

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/165532/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Allen, Naoimi, Lacey, Ben, Lawlor, Deborah, Pell, Jill, Gallacher, John, Smeeth, Liam, Elliott, Paul, Matthews, Paul, Lyons, Ronan, Whetton, Anthony, Lucassen, Anneke, Hurles, Matthew, Chapman, Michael, Roddam, Andrew, Fitzpatrick, Natalie, Hansell, Anna, Hardy, Rebecca, Marioni, Riccardo, O'Donnell, Valerie, Williams, Julie, Lindgren, Cecilia, Effingham, Mark, Sellors, Jonathan, Danesh, John and Collins, Rory 2024. Prospective study design and data analysis in UK Biobank. *Science Translational Medicine* 16 (729), eadf4428. 10.1126/scitranslmed.adf4428

Publishers page: <https://doi.org/10.1126/scitranslmed.adf4428>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



UK Biobank prospective cohort design and analytical considerations

Naomi E. Allen^{1,2}, Ben Lacey^{1,2}, Deborah A. Lawlor^{3,4}, Jill P. Pell⁵, John Gallacher^{6,7}, Liam Smeeth⁸, Paul Elliott^{9,10}, Paul M. Matthews¹², Ronan A. Lyons¹³, Anthony D. Whetton¹⁴, Anneke Lucassen^{15,16}, 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 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

- 25 ¹³ Population Data Science, Swansea University Medical School, Swansea, Wales
- 26 ¹⁴ Veterinary Health Innovation Engine, University of Surrey, Guildford, UK
- 27 ¹⁵ Nuffield Department of Medicine, University of Oxford, Oxford, UK
- 28 ¹⁶ Faculty of Medicine, Southampton University, Southampton, UK
- 29 ¹⁷ Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge, UK
- 30 ¹⁸ NHS Digital, London, UK
- 31 ¹⁹ Our Future Health, London, UK
- 32 ²⁰ Institute of Health Informatics, University College London, London, UK
- 33 ²¹ Centre for Environmental Health and Sustainability, University of Leicester, Leicester, UK
- 34 ²² School of Sport, Exercise and Health Sciences, Loughborough University, Loughborough,
- 35 UK
- 36 ²³ Centre for Genomic and Experimental Medicine, University of Edinburgh, Edinburgh,
- 37 Scotland
- 38 ²⁴ School of Medicine, Cardiff University, Cardiff, Wales
- 39 ²⁵ UK Dementia Research Institute, Cardiff University, Cardiff, Wales
- 40 ²⁶ Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of
- 41 Oxford, Oxford, UK
- 42 ²⁷ British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health
- 43 and Primary Care, University of Cambridge, Cambridge, UK
- 44 ²⁸ Health Data Research UK Cambridge, Wellcome Genome Campus and University of
- 45 Cambridge, Cambridge, UK.
- 46 ²⁹ National Institute for Health Research Cambridge Biomedical Research Centre, University
- 47 of Cambridge and Cambridge University Hospitals, Cambridge, UK.
- 48

49 Corresponding author Emails:

50 naomi.allen@ndph.ox.ac.uk

51

52

53 **Overline: Biobanks**

54

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

57

58 **Abstract**

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

64

65 **Introduction**

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

81 More recently, there has been remarkable progress in research on the genetic
82 determinants of disease. In the early 2000s, the literature was dominated by a plethora of
83 genetic studies that focused on associations with particular conditions within specific
84 “candidate” genes that were of *a priori* interest. Many of these studies involved small numbers
85 of disease cases and yielded false-positive results that failed to replicate, often because of
86 undue emphasis on *post hoc* selective reporting of the more extreme associations that were
87 observed. Subsequently, improvements in assay technology led to genome-wide association
88 studies (GWAS) that allowed hypothesis-free identification across the genome of variants
89 associated with a particular phenotype. Much effort was typically spent on characterising the
90 phenotype under investigation precisely in the belief that outcome misclassification would

91 have a substantial impact on the ability to detect associations. However, when meta-analyses
92 of different studies were performed that yielded much larger numbers of individuals with the
93 outcome of interest (albeit differently defined), small-to-moderate associations between
94 genetic variants and outcomes began to be identified reproducibly after stringent adjustment
95 for multiple testing (8).

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

102 **UK Biobank combines scale, depth, duration and accessibility**

103 UK Biobank is a population-based prospective cohort of 500,000 men and women designed
104 to enable research into the genetic, lifestyle and environmental determinants of a wide range
105 of diseases of middle-to-old age (www.ukbiobank.ac.uk). It was established by the UK Medical
106 Research Council (MRC) and Wellcome, which continue to fund it along with the British Heart
107 Foundation (BHF), Cancer Research UK (CR-UK) and National Institute for Health and Care
108 Research (NIHR). The key design features are its easy accessibility, large-scale prospective
109 nature, depth and range of risk factor data, and comprehensive linkage to health outcomes,
110 which together enable academic and industry researchers worldwide to perform discovery
111 science (Supplementary Table 1).

112 UK Biobank was designed to promote innovative science by maximising access to the
113 data in an equitable and transparent manner. All approved researchers (academic or
114 commercial) can access all of the de-identified data in order to perform any type of health-
115 related research that is in the public interest. This is the key criterion against which applications

116 to access the data are considered, with restrictions only placed on their use for potentially
117 contentious research (for example, investigations that could lead to racial or sexual
118 discrimination). Access to biological samples is currently largely restricted to assays that will
119 be conducted on the whole cohort or large representative samples of the cohort.

120 Ready access to such a large-scale, in-depth resource has encouraged researchers
121 from many disciplines across academia and industry to collaborate to ensure that different
122 types of complex data (e.g., whole-exome and whole-genome sequencing data, magnetic
123 resonance imaging (MRI) scans, accelerometer wave-form data, and electronic health
124 records) are generated and analysed appropriately. The ready accessibility of the data at low
125 cost without requiring collaboration with, or peer review from, the UK Biobank study
126 investigators has led to an exponential increase in research output. By the end of 2023, there
127 were more than 30,000 registered researchers (80% from outside the UK) and about 9,000
128 publications (attracting 270,000 citations), with the number of publications increasing
129 exponentially each year. In particular, the release to the worldwide research community of
130 cohort-wide genome-wide genotyping and imputation data in 2017 has been hugely influential
131 in advancing our understanding of the genetic determinants of disease.

132 The requirement that researchers publish their findings and make available any
133 derived variables that have been generated as part of their research, together with the
134 underlying code that generated the research output, enables the wider scientific community to
135 critique, modify and build upon the work of others in a transparent manner (10). For example,
136 research groups with expertise in signal processing have created derived variables related to
137 the intensity and duration of physical activity from the raw accelerometer data (11, 12).
138 Similarly, academic and commercial research groups with expertise in image analysis have
139 made available variables derived from the MRI scans related to body fat distribution (13), fat
140 and iron content of specific organs (14, 15) and metrics of the structure and function of the
141 brain (16) and heart (17). In this way, complex data that might otherwise only be of use to

142 specialists in a narrow field of research are turned into well-curated derived variables that are
143 integrated with other UK Biobank data and can be used extensively by non-specialists to
144 answer a range of research questions.

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

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

168 UK Biobank has been set up to help minimise random and systematic error so that it
169 can support reliable research into the determinants of disease (Supplementary Table 1),
170 although the general principles of careful study design and appropriate data analysis apply
171 equally to all large-scale, prospective studies. There are a number of trade-offs that need to
172 be considered when designing a cohort study, which relate to the size and heterogeneity of
173 the study population, and to the methods used for its recruitment, data collection and follow-
174 up. UK Biobank has aimed to generate a large-scale, prospective biomedical resource that
175 includes a wide range of exposure and health outcome measures collected as accurately as
176 possible, with easy accessibility to the data. However, as with all prospective studies, it is
177 important to consider, and if possible correct for, potential biases arising from the study design
178 and collection of data.

179 **The importance of a large-scale prospective design**

180 UK Biobank recruited 502,000 volunteers aged 40-69 years at recruitment between
181 2006 and 2010 from across England, Wales and Scotland. This age group was selected to
182 include individuals who were young enough that relatively few would have developed health
183 conditions at the time of recruitment. As a prospective study, UK Biobank has many
184 advantages for investigating the effects of genetic, lifestyle and environmental factors on
185 disease outcomes (21). In particular, information on exposures to potential risk factors can be
186 assessed before disease develops, which avoids bias caused by differential recall of
187 information about past exposures depending on an individual's outcome status (recall bias).
188 The prospective design also allows investigation of factors that might be affected by disease
189 processes or their treatment, or by changes in an individual's behavior following the
190 development of some condition (reverse causation bias). In addition, it can support studies of
191 conditions that cannot readily be investigated retrospectively (e.g. fatal illnesses).
192 Furthermore, by allowing a wide range of different conditions to be studied within the same
193 study population, the full effects of a particular exposure on all aspects of health can be better

194 assessed (e.g. smoking on a wide range of different diseases). Likewise, the effects of many
195 different exposures on a single disease can be determined, provided that sufficient numbers
196 of cases have occurred to allow the separate and combined effects of exposures to be
197 assessed reliably.

198 Prospective studies need to be large, as only a relatively small proportion of the
199 participants will develop any given condition during follow-up. The rationale for recruiting
200 500,000 adults into UK Biobank was that it would enable large numbers of cases of the most
201 common diseases to develop within a reasonable follow-up period (while also allowing detailed
202 exposure information to be collected within funding and organisational constraints). For
203 example, after a median follow-up of 12 years (i.e. by end-2020), linkage to electronic
204 healthcare record data indicated that there had been at least 30,000 incident cases of
205 diabetes, 25,000 cases of depression, 15,000 cases of myocardial infarction, and 10,000
206 cases of breast cancer (Table 1). For the reliable detection of risk ratios of about 1.3 for the
207 main effects of different exposures (ranging from those that are dichotomous variables to
208 those that are continuous measures), about 5,000-10,000 incident cases of a particular
209 disease would be required (22). The need for a large sample size is even more evident when
210 assessing combined effects. For example, when estimating the joint effect of blood pressure
211 and age on the risk of coronary heart disease, the standard error of the estimates (and hence
212 the 95% confidence intervals) are, on average, three times narrower with 500,000 versus
213 50,000 participants (23). As the UK Biobank participants age, the number of incident cases of
214 different diseases is increasing substantially, allowing a wider range of outcomes to be
215 investigated more completely. For example, by 2032 there will be over 50,000 cases of
216 diabetes and chronic obstructive pulmonary disease. The sheer size of the study also means
217 that robust research into less common conditions will also be possible. For example, between
218 2020 and 2027, the number of cases of Alzheimer's disease, hip fracture and Parkinson's
219 disease is expected to more than double to about 17,000, 13,000 and 10,000, respectively
220 (Table 1).

221 **Comparing cohort characteristics with that of the wider population**

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

234 It is this heterogeneity across different levels of exposures (e.g., genetic, lifestyle,
235 sociodemographic and environmental exposures) - and not the relatively low overall
236 participation rate - that largely determines the generalisability of the findings to the broader
237 UK population (25). For example, although individuals from more socio-economic deprived
238 areas are under-represented in UK Biobank (16% versus 33% in the UK population), there
239 are sufficiently large numbers of this group (82,000) to enable reliable assessment of the
240 association of socio-economic deprivation with disease risk. By contrast, although UK Biobank
241 is reasonably representative of the distribution for different ethnic groups, with 29,000
242 participants recruited from Black and other ethnic minority groups (which was about the same
243 proportion, ~5%, as the rest of the UK population at the time) (26), it is insufficient to examine
244 reliably the differences in exposure-disease associations by ethnicity. Even though UK
245 Biobank is currently the largest study in the world with whole-genome sequencing data on

246 individuals of African and South Asian ancestry (27), the numbers are still relatively small (with
247 about 10,000 participants in each ethnic group).

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

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

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

283 The general consistency of research findings in UK Biobank with those in other studies (36-
284 38) – in particular, studies considered to be representative of the underlying population –
285 suggest that many of the exposure-disease associations found in UK Biobank are largely
286 generalizable to other populations. For example, the direction and magnitude of associations
287 of genetic variants with osteoarthritis in UK Biobank are consistent with the associations
288 observed in deCODE, which recruited more than half of Iceland’s adult population (39).
289 Likewise, although the frequency of genetic variants may vary substantially in studies
290 conducted in different populations (resulting in differing statistical power to detect
291 associations), the direction and magnitude of genetic associations are typically similar across
292 populations e.g. the association of specific *GPR75* gene variants with obesity in UK, US and
293 Mexico cohorts (40).

294 Nonetheless, there may be circumstances in which associations between an exposure and
295 disease risk varies across different populations. For example, polygenic risk scores developed
296 and tested in populations of European ancestry often perform less well when applied to African
297 and South Asian populations, owing to differences in allele frequencies and linkage

298 disequilibrium patterns between the ethnic groups (41). As such, other large population
299 cohorts with biological samples are needed around the world to increase the heterogeneity of
300 genetic and non-genetic risk factors for disease (42) (Table 2). For example, studies
301 established in Mexico (150,000 participants) and China (500,000 participants) at about the
302 same time as UK Biobank have enabled reliable investigation into the association between
303 the risk of hypertension with body weight above and below the Western norm (43, 44). Large-
304 scale studies established across Europe and China have also taken advantage of the
305 heterogeneity of dietary and other exposures across different populations (45,46). Genetic
306 and other assays of stored samples in these studies are extending the range of genomic risk
307 factors that can now be investigated. New large-scale prospective studies are now established
308 in the US e.g., All of Us (47) and the Million Veterans Program (48), and are also being
309 established in Asia and parts of Africa e.g., Non-communicable Diseases Genetic Heritage
310 Study in Nigeria (49, 50). This will further increase the ability to assess associations with
311 disease risk across a broad range of genetic (and non-genetic) factors as long as there is
312 sufficient duration of follow-up.

313 **Reliable assessment of a wide range of exposures**

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

322 As a result, UK Biobank has put significant effort into collecting comprehensive, accurate and
323 high-quality data for many different types of exposures. Repeated measures have also been

324 conducted in subsets of participants to address random error in exposure levels and thereby
325 be able to correct for regression-dilution bias. However, there is real value in being able to
326 perform cohort-wide repeat measures that would allow the relevance of individual changes in
327 exposures over time to be assessed.

328 **Depth and breadth of exposure measurement**

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

347 A wide range of physical measurements were also taken for all 500,000 participants,
348 comprising blood pressure, anthropometry (sitting and standing height, weight, waist and hip
349 circumference, and bioimpedance measures), hand grip strength, vision and lung function.

350 Blood and urine samples were also collected for long-term storage (Fig. 1). A proportion of the
351 cohort also underwent a heel ultrasound for bone density, pulse wave velocity of arterial
352 stiffness, a hearing test (180,000 participants), an eye examination (including refractive index),
353 intraocular pressure measurements, a retinal photograph and optical coherence tomography
354 (120,000 participants), a cardio-respiratory fitness test with a 4-lead electrocardiogram (ECG)
355 (78,000 participants), and collection of a saliva sample (~85,000 participants). Since the
356 baseline assessment, UK Biobank continues to collect additional data from large subsets of
357 the cohort. This has included data on physical activity using a 7-day accelerometer (in 100,000
358 participants, with 2,500 undergoing a repeat assessment), a multi-modal imaging assessment
359 (in up to 100,000 participants, with 60,000 undergoing a repeat assessment over the next few
360 years) and a series of web-based questionnaires that cover specific exposures in more depth
361 (e.g. diet, cognition, occupational history).

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

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

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

397 The collection of data on a wide range of measures enables researchers to allow not only for
398 more complete and accurate measurement of exposures, but also for potential confounders
399 (extraneous factors that are associated with the exposure and outcome) and mediators
400 (factors that are on the causal pathway between the exposure and the outcome). This is
401 important, as random error in exposure measures can cause systematic attenuation of any
402 true association, whereas random measurement error of confounders can result in an

403 apparent exposure-disease association, where none really exists. For example, the observed
404 inverse association of fruit and vegetable intake with cardiovascular disease risk in UK
405 Biobank is likely to be due largely to residual confounding by socio-economic factors, which
406 are difficult to assess and therefore subject to measurement error (57). The ability of UK
407 Biobank to obtain more detailed information in the future about socio-economic factors (such
408 as education, occupation and income via linkage to administrative datasets or specific web-
409 based questionnaires) would enable more precise characterisation and, hence, even better
410 adjustment for these important factors.

411 Because all epidemiological studies suffer, to a greater or lesser extent, from imperfect
412 measurement of exposures and their potential confounders, various analytical methods have
413 been developed to quantify and control for this. Perhaps the simplest approach is the
414 comparison of likelihood ratio statistics associated with the exposure of interest in the models
415 before and after adjustment for covariates. Generally speaking, a large proportional reduction
416 in the likelihood ratio chi-square ($LR\chi^2$) test after the addition to the model of covariates is
417 indicative that the association likely remains affected by residual confounding, as adjustment
418 for confounders that are perfectly measured would be expected to reduce the χ^2 even further
419 (6). An increasingly popular approach to distinguish the likely causal effect of an exposure
420 (from that of extraneous confounders) is the use of Mendelian Randomisation – facilitated in
421 analyses of UK Biobank by the extensive genetic information available on all of the participants
422 – whereby specific genetic variants are used as proxies for exposures of interest or their
423 confounders. For example, this approach has provided strong support for a causal role of body
424 fat mass and interleukin-6 in development of cardiovascular conditions (58, 59). Conversely,
425 Mendelian Randomisation has not provided support for a protective effect of vitamin D against
426 COVID-19 (60), cancer or cardiovascular outcomes (61), although it should be noted that
427 Mendelian Randomisation analyses may also be affected by bias in some circumstances (62).
428 When associations of genetic variants with the relevant non-genetic risk factors are weak
429 (such that Mendelian Randomisation may not be effective), the likely impact of residual

430 confounding due to imprecision in measured variables included in the model can be assessed
431 using other analytical approaches such as probabilistic or multiple-bias analysis (34, 63). The
432 use of different analytical strategies to triangulate evidence (for example, comparing results
433 from models that include traditional observational variables with those that use genetic
434 instrumental variables) will enable researchers to assess different biases and their potential
435 impact on causal inference in a more robust manner.

436 **Repeated exposure measurements**

437 Random errors in the measurement of risk factors can lead to substantial underestimation of
438 the strength of their associations with subsequent health outcomes (regression dilution bias)
439 (64, 65), as well as to substantial residual confounding when measurement error is present in
440 confounders (66). These biases may be increased further through random error in risk factor
441 measurements that occur during prolonged follow-up in prospective cohorts. For example, the
442 true association of blood pressure and cholesterol with cardiovascular disease risk may be
443 underestimated by about one-third in the first decade of follow-up and up to two-thirds in the
444 third decade (64). However, despite regression dilution being one of the most important biases
445 in exposure-disease associations, it is often overlooked in analyses of prospective studies,
446 including UK Biobank (with some exceptions) (67-70). It is possible to correct for regression
447 dilution bias by using repeat measures from a relatively small subset of the cohort. UK Biobank
448 performed a repeat assessment on 20,000 participants in 2012-2013 to allow researchers to
449 address this issue specifically. Re-measures collected during the imaging assessment of up
450 to 100,000 UK Biobank participants during 2014-2024 and a repeat assessment of up to
451 60,000 during 2019-2029 can be used to make appropriate time-dependent corrections for the
452 effects of regression dilution bias.

453 In addition to addressing error caused (largely) by random error in baseline risk factors,
454 repeated measures would also enable correction for systematic intra-individual changes in

455 exposures over time, which may lead to either over-estimation or under-estimation of
456 associations depending on the nature and magnitude of misclassification. For example,
457 secular trends in the reduction of smoking or exposure to environmental pollutants may lead
458 to an underestimation of their association with disease risk if solely based on baseline
459 measures. To help address this issue, UK Biobank is exploring opportunities to collect
460 information on longitudinal changes in environmental exposures (e.g. from existing data on
461 changes in participants' residential location or future data collection using smartphone GPS
462 tracking) to enable more accurate inferences to be made about how changes in environmental
463 exposures affect health in the long-term. It is also intended to repeat previous web-based
464 questionnaires in order to capture longitudinal changes in specific lifestyle factors such as diet
465 and sleep.

466 Whereas repeated measures of the baseline assessment are being captured during the
467 imaging assessments in a subset of the UK Biobank cohort, it would be desirable to perform
468 a future repeat assessment of a wide range of exposures in as many of the participants as
469 possible. This would allow investigation of how lifestyle, and physical and biochemical
470 changes over time influence disease risk and progression, thereby helping to determine the
471 temporality of associations and their underlying mechanisms. Data collection for as many
472 surviving participants as possible would also reduce systematic error caused by differential
473 participation rates that are related to the exposures and outcomes under investigation. UK
474 Biobank generally has excellent participant engagement with an ongoing series of repeated
475 web-based questionnaires (with response rates of >50%), physical activity monitoring (45%
476 for the first assessment, of whom 63% also performed a repeat assessment), and imaging
477 assessments (24% for the first assessment and 65% for a repeat assessment). However,
478 researchers should be aware that participants who engage in ongoing data collection activities
479 (including repeat assessments) might not be representative of the cohort as a whole. For
480 example, genetic variants associated with completing UK Biobank online questionnaires and
481 activity monitoring are correlated with several metrics of better health (31). Attrition bias has

482 been documented in other prospective studies (71-73), suggesting that similar factors affect
483 ongoing participant engagement in many cohorts, regardless of their design, recruitment
484 strategy or study population.

485 **Reliable assessment of a wide range of health outcomes**

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

494 **Comprehensive ascertainment of health outcomes**

495 All cohort studies need a comprehensive and efficient way of following participants' health
496 over the long-term to identify a wide range of disease outcomes. Unlike many countries
497 (including the US and most low-to-middle income countries), the UK's National Health Service
498 (NHS) collates and stores electronic health administrative records for clinical care. However,
499 the data content, format and governance requirements may differ for England, Wales and
500 Scotland. To identify a wide range of health outcomes over a prolonged period, UK Biobank
501 has linked to these health administrative records for all participants. This has the advantage
502 of minimising ascertainment bias and reducing loss-to-follow-up or attrition bias by providing
503 cohort-wide follow-up information without the need for active participant re-contact, which may
504 be incomplete. Moreover, the low rate of UK Biobank participants requesting that all of their
505 data and samples be withdrawn from the study (0.2%; most of which occurred soon after

506 recruitment) also minimises systematic bias associated with loss to follow-up from non-
507 random subgroups of the cohort.

508 To date, UK Biobank has linked NHS healthcare data from centralised national cancer and
509 death registries and hospital inpatient admissions for all participants. In contrast, primary care
510 data are not centralised but instead are held by commercial electronic system suppliers under
511 the control of individual general practices, so it has been more challenging to obtain the
512 agreements necessary to obtain these data for all participants. Primary care data are currently
513 available for 45% of the UK Biobank cohort for general research purposes (which represents
514 complete coverage from one primary care system supplier, up to 2016/2017) and for 80% of
515 the cohort for COVID-19 research (complete coverage from two system suppliers in England,
516 up to mid-2021, enabled by emergency legislation to facilitate COVID-19 research). Whereas
517 both subsets are broadly representative of the cohort with respect to the distribution of potential
518 exposures, researchers should be encouraged to check these underlying assumptions prior to
519 analysis. Incorporation of primary care data for all 500,000 participants for all types of health-
520 related research would be of enormous value as it will increase substantially the number of
521 health outcomes that can be detected (thereby increasing statistical power) and their
522 diagnostic accuracy (thereby increasing specificity). For example, whereas addition of primary
523 care data would increase the numbers of myocardial infarction cases identified by less than
524 5%, the numbers of cases identified of diabetes and chronic obstructive pulmonary disease
525 (COPD) would increase by about 40% (Fig. 2). Primary care data are also important for
526 investigating risk factors associated with disease severity, where associations may differ
527 between milder disease subtypes generally captured in primary care records and more severe
528 disease captured in hospital admission data.

529 Whereas linkage to health records ensures comprehensive coverage, there is the
530 possibility of “collider bias” if health outcomes are differentially ascertained based on
531 participant characteristics (e.g., ethnicity), as reported by some researchers in the context of
532 COVID-19 research (74). However, there are a range of analytical approaches that can be

533 used to investigate this type of bias (74-76) and the ascertainment of most health outcomes
534 are not so strongly influenced by these characteristics.

535 **Specificity of health outcomes**

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

549 It is also important to characterise disease subtypes as low biological specificity can limit
550 interpretation of results. To address this, UK Biobank (78-80) and other open-access
551 resources (81) have developed a number of algorithmically defined health outcomes based
552 on inter-operable code lists from electronic healthcare records. Diagnostic codes contained in
553 these records have also been mapped to a common standard (ICD-10) to facilitate broad-
554 brush research. Whereas these coded health outcomes may be sufficient for most research
555 purposes, they may lack specificity to identify disease subtypes. This could lead to materially
556 biased estimates of associations if the determinants of these apparently similar, but
557 etiologically different, disease subtypes differ. For example, while blood pressure is strongly

558 positively associated with the risk of both ischaemic and haemorrhagic stroke (82), the
559 association of cholesterol and certain genetic variants with stroke differ substantially by
560 subtype (83, 84) providing clues to the underlying aetiology of this heterogeneous condition.
561 To increase the specificity of health outcomes beyond the available coded data, UK Biobank
562 intends to collect detailed data on disease sub-types over the next few years. For example,
563 this could include disease-specific registers such as the National Diabetes Audit that collects
564 data on diabetes subtypes, clinical scans to identify stroke sub-types, digitised histopathology
565 slides to determine tumour morphological subtypes, and in-person assessments to
566 characterise dementia subtypes.

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

574 **Long duration of follow-up**

575 For any prospective study, a long duration of follow-up (i.e. decades or more) is needed for
576 sufficiently large numbers of health outcomes to accrue for reliable investigation. It also allows
577 for the identification of recurring events and factors associated with disease progression. While
578 the incidence of common health outcomes during the early years of follow-up in UK Biobank
579 was somewhat lower than in the general population due to the 'healthy volunteer' effect, which
580 is typical of such studies (89), its impact is now reduced as the cohort has aged. With
581 prolonged follow-up, large numbers of incident cases of a wide range of conditions have
582 already occurred. Over the next five to ten years there will be thousands of incident cases of

583 common outcomes (Table 1), enabling reliable investigation of their genetic, lifestyle and
584 environmental determinants.

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

600 **Conclusions**

601 The success of UK Biobank has been due, in large part, to the altruism of the 500,000
602 volunteers, but also the global research community who have been – and continue to be –
603 involved in expanding the range of exposures and outcomes available for research. Such
604 enhancements (e.g. sample assays, linkage to specific healthcare datasets and environmental
605 sources, etc.) help to ensure that the resource fulfils the needs of researchers and remains at
606 the forefront of scientific discovery.

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

615 Easy accessibility of UK Biobank data and research results (including the underlying analytical
616 code) is enabling the community to directly peer review research by undertaking replication
617 analyses, or to apply different methods to the same research question, to confirm or refute the
618 findings of others. In particular, investigation of approaches used to identify and quantify the
619 uncertainty of the results based on sensitivity analyses that examine systematic bias, will
620 provide a level of transparency in the interpretation of findings that has, until now, generally
621 been under-reported.

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

630

631

632

633

634

635

636

637

638

639

640 **References and Notes**

641 (1) R. Doll, A.B. Hill, Lung cancer and other causes of death in relation to smoking; a second
642 report on the mortality of British doctors, *BMJ*, **2**, 1071-1081 (1956).

643

644 (2) J. Truett, J. Cornfield, W. Kannel, A multivariate analysis of the risk of coronary heart
645 disease in Framingham, *J Chronic Dis*, **20**, 511-524 (1967).

646

647 (3) Emerging Risk Factors Collaboration, Lipoprotein(a) concentration and the risk of coronary
648 heart disease, stroke, and nonvascular mortality, *JAMA*, **302**, 412-423 (2009).

649

650 (4) Emerging Risk Factors Collaboration, Separate and combined associations of body-mass
651 index and abdominal adiposity with cardiovascular disease: collaborative analysis of 58
652 prospective studies, *Lancet*, **377**, 1085-1095 (2011).

653

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

656 52,705 women with breast cancer and 108,411 women without breast cancer. *Lancet*, **350**,
657 1047-1059 (1997).

658

659 (6) Collaborative Group on Hormonal Factors in Breast Cancer, Alcohol, tobacco and breast
660 cancer - collaborative reanalysis of individual data from 53 epidemiological studies, including
661 58,515 women with breast cancer and 95,067 women without the disease, *Br J Cancer*, **87**,
662 1234-1245 (2002).

663

664 (7) Collaborative Group on Hormonal Factors in Ovarian Cancer, Ovarian cancer and oral
665 contraceptives: collaborative reanalysis of data from 45 epidemiological studies including
666 23,257 women with ovarian cancer and 87,303 controls, *Lancet*, **371**, 303-314 (2008).

667

668 (8) E. Uffelmann, Q.Q. Huang, N.S. Munung, J. de Vries, Y. Okada, A.R. Martin, C.M. Martin,
669 T. Lappalainen, D. Posthuma, Genome-wide association studies. *Nat Rev Methods Primers*,
670 **1**, 59 (2021).

671

672 (9) J.A. Luan, M.Y. Wong, N.E. Day, N.J. Wareham, Sample size determination for studies of
673 gene-environment interaction, *Int J Epidemiol*, **30**, 1035-1040 (2001).

674

675 (10) M. Conroy, J. Sellors, M. Effingham, T.J. Littlejohns, C. Boultonwood, L. Gillions, C.L.M.
676 Sudlow, R. Collins, N.E. Allen, The advantages of UK Biobank's open access strategy for
677 health research, *J Intern Med*, **286**, 389-397 (2019).

678

679 (11) S. Cassidy, H. Fuller, J. Chau, M. Catt, A. Bauman, M.I. Trenell, Accelerometer-derived
680 physical activity in those with cardio-metabolic disease compared to healthy adults: a UK
681 Biobank study of 52,556 participants. *Acta Diabetologica*, **55**, 975-979 (2018).

682

683 (12) A. Doherty, D. Jackson, N. Hammerla, T. Plötz, P. Olivier, M.H. Granat, T. White, V.T.
684 van Hees, M.I. Trenell, C.G. Owen, S.J. Preece, R. Gillions, S. Sheard, T. Peakman, S. Brage,
685 N.J. Wareham, Large Scale Population Assessment of Physical Activity Using Wrist Worn
686 Accelerometers: The UK Biobank Study. *PloS One*, **12**, e0169649 (2017).
687
688 (13) M. Borga, J. West, J.D. Bell, N.C. Harvey, T. Romu, S.B. Heymsfield, O. Dahlqvist
689 Leinhard, Advanced body composition assessment: from body mass index to body
690 composition profiling. *J Invest Med*, **66**, 1-9 (2018).
691
692 (14) Y. Liu, N. Basty, B. Witcher, J.D. Bell, E.P. Sorokin, N. van Bruggen, E.L. Thomas, M.
693 Cule, Genetic architecture of 11 organ traits derived from abdominal MRI using deep learning,
694 *eLife*, **10**, e65554 (2021).
695
696 (15) A. McKay, H.R. Wilman, A. Dennis, M. Kelly, M.L. Gyngell, S. Neubauer, J.D. Bell, R.
697 Banerjee, E.L. Thomas, Measurement of liver iron by magnetic resonance imaging in the UK
698 Biobank population. *PloS One*, **13**, e0209340 (2018).
699
700 (16) F. Alfaro-Almagro, M. Jenkinson, N.K. Bangerter, J.L.R. Andersson, L. Griffanti, G.
701 Douaud, S.N. Sotiropoulos, S. Jbabdi, M. Hernandez-Fernandez, E. Vallee, D. Vidaurre, M.
702 Webster, P. McCarthy, C. Rorden, A. Daducci, D.C. Alexander, H. Zhang, I. Dragonu, P.M.
703 Matthews, K.L. Miller, S.M. Smith, Image processing and Quality Control for the first 10,000
704 brain imaging datasets from UK Biobank. *NeuroImage*, **166**, 400-424 (2018).
705
706 (17) W. Bai, H. Suzuki, J. Huang, C. Francis, S. Wang, G. Tarroni, F. Guitton, N. Aung, K.
707 Fung, S.E. Petersen, S.K. Piechnik, S. Neubauer, E. Evangelou, A. Dehghan, D.P. O'Regan,
708 M.R. Wilkins, Y. Guo, P.M. Matthews, D. Rueckert, A population-based phenome-wide
709 association study of cardiac and aortic structure and function. *Nature Med*, **26**, 1654-1662
710 (2020).

711

712 (18) O. Canela-Xandri, K. Rawlik, A. Tenesa, An atlas of genetic associations in UK Biobank.
713 *Nature Genet*, **50**, 1593-1599 (2018).

714

715 (19) G. McInnes, Y. Tanigawa, C. DeBoever, A. Lavertu, J.E. Olivieri, M. Aguirre, M.A. Rivas,
716 Global Biobank Engine: enabling genotype-phenotype browsing for biobank summary
717 statistics. *Bioinformatics*, **35**, 2495-2497 (2019).

718

719 (20) T.G. Richardson, S. Harrison, G. Hemani, G. Davey Smith, An atlas of polygenic risk
720 score associations to highlight putative causal relationships across the human phenome.
721 *eLife*, **8**, e43657 (2019).

722

723 (21) D.A. Grimes, K.F. Schulz, Cohort studies: marching towards outcomes. *Lancet*, **359**, 341-
724 345 (2002).

725

726 (22) P.R. Burton, A.L. Hansell, I. Fortier, T.A. Manolio, M.J. Khoury, J. Little, P. Elliott, Size
727 matters: just how big is BIG?: Quantifying realistic sample size requirements for human
728 genome epidemiology. *Int J Epidemiol*, **38**, 263-273 (2009).

729

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

731

732 (24) A. Fry, T.J. Littlejohns, C. Sudlow, N. Doherty, L. Adamska, T. Sprosen, R. Collins, N.E.
733 Allen, Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank
734 Participants with the General Population. *Am J Epidemiol*, **186**, 1026-1034 (2017).

735

736 (25) K.J. Rothman, J.E. Gallacher, E.E. Hatch, Why representativeness should be avoided.
737 *Int J Epidemiol*, **42**, 1012-1014 (2013).

738

739 (26) C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott,
740 J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen,
741 T. Peakman, R. Collins, UK biobank: an open access resource for identifying the causes of a
742 wide range of complex diseases of middle and old age. *PLoS Med*, **12**, e1001779 (2015).

743

744 (27) B.V. Halldorsson, H.P. Eggertsson, K.H.S. Moore, H. Hauswedell, O. Eiriksson, M.O.
745 Ulfarsson, G. Palsson, M.T. Hardarson, A. Oddsson, B.O. Jensson, S. Kristmundsdottir, B.D.
746 Sigurpalsdottir, O.A. Stefansson, D. Beyter, G. Holley, V. Tragante, A. Gylfason, P.I. Olason,
747 F. Zink, M. Asgeirsdottir, S.T. Sverrisson, B. Sigurdsson, S.A. Gudjonsson, G.T. Sigurdsson,
748 G.H. Halldorsson, G. Sveinbjornsson, K. Norland, U. Styrkarsdottir, D.N. Magnusdottir, S.
749 Snorraddottir, K. Kristinsson, E. Sobech, H. Jonsson, A.J. Geirsson, I. Olafsson, P. Jonsson,
750 O.B. Pedersen, C. Erikstrup, S. Brunak, S.R. Ostrowski, G. Thorleifsson, F. Jonsson, P.
751 Melsted, I. Jonsdottir, T. Rafnar, H. Holm, H. Stefansson, J. Saemundsdottir, D.F.
752 Gudbjartsson, O.T. Magnusson, G. Masson, U. Thorsteinsdottir, A. Helgason, H. Jonsson, P.
753 Sulem, K. Stefansson, The sequences of 150,119 genomes in the UK Biobank. *Nature*, **607**,
754 732-740 (2022).

755

756 (28) E. Stamatakis, K.B. Owen, L. Shepherd, B. Drayton, M. Hamer, A.E. Bauman, Is Cohort
757 Representativeness Passé? Poststratified Associations of Lifestyle Risk Factors with Mortality
758 in the UK Biobank, *Epidemiol*, **32**, 179-188 (2021).

759

760 (29) J.R. Emberson, D.A. Bennett, Effect of alcohol on risk of coronary heart disease and
761 stroke: causality, bias, or a bit of both?, *Vasc Health Risk Manag*, **2**, 239-249 (2006).

762

763 (30) S. Ebrahim, G. Davey Smith, Commentary: Should we always deliberately be non-
764 representative?, *Int J Epidemiol*, **42**, 1022-1026 (2013).

765

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. Spiegelhalter, 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

794 for self-reported habitual sleep duration supported by accelerometer-derived estimates,
795 *Nature Comms*, **10**, 1100 (2019).

796

797 (38) J. Deelen, D.S. Evans, D.E. Arking, N. Tesi, M. Nygaard, X. Liu, M.K. Wojczynski, M.L.
798 Biggs, A. van der Spek, G. Atzmon, E.B. Ware, C. Sarnowski, A.V. Smith, I. Seppälä, H.J.
799 Cordell, J. Dose, N. Amin, A.M. Arnold, K.L. Ayers, N. Barzilai, E.J. Becker, M. Beekman, H.
800 Blanché, K. Christensen, L. Christiansen, J.C. Collerton, S. Cubaynes, S.R. Cummings, K.
801 Davies, B. Debrabant, J.F. Deleuze, R. Duncan, J.D. Faul, C. Franceschi, P. Galan, V.
802 Gudnason, T.B. Harris, M. Huisman, M.A. Hurme, C. Jagger, I. Jansen, M. Jylhä, M. Kähönen,
803 D. Karasik, S.L.R. Kardia, A. Kingston, T.B.L. Kirkwood, L.J. Launer, T. Lehtimäki, W. Lieb,
804 L.P. Lytykäinen, C. Martin-Ruiz, J. Min, A. Nebel, A.B. Newman, C. Nie, E.A. Nohr, E.S.
805 Orwoll, T.T. Perls, M.A. Province, B.M. Psaty, O.T. Raitakari, M.J.T. Reinders, J.M. Robine,
806 J.I. Rotter, P. Sebastiani, J. Smith, T.I.A. Sørensen, K.D. Taylor, A.G. Uitterlinden, W. van der
807 Flier, S.J. van der Lee, C.M. van Duijn, D. van Heemst, J.W. Vaupel, D. Weir, K. Ye, Y. Zeng,
808 W. Zheng, H. Holstege, D.P. Kiel, K.L. Lunetta, P.E. Slagboom, J.M. Murabito, A meta-
809 analysis of genome-wide association studies identifies multiple longevity genes. *Nature*
810 *Comm*, **10**, 3669 (2019).

811

812 (39) U. Styrkarsdottir, S.H. Lund, G. Thorleifsson, F. Zink, O.A. Stefansson, J.K. Sigurdsson,
813 K. Juliusson, K. Bjarnadottir, S. Sigurbjornsdottir, S. Jonsson, K. Norland, L. Stefansdottir, A.
814 Sigurdsson, G. Sveinbjornsson, A. Oddsson, G. Bjornsdottir, R.L. Gudmundsson, G.H.
815 Halldorsson, T. Rafnar, I. Jonsdottir, E. Steingrimsson, G.L. Norddahl, G. Masson, P. Sulem,
816 H. Jonsson, T. Ingvarsson, D.F. Gudbjartsson, U. Thorsteinsdottir, K. Stefansson, Meta-
817 analysis of Icelandic and UK data sets identifies missense variants in SMO, IL11, COL11A1
818 and 13 more new loci associated with osteoarthritis. *Nature Genet*, **50**, 1681-1687 (2018).

819

820 (40) P. Akbari, A. Gilani, O. Sosina, J.A. Kosmicki, L. Khrimian, Y.Y. Fang, T. Persaud, V.
821 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. Meijssen, 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

849 (45) E. Riboli, R. Kaaks, The EPIC Project: rationale and study design. European Prospective
850 Investigation into Cancer and Nutrition. *Int J Epidemiol*, **26** Suppl 1, S6-14 (1997).
851

852 (46) Z. Chen, J. Chen, R. Collins, Y. Guo, R. Peto, F. Wu, L. Li, China Kadoorie Biobank of
853 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int J*
854 *Epidemiol*, **40**, 1652-1666 (2011).
855

856 (47) J.C. Denny, J.L. Rutter, D.B. Goldstein, A. Philippakis, J.W. Smoller, G. Jenkins, E.
857 Dishman, The "All of Us" Research Program. *New Engl J Med*, **381**, 668-676 (2019).
858

859 (48) K.M. Harrington, X.T. Nguyen, R.J. Song, K. Hannagan, R. Quaden, D.R. Gagnon, K.
860 Cho, J.E. Deen, S. Muralidhar, T.J. O'Leary, J.M. Gaziano, S.B. Whitbourne, Gender
861 Differences in Demographic and Health Characteristics of the Million Veteran Program Cohort.
862 *Womens Health Issues*, **29** Suppl 1, S56-66 (2019).
863

864 (49) T. Chikowore, A.B. Kamiza, O.H. Oduaran, T. Machipisa, S. Fatumo, Non-communicable
865 diseases pandemic and precision medicine: Is Africa ready? *EBioMedicine*, **65**, 103260
866 (2021).
867

868 (50) P. Song, A. Gupta, I.Y. Goon, M. Hasan, S. Mahmood, R. Pradeepa, S. Siddiqui, G.S.
869 Frost, D. Kusuma, M. Miraldo, F. Sassi, N.J. Wareham, S. Ahmed, R.M. Anjana, S. Brage,
870 N.G. Forouhi, S. Jha, A. Kasturiratne, P. Katulanda, K.I. Khawaja, M. Loh, M.K. Mridha, A.R.
871 Wickremasinghe, J.S. Kooner, J.C. Chambers, Data Resource Profile: Understanding the
872 patterns and determinants of health in South Asians - the South Asia Biobank. *Int J Epidemiol*,
873 **50**, 717-718e (2021).
874

875 (51) T.J. Littlejohns, J. Holliday, L.M. Gibson, S. Garratt, N. Oesingmann, F. Alfaro-Almagro,
876 J.D. Bell, C. Boulwood, 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

904 (57) Q. Feng, J.H. Kim, W. Omiyale, J. Bešević, M. Conroy, M. May, Z. Yang, S.Y. Wong, K.K.
905 Tsoi, N. Allen, B. Lacey, Raw and cooked vegetable consumption and risk of cardiovascular
906 disease: a study of 400,000 adults in UK Biobank. *Frontiers in Nutr*, **9**, 831470 (2022).
907

908 (58) M.K. Georgakis, R. Malik, D. Gill, N. Franceschini, C.L.M. Sudlow, M. Dichgans,
909 Interleukin-6 Signaling Effects on Ischemic Stroke and Other Cardiovascular Outcomes: A
910 Mendelian Randomization Study, *Circ Genom Precis Med*, **13**, e002872 (2020).
911

912 (59) S.C. Larsson, M. Bäck, J.M.B. Rees, A.M. Mason, S. Burgess, Body mass index and
913 body composition in relation to 14 cardiovascular conditions in UK Biobank: a Mendelian
914 randomization study. *Europ Heart J*, **41**, 221-226 (2020).
915

916 (60) G. Butler-Laporte, T. Nakanishi, V. Mooser, D.R. Morrison, T. Abdullah, O. Adeleye, N.
917 Mamlouk, N. Kimchi, Z. Afrasiabi, N. Rezk, A. Giliberti, A. Renieri, Y. Chen, S. Zhou, V.
918 Forgetta, J.B. Richards, Vitamin D and COVID-19 susceptibility and severity in the COVID-19
919 Host Genetics Initiative: A Mendelian randomization study. *PLoS Med*, **18**, e1003605 (2021).
920

921 (61) X. Meng, X. Li, M.N. Timofeeva, Y. He, A. Spiliopoulou, W.Q. Wei, A. Gifford, H. Wu, T.
922 Varley, P. Joshi, J.C. Denny, S.M. Farrington, L. Zgaga, M.G. Dunlop, P. McKeigue, H.
923 Campbell, E. Theodoratou, Phenome-wide mendelian-randomization study of genetically
924 determined vitamin D on multiple health outcomes using the UK Biobank study. *Int J*
925 *Epidemiol*, **48**, 1425-1434 (2019).
926

927 (62) G.D. Smith, Mendelian randomisation and vitamin D: the importance of model
928 assumptions, *Lancet Diabetes Endocrinol*, **11**, 14 (2023).
929

930 (63) S. Greenland, Multiple-bias modelling for analysis of observational data, *J Royal Stat Soc:*
931 *Series A*, **168**, 267-306, (2005).

932

933 (64) R. Clarke, M. Shipley, S. Lewington, L. Youngman, R. Collins, M. Marmot, R. Peto,
934 Underestimation of risk associations due to regression dilution in long-term follow-up of
935 prospective studies. *Am J Epidemiol*, **150**, 341-353 (1999).

936

937 (65) S. MacMahon, R. Peto, J. Cutler, R. Collins, P. Sorlie, J. Neaton, R. Abbott, J. Godwin,
938 A. Dyer, J. Stamler, Blood pressure, stroke, and coronary heart disease. Part 1, Prolonged
939 differences in blood pressure: prospective observational studies corrected for the regression
940 dilution bias. *Lancet*, **335**, 765-774 (1990).

941

942 (66) A.N. Phillips, G.D. Smith, How independent are "independent" effects? Relative risk
943 estimation when correlated exposures are measured imprecisely. *J Clin Epidemiol*, **44**, 1223-
944 1231 (1991).

945

946 (67) V. Codd, Q. Wang, E. Allara, C. Musicha, S. Kaptoge, S. Stoma, T. Jiang, S.E. Hamby,
947 P.S. Braund, V. Bountziouka, C.A. Budgeon, M. Denniff, C. Swinfield, M. Papakonstantinou,
948 S. Sheth, D.E. Nanus, S.C. Warner, M. Wang, A.V. Khera, J. Eales, W.H. Ouwehand, J.R.
949 Thompson, E. Di Angelantonio, A.M. Wood, A.S. Butterworth, J.N. Danesh, C.P. Nelson, N.J.
950 Samani, Polygenic basis and biomedical consequences of telomere length variation. *Nature*
951 *Genet*, **53**, 1425-1433 (2021).

952

953 (68) C.E. Rutter, L.A.C. Millard, M.C. Borges, D.A. Lawlor, Exploring regression dilution bias
954 using repeat measurements of 2858 variables in up to 49,000 UK Biobank participants, *Int J*
955 *Epidemiol*, **52**, 1545-1556 (2022).

956

957 (69) S. Tin Tin, G.K. Reeves, T.J. Key, Endogenous hormones and risk of invasive breast
958 cancer in pre- and post-menopausal women: findings from the UK Biobank. *Br J Cancer*, **125**,
959 126-134 (2021).

960

961 (70) K.A. Wartolowska, A.J.S. Webb, Midlife blood pressure is associated with the severity of
962 white matter hyperintensities: analysis of the UK Biobank cohort study. *Europ Heart J*, **42**,
963 750-757 (2021).

964

965 (71) M.J. Adams, W.D. Hill, D.M. Howard, H.S. Dashti, K.A.S. Davis, A. Campbell, T.K. Clarke,
966 I.J. Deary, C. Hayward, D. Porteous, M. Hotopf, A.M. McIntosh, Factors associated with
967 sharing e-mail information and mental health survey participation in large population cohorts.
968 *Int J Epidemiol*, **49**, 410-421 (2020).

969

970 (72) J. Beller, S. Geyer, J. Epping, Health and study dropout: health aspects differentially
971 predict attrition. *BMC Med Res Methodol*, **22**, 31 (2022).

972

973 (73) A.E. Taylor, H.J. Jones, H. Sallis, J. Euesden, E. Stergiakouli, N.M. Davies, S. Zammit,
974 D.A. Lawlor, M.R. Munafò, G. Davey Smith, K. Tilling, Exploring the association of genetic
975 factors with participation in the Avon Longitudinal Study of Parents and Children. *Int J*
976 *Epidemiol*, **47**, 1207-1216 (2018).

977

978 (74) G.J. Griffith, T.T. Morris, M.J. Tudball, A. Herbert, G. Mancano, L. Pike, G.C. Sharp, J.
979 Sterne, T.M. Palmer, G. Davey Smith, K. Tilling, L. Zuccolo, N.M. Davies, G. Hemani, Collider
980 bias undermines our understanding of COVID-19 disease risk and severity. *Nature Comms*,
981 **11**, 5749 (2020).

982

983 (75) M. Chadeau-Hyam, B. Bodinier, J. Elliott, M.D. Whitaker, I. Tzoulaki, R. Vermeulen, M.
984 Kelly-Irving, C. Delpierre, P. Elliott, Risk factors for positive and negative COVID-19 tests: a
985 cautious and in-depth analysis of UK Biobank data, *Int J Epidemiol*, **49**, 1454-1467 (2020).

986

987 (76) L.A.C. Millard, A. Fernández-Sanlés, A.R. Carter, R.A. Hughes, K. Tilling, T.P. Morris, D.
988 Major-Smith, G.J. Griffith, G.L. Clayton, E. Kawabata, G. Davey Smith, D.A. Lawlor, M.C.
989 Borges, Exploring the impact of selection bias in observational studies of COVID-19: a
990 simulation study, *Int J Epidemiol*, **52**, 44-57 (2023).

991

992 (77) K.A.S. Davis, J.R.I. Coleman, M. Adams, N. Allen, G. Breen, B. Cullen, C. Dickens, E.
993 Fox, N. Graham, J. Holliday, L.M. Howard, A. John, W. Lee, R. McCabe, A. McIntosh, R.
994 Pearsall, D.J. Smith, C. Sudlow, J. Ward, S. Zammit, M. Hotopf, Mental health in UK Biobank
995 - development, implementation and results from an online questionnaire completed by
996 157,366 participants: a reanalysis. *BJPsych Open*, **6**, e18 (2020).

997

998 (78) K. Rannikmäe, K. Ngoh, K. Bush, R. Al-Shahi Salman, F. Doubal, R. Flaig, D.E. Henshall,
999 A. Hutchison, J. Nolan, S. Osborne, N. Samarasekera, C. Schnier, W. Whiteley, T. Wilkinson,
1000 K. Wilson, R. Woodfield, Q. Zhang, N. Allen, C.L.M. Sudlow, Accuracy of identifying incident
1001 stroke cases from linked health care data in UK Biobank. *Neurology*, **95**, e697-e707 (2020).

1002

1003 (79) B. Rubbo, N.K. Fitzpatrick, S. Denaxas, M. Daskalopoulou, N. Yu, R.S. Patel, H.
1004 Hemingway, Use of electronic health records to ascertain, validate and phenotype acute
1005 myocardial infarction: A systematic review and recommendations. *Int J Cardiol*, **187**, 705-11
1006 (2015).

1007

1008 (80) T. Wilkinson, C. Schnier, K. Bush, K. Rannikmae, D.E. Henshall, C. Lerpiniere, N.E. Allen,
1009 R. Flaig, T.C. Russ, D. Bathgate, S. Pal, J.T. O'Brien, C.L.M. Sudlow, Identifying dementia
1010 outcomes in UK Biobank: a validation study of primary care, hospital admissions and mortality
1011 data. *Eur J Epidemiol*, **34**, 557-565 (2019).

1012

1013 (81) V. Kuan, S. Denaxas, A. Gonzalez-Izquierdo, K. Direk, O. Bhatti, S. Husain, S. Sutaria,
1014 M. Hingorani, D. Nitsch, C.A. Parisinos, R.T. Lumbers, R. Mathur, R. Sofat, J.P. Casas, I.C.K.

1015 Wong, H. Hemingway, A.D. Hingorani, A chronological map of 308 physical and mental health
1016 conditions from 4 million individuals in the English National Health Service. *Lancet*, **1**, e63-
1017 e77 (2019).

1018

1019 (82) S. Lewington, R. Clarke, N. Qizilbash, R. Peto, R. Collins, Age-specific relevance of usual
1020 blood pressure to vascular mortality: a meta-analysis of individual data for one million adults
1021 in 61 prospective studies. *Lancet*, **360**, 1903-1913 (2002).

1022

1023 (83) S. Lewington, G. Whitlock, R. Clarke, P. Sherliker, J. Emberson, J. Halsey, N. Qizilbash,
1024 R. Peto, R. Collins, Blood cholesterol and vascular mortality by age, sex, and blood pressure:
1025 a meta-analysis of individual data from 61 prospective studies with 55,000 vascular deaths.
1026 *Lancet*, **370**, 1829-1839 (2007).

1027

1028 (84) CHARGE and ISGC Consortium, Identification of additional risk loci for stroke and small
1029 vessel disease: a meta-analysis of genome-wide association studies. *Lancet Neurol*, **15**, 695-
1030 707 (2016).

1031

1032 (85) W. Luo, L. Gong, X. Chen, R. Gao, B. Peng, Y. Wang, T. Luo, Y. Yang, B. Kang, C. Peng,
1033 L. Ma, M. Mei, Z. Liu, Q. Li, S. Yang, Z. Wang, J. Hu, Lifestyle and chronic kidney disease: a
1034 machine learning modeling study, *Frontiers Nutr*, **9**, 918576 (2022).

1035

1036 (86) R.A. Shah, B. Asatryan, G. Sharaf Dabbagh, N. Aung, M.Y. Khanji, L.R. Lopes, S. van
1037 Duijvenboden, A. Holmes, D. Muser, A.P. Landstrom, A.M. Lee, P. Arora, C. Semsarian, V.K.
1038 Somers, A.T. Owens, P.B. Munroe, S.E. Petersen, C.A.A. Chahal, Frequency, penetrance,
1039 and variable expressivity of dilated cardiomyopathy-associated putative pathogenic gene
1040 variants in UK Biobank participants. *Circulation*, **146**, 110-124 (2022).

1041

1042 (87) D. Chahal, D. Sharma, S. Keshavarzi, F.A.Q. Arisar, K. Patel, W. Xu, M. Bhat, Distinctive
1043 clinical and genetic features of lean vs overweight fatty liver disease using the UK Biobank.
1044 *Hepatol Int*, **16**, 325-336 (2022).
1045
1046 (88) N.J. Thomas, S.E. Jones, M.N. Weedon, B.M. Shields, R.A. Oram, A.T. Hattersley,
1047 Frequency and phenotype of type 1 diabetes in the first six decades of life: a cross-sectional,
1048 genetically stratified survival analysis from UK Biobank, *Lancet Diabetes Endocrinol*, **6**, 122-
1049 129 (2018).
1050
1051 (89) P.R. Burton, A.L. Hansell, UK Biobank: the expected distribution of incident and prevalent
1052 cases of chronic disease and the statistical power of nested case-control studies, *UK Biobank*
1053 *Technical Reports*, (2005).
1054
1055 (90) K. Papier, A. Knuppel, A. Perez-Cornago, E.L. Watts, T.Y.N. Tong, J.A. Schmidt, N. Allen,
1056 T.J. Key, R.C. Travis, Circulating insulin-like growth factor-I and risk of 25 common conditions:
1057 outcome-wide analyses in the UK Biobank study. *Europ J Epidemiol*, **37**, 25-34 (2022).
1058
1059 (91) S. Floud, R.F. Simpson, A. Balkwill, A. Brown, A. Goodill, J. Gallacher, C. Sudlow, P.
1060 Harris, A. Hofman, S. Parish, G.K. Reeves, J. Green, R. Peto, V. Beral, Body mass index,
1061 diet, physical inactivity, and the incidence of dementia in 1 million UK women. *Neurology*, **94**,
1062 e123-e132 (2020).
1063
1064 (92) T. Strain, K. Wijndaele, S.J. Sharp, P.C. Dempsey, N. Wareham, S. Brage, Impact of
1065 follow-up time and analytical approaches to account for reverse causality on the association
1066 between physical activity and health outcomes in UK Biobank, *Int J Epidemiol*, **49**, 162-172
1067 (2020).
1068

1069 (93) K. Bleicher, R. Summerhayes, S. Baynes, M. Swarbrick, T. Navin Cristina, H. Luc, G.
1070 Dawson, A. Cowle, X. Dolja-Gore, M. McNamara, Cohort Profile Update: The 45 and Up
1071 Study. *Int J Epidemiol*, **52**, e92-e101 (2023).

1072

1073 (94) T.J.B. Dummer, P. Awadalla, C. Boileau, C. Craig, I. Fortier, V. Goel, J.M.T. Hicks, S.
1074 Jacquemont, B.M. Knoppers, N. Le, T. McDonald, J. McLaughlin, A.M. Mes-Masson, A.M.
1075 Nuyt, L.J. Palmer, L. Parker, M. Purdue, P.J. Robson, J.J. Spinelli, D. Thompson, J. Vena, M.
1076 Zawati, The Canadian Partnership for Tomorrow Project: a pan-Canadian platform for
1077 research on chronic disease prevention. *CMAJ*, **190**, e710-717 (2018).

1078

1079 (95) H.S. Feigelson, C.L. Clarke, S.K. Van Den Eeden, S. Weinmann, A.N. Burnett-Hartman,
1080 S. Rowell, S.G. Scott, L.L. White, M. Ter-Minassian, S.A.A. Honda, D.R. Young, A. Kamineni,
1081 T. Chinn, A. Lituev, A. Bauck, E.A. McGlynn, The Kaiser Permanente Research Bank Cancer
1082 Cohort: a collaborative resource to improve cancer care and survivorship. *BMC Cancer*, **22**,
1083 209 (2022).

1084

1085

1086

1087

1088

1089 **Acknowledgements**

1090 We thank Jenny Mills and Alicia Motley for constructing the figures and George Davey Smith
1091 for helpful suggestions. Additional thanks to the UK Biobank Access team for their tireless

1092 work on research registrations, applications and output. The authors would like to thank the
1093 500,000 participants in the UK Biobank study for their enormous generosity and altruism and
1094 their continued interest, support and involvement.

1095 **Funding:** UK Biobank has core funding from the Medical Research Council, Wellcome, British
1096 Heart Foundation, Cancer Research UK and National Institute for Health Research.

1097 **Competing interests:** All authors are past or present members of the UK Biobank Strategic
1098 Oversight Committee or the UK Biobank Senior Team. J.D. serves on scientific advisory
1099 boards for AstraZeneca and Novartis and consults; British Heart Foundation Centre of
1100 Research Excellence, University of Cambridge; the National Institute for Health and Care
1101 Research Blood and Transplant Research Unit in Donor Health and Behaviour, University of
1102 Cambridge; Health Data Research UK; Wellcome Genome Campus and University of
1103 Cambridge; Department of Human Genetics, Wellcome Sanger Institute, Hinxton, UK. R.C. is
1104 named on US patent #9957563B2 regarding a statin-related myopathy genetic test but any
1105 share in royalty and other payments has been waived in favour of the Nuffield Department of
1106 Population Health, University of Oxford. Financial Relationships: UK Biobank Consortium
1107 Funding and Enhancements (Novartis, Regeneron Pharmaceuticals, Merck, AstraZeneca). R.
1108 E. M. is a scientific advisor to Optima Partners and the Epigenetic Clock Development
1109 Foundation and has received a speaker fee from Illumina. P.M has received consultancy or
1110 speaker fees from Roche, Merck, Biogen, Rejuveron, Sangamo, Nodthera, Novartis and
1111 Biogen. P.M. has received research or educational funds from Biogen, Novartis, Merck and
1112 GlaxoSmithKline.

1113

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 data¹

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

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)