J. Rothman, J.E. Gallacher, E.E. Hatch, Why representativeness should be avoided. 


33


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

(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

(78) K. Rannikmäe, K. Ngoh, K. Bush, R. Al-Shahi Salman, F. Doubal, R. Flaig, D.E. Henshall,
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,

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

(81) V. Kuan, S. Denaxas, A. Gonzalez-Izquierdo, K. Direk, O. Bhatti, S. Husain, S. Sutaria,


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
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

<table>
<thead>
<tr>
<th>Condition</th>
<th>Observed</th>
<th>Year of diagnosis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2020</td>
</tr>
<tr>
<td>Diabetes</td>
<td>31,000</td>
<td>54,000</td>
</tr>
<tr>
<td>Myocardial infarction</td>
<td>15,000</td>
<td>30,000</td>
</tr>
<tr>
<td>Stroke</td>
<td>12,000</td>
<td>25,000</td>
</tr>
<tr>
<td>COPD</td>
<td>25,000</td>
<td>47,000</td>
</tr>
<tr>
<td>Depression</td>
<td>25,000</td>
<td>39,000</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>9,000</td>
<td>14,000</td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td>5,000</td>
<td>8,000</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>4,000</td>
<td>6,000</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>10,000</td>
<td>16,000</td>
</tr>
<tr>
<td>Hip fracture</td>
<td>5,000</td>
<td>13,000</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>4,000</td>
<td>6,000</td>
</tr>
<tr>
<td>Alzheimer’s disease</td>
<td>5,000</td>
<td>17,000</td>
</tr>
<tr>
<td>Parkinson’s disease</td>
<td>4,000</td>
<td>10,000</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Study name</th>
<th>Recruitment dates (range)</th>
<th>Location</th>
<th>Sample size</th>
<th>Sex; Age at recruitment</th>
<th>Population from which the sample was recruited</th>
<th>Participation rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>23andMe (<a href="http://www.23andme.com">www.23andme.com</a>)</td>
<td>2007 - present</td>
<td>Global (but mainly USA)</td>
<td>6.8M</td>
<td>MF; 13+</td>
<td>Customers of a personal genetics company</td>
<td>not known</td>
</tr>
<tr>
<td>45 and Up (93)</td>
<td>2006 - 2009</td>
<td>Australia</td>
<td>267,000</td>
<td>MF; 45+</td>
<td>New South Wales residents enrolled in Medicare, recruited through direct invitations</td>
<td>19%</td>
</tr>
<tr>
<td>All of Us (47)</td>
<td>2018 - present</td>
<td>USA</td>
<td>Ongoing.</td>
<td>MF; 18+</td>
<td>Varied approaches, many of which are targeted at underrepresented groups via direct and indirect means</td>
<td>not known</td>
</tr>
<tr>
<td>Canadian Partnership for Tomorrow’s Health (CanPath) (94)</td>
<td>2008 - present</td>
<td>Canada</td>
<td>330,000</td>
<td>MF; 30-74</td>
<td>Residents across Canada recruited into 7 regional cohorts using varied approaches</td>
<td>not known</td>
</tr>
<tr>
<td>Study Name</td>
<td>Years</td>
<td>Country</td>
<td>Participants</td>
<td>Gender Age</td>
<td>Participants Description</td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------------------------------------</td>
<td>-------------</td>
<td>---------</td>
<td>--------------</td>
<td>------------</td>
<td>------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>China Kadoorie Biobank (46)</td>
<td>2004 - 2008</td>
<td>China</td>
<td>510,000</td>
<td>MF; 30-70</td>
<td>Residents of 10 geographically defined regions across China, recruited through direct invitations</td>
<td></td>
</tr>
<tr>
<td>European Prospective Investigation into Cancer, Chronic Diseases, Nutrition and Lifestyle (EPIC) (45)</td>
<td>1992 - 2000</td>
<td>10</td>
<td>520,000</td>
<td>MF; 35-70</td>
<td>Residents from 23 centres located in 10 European countries, recruited through direct invitations</td>
<td></td>
</tr>
<tr>
<td>Kaiser Permanente Research Bank (95)</td>
<td>2007 - 2013</td>
<td>USA</td>
<td>400,000</td>
<td>MF; 18+</td>
<td>Members of Kaiser Permanente health plan recruited through direct invitations, in-person communication and via website, 20-50% of each areas' insured population</td>
<td></td>
</tr>
<tr>
<td>Million Veterans Program (48)</td>
<td>2011 - present</td>
<td>USA</td>
<td>Ongoing.</td>
<td>MF; 18+</td>
<td>Members of the Veterans Health Administration System recruited through direct invitations and indirect (promotional materials) methods</td>
<td></td>
</tr>
<tr>
<td>UK Biobank (26)</td>
<td>2006 - 2010</td>
<td>UK</td>
<td>500,000</td>
<td>MF; 40-69</td>
<td>Residents living close to 22 assessment centres in the UK, recruited via direct invitations</td>
<td></td>
</tr>
</tbody>
</table>

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

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).