Comparison of sociodemographic and health-related characteristics of UK Biobank participants with the general population

Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants with the General Population

Anna Fry, Thomas J Littlejohns*, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E Allen

Anna Fry and Thomas J Littlejohns contributed equally to this work.

*Correspondence to Dr. Thomas J. Littlejohns, Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK (e-mail: thomas.littlejohns@ndph.ox.ac.uk; phone: +44 (0) 1865 743645)

Author affiliations: Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK (Anna Fry, Thomas J Littlejohns, Ligia Adamska, Tim Sprosen, Rory Collins and Naomi E Allen); Cancer Research UK, London, UK (Anna Fry); Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK (Cathie Sudlow); and UK Biobank Coordinating Centre, Stockport, UK (Nicola Doherty).

Funding: UK Biobank was established by the Wellcome Trust; Medical Research Council; Department of Health and the Scottish Government. UK Biobank has also received funding from the Welsh Assembly Government; British Heart Foundation and Diabetes UK.

The authors have no conflicts of interest to declare.

Running head: The Representativeness of UK Biobank

Abbreviations: Health Survey for England; HSE: International Classification of Diseases-10; ICD-10

Abstract

UK Biobank is a population-based cohort of 500,000 participants recruited between 2006 and 2010. Approximately 9.2 million individuals aged 40-69 years who lived within 25 miles of the 22 assessment centres in England, Wales and Scotland were invited, and 5.4% participated in the baseline assessment. The representativeness of the UK Biobank cohort was investigated by comparing demographic characteristics between non-responders and responders. Sociodemographic, physical, lifestyle and health-related characteristics of the cohort were compared with nationally representative data sources. UK Biobank participants were more likely to be older, women and to live in less socioeconomically deprived areas than non-participants. Compared with the general population, participants were less likely to be obese, smoke, drink alcohol on a daily basis and had fewer self-reported health outcomes. Rates of all-cause mortality and total cancer incidence (at age 70-74 years) were 46.2% and 11.8% lower in men, and 55.5% and 18.1% lower in women, respectively, than the general population of the same age. UK Biobank is not representative of the sampling population, with evidence of a 'healthy volunteer' selection bias. Nonetheless, the valid assessment of exposure-disease relationships may be widely generalizable and does not require participants to be representative of the population at large.

Keywords: UK Biobank; Representativeness; Lifestyle; Sociodemographics; Cancer; Mortality

Word count: Abstract - 199; Main text – 2,616

UK Biobank is a large prospective study, primarily established to investigate the genetic and lifestyle determinants of a wide range of diseases of middle and later life (1). The open access resource involves 500,000 men and women who were aged 40-69 years when recruited during 2006 and 2010 throughout England, Wales and Scotland. Extensive questionnaire data, physical measures and biological samples were collected at recruitment, with ongoing enhanced data collection in large subsets of the cohort, including a repeat baseline assessment, genotyping, biochemical assays, web-based questionnaires, physical activity monitoring and multimodal imaging. All participants are followed-up for health outcomes through linkage to national electronic health-related datasets.

The aim of the current study is to examine and quantify whether the UK Biobank cohort differs from the sampling frame on a range of characteristics due to the 'healthy volunteer effect' (2), whereby volunteering participants, tend to be, on average, more health-conscious than non-participants (3). To investigate this, the distribution of a range of sociodemographic, physical, lifestyle and health-related characteristics was compared between UK Biobank participants and (a) those invited to join UK Biobank and (b) findings from nationally representative surveys.

METHODS

UK Biobank sent postal invitations to 9,238,453 individuals registered with the National Health Service who were aged 40-69 years and lived within approximately 25 miles of one of 22 assessment centres located throughout England, Wales and Scotland. Approval was

obtained from the National Information Governance Board for Health and Social Care and the North-West Multicentre Research Ethics Committee for local National Health Service Primary Care Trusts to provide UK Biobank with contact details of people within the eligible age range and for us to retain limited information on non-responders. Overall, 503,317 participants consented to join the study and attended an assessment centre between 2006 and 2010, resulting in a participation rate of 5.44% (see Web Figure 1 for flow chart demonstrating responses to invitations).

Anonymised data on sex, month and year of birth, Townsend deprivation index (an indicator of socioeconomic status) and geographic location are stored by UK Biobank and are available for 8,761,869 out of 9,238,453 (94.8%) of the individuals sent an invitation letter, allowing these characteristics to be compared between non-participating invitees and participants. The distribution of a range of sociodemographic, physical, lifestyle and health-related characteristics of the UK Biobank cohort was also compared with publicly available summary data from nationally representative population-based surveys and the UK Census. We selected summary survey data that matched the UK Biobank cohort as closely as possible with regard to population demographics (i.e. mixed gender and aged 40-69 years) and period of data collection (2006 to 2010). Where certain characteristics from the national survey summary data were only available in pre-specified aggregated age and sex subgroups, UK Biobank data was stratified into similar groups for comparative purposes. Formal statistical tests of the difference in characteristics between UK Biobank and national data were not performed because of the lack of variance measures required to test for differences between means, such as standard deviations, from the comparison populations.

The UK Census collects individual and household-level demographic data every 10 years for the whole UK population. Data on ethnicity was obtained from the 2001 and the 2011 UK Census for England, Scotland and Wales (as these reflect the census years before and

immediately after the recruitment period) (4,5). Data on property ownership status was obtained from the 2001 UK Census for England and Wales only (as 2011 UK Census data on property ownership was not available for the appropriate age groups). Data on anthropometric measures, smoking status, alcohol consumption and prevalence of self-reported health outcomes were obtained from the Health Survey for England (HSE) performed in 2006, 2008, 2009 and 2010 (6–9). The HSE consists of an annual cross-sectional survey of a small (n=~5,000 to ~15,000) representative population of England through a two-stage random probability sampling process, with different data items collected on a different population each year (10,11). Since 2003, HSE has incorporated weighting to account for non-response bias (12). This includes different weights for non-responding households, non-responding individuals in responding households and non-response at different stages of data collection. For a detailed description of the data collection methods in UK Biobank and national surveys, see Web Table 1.

Data on age and sex-specific all-cause mortality and cancer incidence rates for England were obtained from the Office for National Statistics for 2012 (as this date represents the midpoint of the follow-up period for UK Biobank participants) (13,14). For all-cause mortality, follow-up time (person-years) in the UK Biobank cohort was calculated as the age at recruitment to the age at death or date of complete follow-up (30th November 2015), whichever came first; for cancer incidence rates, follow-up time was defined as the age at recruitment to the age at first cancer diagnosis, death or date of complete follow-up (30th September 2014), whichever came first (among individuals with no cancer at recruitment based on cancer registry data). Incidence rates were calculated for total cancer (excluding non-melanoma skin cancer), defined using International Classification of Diseases-10 (ICD-10) C00-C97 excluding C44, and common subtypes; prostate (ICD-10; C61), breast (ICD-10; C50), colorectal (ICD-10; C18-20), lung (ICD-10; C33-34), endometrium (ICD-10; C54) and kidney (ICD-10; C64).

Ethical statement

UK Biobank received approval from the National Information Governance Board for Health and Social Care and the National Health Service North West Centre for Research Ethics Committee (Ref: 11/NW/0382).

RESULTS

Characteristics of UK Biobank participants versus non-participating invitees

Of the 9,238,453 men and women invited to join UK Biobank, 503,317 (5.45%) consented and were recruited between 2006-2010. Overall, the participation rate was higher in women (6.4% and 5.1% in women and men, respectively) (Figure 1A), higher in older age groups (9% in those aged ≥60 years and 3% in those aged 40-44 years) (Figure 1B) and higher in less socioeconomically deprived areas (8.3% in those from the least deprived areas and 3.1% in those from the most deprived areas) (Figure 1C). Participation rates showed regional differences, being highest in South West England (9.6%) and East Scotland (8.2%) and lowest in West Scotland (4.3%), London, West Midlands and North West England (all 4.7%) (Figure 1D, also see Web Table 2 for further details).

Characteristics of UK Biobank participants compared with national survey data

Sociodemographic factors

In UK Biobank, 94.6% of participants were of white ethnic background, which was similar to that of the national population of the same age range taken from the 2001 UK Census (94.5%), but somewhat higher than the 2011 Census (91.3%; Table 1). UK Biobank participants were also more likely to own their property outright and were less likely to have a

mortgage or loan, have shared ownership or live in rental accommodation than the general population of the same age range (Table 2).

Physical characteristics

UK Biobank participants were, on average, taller, leaner and had a smaller waist circumference than the general population, based on the HSE 2008 (Table 3). For example, mean body mass index (defined as weight [kg]/height $[m]^2$) in UK Biobank men and women aged 55-64 years was 27.9 and 27.7, respectively, compared with 28.5 and 28.0 in the general population, based on data from the HSE 2008. UK Biobank men and women were also less likely to be obese (defined as body mass index ≥30 kg/m^2) across all age groups examined, compared with the general population. For example, for men aged 45-54 years, the prevalence of obesity was 25.6% in UK Biobank and 31.5% in the general population, with corresponding values of 23.0% and 32.2% for women, respectively (Web Table 3).

Lifestyle characteristics

UK Biobank men and women were less likely to be current smokers across all ages than the general population based on data from the HSE 2008 (Figure 2). For example, for men aged 45-54 years, the prevalence of current smoking was 15% in UK Biobank and 22% in the general population (Figure 2C); the corresponding values for women were 11% and 20%, respectively (Figure 2D). However, younger smokers (aged 45-54 years) in UK Biobank smoked more heavily (≥20 cigarettes per day) than the general population (46% and 41%, respectively for men; 32% and 28%, respectively for women). This difference persisted for older women aged 55-64 years (31% and 23% in UK Biobank and the general population, respectively) but not for older men (47% and 49%, respectively; Web Figure 2). UK Biobank

participants were also less likely to be never-drinkers but were less likely to drink every day compared with the general population included in the HSE 2008 (Table 4).

Self-reported health outcomes

UK Biobank participants had a lower prevalence of self-reported health outcomes, including cardiovascular disease, stroke, hypertension, diabetes, chronic kidney disease and respiratory disease compared with the general population, as obtained from various surveys of the HSE performed in 2006, 2009 and 2010 (Table 5). For example, for those aged 45-54 years, the prevalence of self-reported cardiovascular disease in UK Biobank compared with the general population was 4.6% and 10.9%, respectively for men and 2.4% and 10.3%, respectively for women.

All-cause mortality and cancer incidence rates

UK Biobank participants were followed up for a mean of 6.77 years (SD=1.01) and 5.53 years (SD=1.10) for all-cause mortality and incident cancer, respectively. Compared with national death rates in those aged 70-74 years, all-cause mortality in UK Biobank was 46.2% lower in men and 55.5% lower in women (Figure 3A and B; also see Web Table 4 for further details of age-specific mortality rates). The total cancer incidence rate was also lower than the general population, being 11.8% and 18.1% lower at ages 70-74 years in men and women, respectively (Figure 4A and B; also see Web Table 5 for further details of age-specific cancer incidence rates). A similar pattern was observed for cancers of the colorectum, kidney and endometrium (Web Figure 3). Lung cancer incidence rates in UK Biobank were markedly lower for both men and women, whilst rates of female breast cancer were similar to the national average, with the exception of ages 45-49 years, where the rate was higher in the UK

Biobank cohort. In contrast, prostate cancer incidence was higher in UK Biobank compared with national rates across all age groups examined.

DISCUSSION

The participation rate in UK Biobank was higher among women, older age groups and among those living in less socioeconomically deprived areas. For example, men aged 45-54 years were less likely to be obese (25.6% in UK Biobank versus 31.5% in the general population), and less likely to be current smokers (15% versus 22%), with similar findings observed for women and older age groups. Furthermore, compared with the general population of the same age, UK Biobank participants were less likely to drink on a daily basis and had fewer self-reported health outcomes. Linkage of UK Biobank participants with their health records during an average of 6-7 years follow-up also showed lower rates of all-cause mortality and total cancer incidence than the general population of the same age group.

These findings are consistent with the well-established "healthy volunteer" effect, which has been demonstrated in other volunteer-based cohort studies (15–17). Other prospective studies have also reported lower all-cause mortality and incident cancer rates compared with national rates (18–21). The only health outcome examined that was higher in UK Biobank than the general population was prostate cancer, which might reflect higher rates of voluntary prostate-specific antigen testing (and subsequent prostate cancer diagnosis) among health-conscious men. In contrast, lung cancer incidence rates were markedly lower in UK Biobank across all age and sex groups, almost certainly caused by the lower prevalence of smoking compared with the general population.

Because UK Biobank participants are, on average, more health-conscious than the general population, this cohort is not best-placed to estimate generalizable prevalence or incidence rates of disease (although some health-related characteristics of the UK Biobank cohort, such

as the prevalence of self-reported pain, have previously been shown to be similar to that of the national population (22)). In order for a cohort study to produce generalizable associations of exposures with disease, what is important is that sufficiently large numbers of individuals with different levels of exposures are investigated with high internal validity (23–26). Indeed, if one was interested in investigating the association of ethnicity with subsequent disease risk, the most appropriate study design would be to recruit a large number of people from different ethnic backgrounds rather than have a representative, largely white population. As UK Biobank is primarily designed for investigating exposure-disease associations, the lack of representativeness should not be regarded as a limitation (27,28). As with all observational studies, it is incumbent on researchers to acknowledge potential sources of bias on a case-by-case basis that might affect the generalisability of exposure-disease associations, such as residual confounding, reverse causation and self-selection bias (24,29). Indeed, although still in the early stages as a prospective study, initial publications show expected associations between cardiometabolic morbidity, self-reported health and smoking with mortality risk (30,31).

This study provides an overview of the representativeness of the UK Biobank cohort on a variety of key characteristics in comparison with the general population using data from nationally representative surveys. We expect these findings will be used by researchers to inform the interpretation of results or, in some instances, to help generate weighted results (for example, in order to estimate nationally-representative disease rates). We were able to compare participation rates for key sociodemographic characteristics (such as age, sex, socioeconomic status and geographic location) due to the availability of such data for the total sampling frame. The availability of follow-up health data enabled us to compare death and cancer incidence rates with age- and sex-specific national rates and the large size of the cohort meant that sufficient numbers of cases had accrued to investigate common cancer subtypes.

All participants are flagged by national death and cancer registries, and loss to follow-up due to emigration is minimal (0.3% of the cohort). Further follow-up is required to determine whether this 'healthy volunteer effect' attenuates over time (owing to the development of chronic disease as the cohort ages), which has been observed in previous studies (18,20,32).

One limitation of the study is that the national survey data (available from the UK Census and the HSE) were presented in pre-specified age groups, thereby restricting the comparisons that could be performed. For the majority of characteristics, comparable national survey data was only available for England, although only 11% of participants were recruited in Wales and Scotland and the distribution of most characteristics is similar across the countries. It is also possible that differences in the wording of questions, answer choices and data collection methods might have influenced the comparability of certain characteristics between the national surveys and the UK Biobank cohort. For example, HSE primarily consisted of a verbal interview that enabled the interviewer to probe the participant for further information, whereas all of the characteristics presented here in UK Biobank were collected via a touchscreen questionnaire with the exception of self-reported illnesses, which was collected through a verbal interview with a trained nurse.

In conclusion, the UK Biobank cohort is not representative of the general population on a number of sociodemographic, physical, lifestyle and health-related characteristics. UK Biobank participants generally live in less socioeconomically deprived areas and are less likely to be obese, to smoke, to drink on a daily basis and to have fewer self-reported diseases. All-cause mortality is approximately half and total cancer incidence rates are approximately 10-20% lower that of the UK population as a whole. Although UK Biobank is not suitable for deriving generalizable disease prevalence and incidence rates, its large size and heterogeneity of exposure measures provide valid scientific inferences of associations between exposures and health outcomes that are generalizable to other populations.

Acknowledgements

Author affiliations: Clinical Trial Service Unit and Epidemiological Studies Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK (Anna Fry, Thomas J Littlejohns, Ligia Adamska, Tim Sprosen, Rory Collins, Naomi E Allen); Cancer Research UK, London, UK (Anna Fry); Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK (Cathie Sudlow); UK Biobank Coordinating Centre, Stockport, UK (Nicola Doherty).

This research has been conducted using the UK Biobank resource. The England and Wales censuses are undertaken by the Office for National Statistics and the Scottish census is organised by the National Records of Scotland, formally the General Register Office for Scotland. The Health Survey for England is carried out by NatCen Social Research on behalf of the Health and Social Care Information Centre and is funded by NHS digital.

Special thanks to Professor Sir Andy Haines for providing valuable comments when drafting the manuscript. Acknowledgements to the members of the UK Biobank Steering Committee; Professor John Danesh, Professor Paul Elliot, Professor John Gallacher, Professor Jane Green, Professor Paul Matthews, Dr Tim Peakman and Professor Jill Pell.

The authors have no conflicts of interest to declare.

References

1. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Med.* 2015;12(3):1–10.
2. Delgado-Rodriguez M. Bias. *J. Epidemiol. Community Heal.* 2004;58(8):635–641.
3. Manolio TA, Weis BK, Cowie CC, et al. New models for large prospective studies: Is

there a better way? *Am. J. Epidemiol.* 2012;175(9):859–866.

4. Office for National Statistics ; General Register Office for Scotland ; Northern Ireland Statistics and Research Agency (2005): 2001 Census aggregate data. (https://discover.ukdataservice.ac.uk/doi/2001-census-aggregate). Updated June 2016, Accessed September 1, 2016.
5. Office for National Statistics ; National Records of Scotland ; Northern Ireland Statistics and Research Agency: 2011 Census aggregate data. (https://discover.ukdataservice.ac.uk/doi/2011-census-aggregate). Published 2005. Accessed September 1, 2016.
6. Health Survey for England - 2006: CVD and risk factors for adults, obesity and risk factors for children. (http://content.digital.nhs.uk/catalogue/PUB01213). Published January 31, 2008. Accessed December 22, 2015.
7. Health Survey for England - 2008: Physical activity and fitness. (http://content.digital.nhs.uk/catalogue/PUB00430). Published December 17, 2009. Accessed December 22, 2015.
8. Health Survey for England - 2009: Health and lifestyles. (http://content.digital.nhs.uk/catalogue/PUB00414). Published December 16, 2010. Accessed December 22, 2015.
9. Health Survey for England - 2010: Respiratory health. (http://content.digital.nhs.uk/catalogue/PUB03023). Published December 15, 2011. Accessed December 22, 2015.
10. Mindell J, Aresu M, Becares L, et al. Representativeness of participants in a cross-sectional health survey by time of day and day of week of data collection. *Eur. J. Public Health.* 2012;22(3):364–369.
11. Mindell J, Biddulph JP, Hirani V, et al. Cohort profile: the health survey for England.

Int. J. Epidemiol. 2012;41(6):1585–93.

12. Department of Health. Health Survey for England 2003 - Vol 3. Methodology and documentation. 2004;1–197. (http://webarchive.nationalarchives.gov.uk/). Published 17 December 2004. Accessed December 22, 2015.
13. Office for National Statistics, Death Registration Summary Tables - England and Wales, 2012. (http://www.ons.gov.uk/). Published July 10, 2013. Accessed May 1, 2016.
14. Office for National Statistics, Cancer Statistics Registrations, England, 2012. (http://www.ons.gov.uk/). Published June 19, 2014. Accessed May 1, 2016.
15. Andreeva VA, Salanave B, Castetbon K, et al. Comparison of the sociodemographic characteristics of the large NutriNet-Santé e-cohort with French Census data: the issue of volunteer bias revisited. *J. Epidemiol. Community Health*. 2015;69(9):893–8.
16. Mishra GD, Hockey R, Powers J, et al. Recruitment via the Internet and social networking sites: the 1989-1995 cohort of the Australian Longitudinal Study on Women's Health. *J. Med. Internet Res.* 2014;16(12):e279.
17. Brown WJ, Bryson L, Byles JE, et al. Women's Health Australia: Recruitment for a National Longitudinal Cohort Study. *Women Health*. 1999;28(1):23–40.
18. Struijk E, May A, Beulens J, et al. Mortality and cancer incidence in the EPIC-NL cohort: impact of the healthy volunteer effect. *Eur. J. Public Health*. 2014;25(1):144–149.
19. Otto SJ, Schroder FH, de Koning HJ. Low all-cause mortality in the volunteer-based Rotterdam section of the European randomised study of screening for prostate cancer: self-selection bias? *J Med Screen*. 2004;11(2):89–92.
20. Pinsky PF, Miller A, Kramer BS, et al. Evidence of a healthy volunteer effect in the prostate, lung, colorectal, and ovarian cancer screening trial. *Am. J. Epidemiol.*

2007;165(8):874–881.

21. Lindsted KD, Fraser GE, Steinkohl M, et al. Healthy volunteer effect in a cohort study: Temporal resolution in the Adventist Health Study. *J. Clin. Epidemiol.* 1996;49(7):783–790.

22. Macfarlane GJ, Beasley M, Smith BH, et al. Can large surveys conducted on highly selected populations provide valid information on the epidemiology of common health conditions? An analysis of UK Biobank data on musculoskeletal pain. *Br. J. Pain.* 2015;9(4):203–212.

23. Rothman K, Gallacher J, Hatch E. Why representativeness should be avoided. *Int. J. Epidemiol.* 2013;42(4):1012–1014.

24. Ebrahim S, Davey Smith G. Should we always deliberately be non-representative? *Int. J. Epidemiol.* 2013;42(4):1022–1026.

25. Elwood J. On representativeness. *Int. J. Epidemiol.* 2013;42:1014–1015.

26. Richiardi L, Pizzi C, Pearce N. Representativeness is usually not necessary and often should be avoided. *Int. J. Epidemiol.* 2013;42(4):1018–1022.

27. Allen N, Sudlow C, Downey P, et al. UK Biobank: Current status and what it means for epidemiology. *Heal. Policy Technol.* 2012;1(3):123–126.

28. Collins R. What makes UK Biobank special? *Lancet*. 2012;379(9822):1173–4.

29. Hernán M a., Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15(5):615–25.

30. Emerging Risk Factors Collaboration, Di Angelantonio E, Kaptoge S, et al. Association of Cardiometabolic Multimorbidity With Mortality. *JAMA.* 2015;314(1):52–60.

31. Ganna A, Ingelsson E. 5 year mortality predictors in 498 103 UK Biobank participants: a prospective population-based study. *Lancet*. 2015;6736(15):1–8.

32. Burnell M, Gentry-Maharaj A, Ryan A, et al. Impact on mortality and cancer incidence rates of using random invitation from population registers for recruitment to trials. *Trials*. 2011;12(1):61.

Figure legends

Figure 1. Participation rate according to A) sex, B) age at recruitment, C) Townsend deprivation score and D) region of residence. For numerators and denominators see Web Table 1. Participants were assigned a Townsend deprivation score corresponding to the output area of their residential postcode; most deprived = ≥2.00, average = −2.00 to 1.99, least deprived = <−2.00.

Figure 2. Comparison of smoking status in UK Biobank participants with data from the Health Survey for England (HSE), 2008 for A) men aged 45-54 years, B) women aged 45-54 years, C) men aged 55-64 years and D) women aged 55-64 years. HSE estimates weighted for non-response bias. Excludes 1,899 UK Biobank participants aged 45-64 with missing data for smoking status or who responded 'prefer not to answer'. Number of participants: A) UK Biobank = 62,004, HSE = 1,206; B) UK Biobank = 79,755, HSE = 1,233; C) UK Biobank = 94,907, HSE = 1,085; D) UK Biobank = 116,246, HSE = 1,123. See reference 9 for further information about HSE data.

Figure 3. Comparison of (A,B) mortality rates per 1000 person-years by age at death for UK Biobank participants and the population of England and Wales in 2012 from the Office for National Statistics for A) men and B) women. Total number of deaths in UK Biobank participants aged 45-74: men = 8,291, women = 5,380. See reference 15 for further information about death registration data.

Figure 4. Comparison of incidence rates for all cancers excluding non-melanoma skin cancer (NMSC) per 100,000 person-years by age at cancer diagnosis for UK Biobank participants and the population of England in 2012 from the Office for National Statistics for A) men and B) women. Total number of all incident cancers excluding NMSC in UK Biobank participants aged 45-74: men = 11,436, women = 10,592. See reference 16 for further information about cancer registration data.

Table 1. Comparison of Self-Reported Ethnic Origin of UK Biobank Participants With Census Data for the Age Group 40-69 Years in England, Wales and Scotland in 2001 and 2011[a,b]

	UK Biobank (n=499,877)		**2001 UK Census (n=20,198,307)**		**2011 UK Census (n=23,146,612)**	
Ethnicity	**N**	**%**	**N**	**%**	**N**	**%**
White[c]	472,837	94.6	19,085,322	94.5	21,133,317	91.3
Black or black British[d]	8,066	1.6	302,073	1.5	565,777	2.4
Mixed[e]	2,958	0.6	82,389	0.4	191,085	0.8
Indian	5,951	1.2	325,651	1.6	442,338	1.9
Pakistani	1,837	0.4	147,695	0.7	239,166	1.0
Bangladeshi	236	0.0	46,220	0.2	75,919	0.3
Chinese	1,574	0.3	70,572	0.3	109,412	0.5
Other Asian	1,858	0.4	73,917	0.4	240,324	1.0
Other ethnic group	4,560	0.9	64,468	0.3	149,274	0.6

[a] See references 4-7 for further information about census data.

[b] Excludes 2,778 UK Biobank participants aged 40-69 with missing data for ethnicity, who responded ‘prefer not to answer’ or who responded ‘do not know’.

[c] Includes the following categories; white British, white Irish and other white background.

[d] Includes the following categories; Caribbean, African and other black background.

[e] Includes the following categories; white and black Caribbean, white and black African, white and Asian and other mixed background.

Table 2. Comparison of Property Ownership Status of UK Biobank Participants With Census Data for Age Group 50-64 Years in England and Wales in 2001[a,b]

	UK Biobank (n=284,400)		**2001 UK census (n=9,098,70)**	
Property ownership status	**N**	**%**	**N**	**%**
Owns outright	161,318	56.7	3,690,996	40.6
Owns with mortgage or loan	96,427	33.9	3,599,560	39.6
Shared ownership	682	0.2	33,971	0.4
Rented from council (local authority), housing association or registered social landlord	16,407	5.8	1,187,422	13.1
Rented from private landlord /letting agency	7,514	2.6	418,900	4.6
Live in accommodation rent-free	2,052	0.7	117,344	1.3
Living in a communal establishment[c]	N/A	N/A	49,877	0.5

N/A; not available

[a] See reference 4 for further information about census data.

[b] Excludes 4,313 UK Biobank participants aged 50-64 with missing data for property ownership status, who responded 'none of the above' or who responded 'prefer not to answer'.

[c] Category not included in UK Biobank questionnaire.

Table 3. Comparison of Measured Mean Weight, Height and Waist Circumference by Age and Sex for UK Biobank Participants (Recruited 2006-2010) With Data From the Health Survey for England, 2008[a,b]

	Age 45–54 years				Age 55–64 years			
	UK Biobank		HSE		UK Biobank		HSE	
Anthropometric measures by sex	**N**	**Mean (SD)[c]**	**N**	**Mean**	**N**	**Mean (SD)[c]**	**N**	**Mean**
Men								
Body mass index[d]	61,860	27.8 (4.4)	1,059	28.1	94,776	27.9 (4.3)	968	28.5
Weight (kg)[e]	61,929	86.9 (15.1)	1,079	86.4	94,875	86.0 (14.3)	980	86.7
Height (cm)[f]	61,919	176.5 (6.9)	1,076	175.1	94,901	175.4 (6.7)	981	174.0
Waist (cm)[g]	62,010	96.1 (11.5)	845	100.3	95,031	97.7 (11.4)	755	102.9
Women								
Body mass index[d]	79,714	26.9 (5.4)	1,057	27.7	116,303	27.3 (5.1)	985	28.0
Weight (kg)[e]	79,738	71.8 (14.8)	1,067	72.8	116,344	71.6 (13.8)	995	72.3
Height (cm)[f]	79,792	163.4 (6.3)	1,097	162.0	116,429	162.0 (6.2)	1,016	160.5
Waist (cm)[g]	79,809	83.6 (12.8)	850	89.3	116,471	85.5 (12.5)	784	91.6

[a] See reference 9 for further information about HSE data.

[b] HSE data is weighted for nonresponse bias.

[c] Standard deviation values were not available from the HSE.

[d] Excludes 2,158 UK Biobank participants aged 45-64 with missing data for body mass index.

[f] Excludes 1,925 UK Biobank participants aged 45-64 with missing data for weight.

[e] Excludes 1,770 UK Biobank participants aged 45-64 with missing data for height.

[g] Excludes 1,482 UK Biobank participants aged 45-64 with missing data for waist circumference and 8 people for whom values outside the range 50–180 cm were obtained.

Table 4. Comparison of Frequency of Alcohol Consumption by Age and Sex in UK Biobank Participants With Data From the Health Survey for England 2008[a,b,c]

	Men				Women			
	Age 45–54 years		Age 55–64 years		Age 45–54 years		Age 55–64 years	
Alcohol consumption	UK Biobank (n=62,082) %	HSE (n=1,204) %	UK Biobank (n=95,207) %	HSE (n=1,085) %	UK Biobank (n=79,904) %	HSE (n=1,232) %	UK Biobank (n=116,605) %	HSE (n=1,123) %
Daily[d]	21.2	24	28.3	30	14.5	16	17.6	18
3–4 days a week	26.8	21	26.9	15	21.9	16	20.9	15
1–2 days a week	28.2	29	24.2	26	27.6	26	24.9	23
1–3 times a month	10.0	10	8.0	9	13.9	12	12.2	11
Special occasions[e]	7.4	9	6.8	11	13.8	16	15.0	21
Never[f]	6.6	8	5.8	9	8.3	12	9.5	12

[a] See reference 9 for further information about HSE data.

[b] HSE estimates are weighted for nonresponse bias.

[c] Excludes 1,013 UK Biobank participants aged 45-64 with missing data for alcohol intake or responded 'prefer not to answer'.

[d] HSE categories 'almost every day' and '5 or 6 days a week' defined as 'daily'.

[e] HSE categories 'once every couple of months' and 'once or twice in the past year' defined as 'special occasions'.

[f] HSE category 'not at all in the last 12 months/non–drinker' defined as 'never.

Table 5. Comparison of Self-Reported Disease by Age and Sex in UK Biobank Participants With Data From the Health Survey for England Performed in 2006, 2009 or 2010[a,b,c]

	Men				**Women**			
	Age 45–54 years		**Age 55–64 years**		**Age 45–54 years**		**Age 55–64 years**	
	UK Biobank	**HSE**	**UK Biobank**	**HSE**	**UK Biobank**	**HSE**	**UK Biobank**	**HSE**
Self-reported disease	**%**	**%**	**%**	**%**	**%**	**%**	**%**	**%**
Cardiovascular disease[d]	4.6	10.9	11.5	18.5	2.4	10.3	5.0	15.2
Ischaemic heart disease[e]	2.8	3.6	7.9	10.6	0.9	1.3	2.6	3.5
Stroke	0.8	1.2	1.9	3.0	0.6	0.9	1.0	2.3
Angina	1.8	2.4	5.3	8.0	0.7	1.2	2.1	3.2
Myocardial infarction	1.7	2.1	4.5	6.3	0.3	0.7	0.9	1.6
Abnormal heart	1.5	5.7	3.1	6.3	1.4	5.7	2.2	7.3

rhythm								
Hypertension[f]	21.2	27	34.4	39	15.4	16	27.4	29
Diabetes	4.5	8.1	7.8	10.5	2.4	3.5	6.3	8.0
Chronic kidney disease	0.2	1.1	0.3	1.5	0.2	1.2	0.2	1.9
Asthma[f]	11.7	12	9.9	13	13.0	16	11.8	15
Chronic obstructive pulmonary disease	0.1	1	0.4	3	0.1	0	0.4	2

[a] See references 8, 10 and 11 for further information about HSE data.

[b] HSE estimates are weighted for nonresponse bias.

[c] HSE 2006 data were used for CVD, IHD, stroke, angina, MI and abnormal heart rhythm (n=1,123, n=1,015, n=1,141, n=1,050 respectively). 2009 estimates were used for hypertension (n=274, n=244, n=280, n=253 respectively) and diabetes (n=391, n=345, n=398, n=358 respectively). 2010 estimates were used for asthma (n=720, n=608, n=730, n=630 respectively) and COPD (n=720, n=608, n=730, n=631 respectively). Both 2009 and 2010 estimates (n=1,112, n=1,128, n=953, n=989 respectively) were used for CKD.

[d] CVD includes angina, heart attack, stroke, heart murmur and irregular heart rhythm.

[e] IHD includes heart attack or angina. [f] HSE estimates only available to the nearest integer.