Comparison of risk factor associations in UK Biobank against representative, general population based studies with conventional response rates: prospective cohort study and individual participant meta-analysis

G David Batty,[1,2] Catharine R Gale,[3,4] Mika Kivimäki,[1] Ian J Deary,[4] Steven Bell[5,6,7]

For numbered affiliations see end of the article.

Correspondence to: D Batty
david.batty@ucl.ac.uk
ORCID 0000-0003-1822-5753)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2020;368:m131
http://dx.doi.org/10.1136/bmj.m131

Accepted: 16 December 2019

ABSTRACT

OBJECTIVE

To compare established associations between risk factors and mortality in UK Biobank, a study with an exceptionally low rate of response to its baseline survey, against those from representative studies that have conventional response rates.

DESIGN

Prospective cohort study alongside individual participant meta-analysis of other cohort studies.

SETTING

United Kingdom.

PARTICIPANTS

Analytical sample of 499 701 people (response rate 5.5%) in analyses in UK Biobank; pooled data from the Health Surveys for England (HSE) and the Scottish Health Surveys (SHS), including 18 studies and 89 895 people (mean response rate 68%). Both study populations were linked to the same nationwide mortality registries, and the baseline age range was aligned at 40-69 years.

MAIN OUTCOME MEASURE

Death from cardiovascular disease, selected malignancies, and suicide. To quantify the difference between hazard ratios in the two studies, a ratio of the hazard ratios was used with HSE-SHS as the referent.

RESULTS

Risk factor levels and mortality rates were typically more favourable in UK Biobank participants relative to the HSE-SHS consortium. For the associations between risk factors and mortality endpoints, however, close agreement was seen between studies. Based on 14 288 deaths during an average of 7.0 years of follow-up in UK Biobank and 7861 deaths over 10 years of mortality surveillance in HSE-SHS, for cardiovascular disease mortality, for instance, the age and sex adjusted hazard ratio for ever having smoked cigarettes (versus never) was 2.04 (95% confidence interval 1.87 to 2.24) in UK Biobank and 1.99 (1.78 to 2.23) in HSE-SHS, yielding a ratio of hazard ratios close to unity (1.02, 0.88 to 1.19). The overall pattern of agreement between studies was essentially unchanged when results were compared separately by sex and when baseline years and censoring dates were aligned.

CONCLUSION

Despite a very low response rate, risk factor associations in the UK Biobank seem to be generalisable.

WHAT IS ALREADY KNOWN ON THIS TOPIC

The primary objective of UK Biobank is to identify risk factors for chronic diseases and injuries of public health importance

That the baseline response rate was an order of magnitude lower than is conventional has led to debate as to the generalisability of its findings

Relative to studies with higher response rates and national statistics, baseline risk factor profile and mortality rates in UK Biobank are more favourable, but the impact, if any, on risk factor associations is unknown

WHAT THIS STUDY ADDS

This is the first study to directly compare risk factor associations in UK Biobank with nationally representative cohort studies with conventional response rates

Associations of a wide range of risk factors with mortality outcomes showed close agreement between studies

Risk factor associations in UK Biobank seem to be generalisable

Introduction

Well designed prospective cohort studies have considerable utility in identifying genetic and environmental risk factors for an array of somatic and psychiatric disorders. In the many contexts in which randomised controlled trials are not feasible owing to financial, ethical, or logistical constraints, this type of observational study provides the best approximation of causality. Although well phenotyped cohort studies have existed for decades, recent technological advances have led to low cost, high throughput methods to quantify genetic variation. Simultaneously expanding prospects for linkage to medical and non-medical electronic records has allowed construction of studies with the capacity to explore the effect of gene-environment combinations on health endpoints at a previously unheralded scale. Several countries have established such national "biobanks,"[1 2] are in the process of their formulation,[3-5] or are planning such an endeavour.[6]

One such leading resource is UK Biobank, a prospective cohort study comprising around 500 000 middle aged and older people.[7] Unusually in the context of medical research, baseline data have been open access since completion of curation in 2012,[8] and, to date, the study has yielded in excess of 1000 publications.[9] Although UK Biobank is rare in its combination of size and content, it also had an uncommonly low response to its baseline survey: of more than nine million people sent an invitation to participate, only around 6% did so.[10] This achieved response rate was driven by the cost and time saving

decision not to re-contact undecided potential participants.[11] Presumably as a consequence, the project came in under budget and ahead of schedule.

Whereas such an approach is doubtless procedurally efficient, the long held view is that epidemiological studies need to achieve considerably higher response rates if their findings are to be credible.[12 13] Debates about the effect of non-response on estimations of chronic disease determinants in UK Biobank—its primary objective—and the wider necessity for representativeness have followed.[14-23] Despite more favourable baseline risk factor levels and mortality rates in UK Biobank relative to studies achieving a greater response,[24] its principal investigators have consistently maintained that, because the exposures of interest have sufficient variance and the study sample is large, the generalisability of associations between risk factors and health outcomes is assured.[11 25 26] Although findings from cohort studies sampled from highly select groups—Framingham residents and British civil servants,[27 28] among many others[29]—provide indirect support for this assertion, to our knowledge it has yet to be tested empirically.

To examine whether risk factor associations in UK Biobank are generalisable, in analyses of raw data from the study, we compared effect estimates for characteristics known to be linked to major causes of mortality against those from a pooling of data from nationally sampled cohort studies drawn from England and Scotland, all of which had a conventional response to their baseline surveys (mean 68%).[30] With UK Biobank data being deployed across a range of scientific disciplines, we also chose an array of mortality endpoints and exposures. Given the nature of our research question, our focus was not on discovery of risk factors; rather, our aim was to test risk factor-endpoint associations that are well established on the basis of strong observational and/or experimental evidence. We therefore related demographic, social, behavioural, and biomedical risk factors to cardiovascular disease,[31 32] physical stature to cardiovascular disease and cancer,[33-35] and educational attainment to suicide risk.[36-39]

Methods

We used individual level data from both UK Biobank,[7] a prospective cohort study, and a pooling of 18 other prospective cohort studies with identical core protocols: the Health Survey for England (HSE; 15 studies)[40] and the Scottish Health Surveys (SHS; three studies)[41] (hereafter, HSE-SHS). The sampling and procedures of these studies have been well described.[42 43] In brief, baseline data collection in UK Biobank took place between 2006 and 2010 in 22 research assessment centres across the UK, resulting in a sample of 502 655 people aged 40 to 69 years (response rate 5.5%).[7] In HSE and SHS, a total of 193 842 people aged 16-102 years (mean response rate 68%; range 58-93%[30]) participated in home based data collection between 1994 and 2008. For the purposes of this comparison, we restricted HSE-SHS data to the 89 895 people (48 364 women) with a baseline age range that matched UK Biobank. Participants in both studies gave informed consent.

Assessment of baseline characteristics

In both UK Biobank and HSE-SHS, the following data were self-reported using identical or near identical enquiries: diagnosis by a physician of chronic disease (diabetes, hypertension, cardiovascular disease); use of multivitamins, lipid lowering drugs, blood glucose lowering drugs, and antihypertensive drugs; educational attainment; cohabitation status; and cigarette smoking habit. Although physical activity and alcohol intake were collected using somewhat different questions, we were able to derive comparable binary categories (current non-drinker versus the rest; physically inactive versus the rest).

During medical examinations, waist and hip circumference, as well as height and weight, were measured directly using standard protocols. Elevated waist:hip ratio was denoted by values of 0.90 or greater for men and 0.85 or greater for women[44]; obesity was indicated by a body mass index of 30 or above.[45] Forced expiratory volume in one second, a measure of pulmonary function, was quantified using spirometry with the best of three (UK Biobank) or five (HSE-SHS) technically satisfactory exhalations used in our analyses. In UK Biobank, seated systolic and diastolic blood pressure measurements were made twice using the Omron HEM-7015IT digital blood pressure monitor (Omron Healthcare)[20] or, exceptionally, a manual sphygmomanometer (6652 people); we used an average of the two readings. In HSE-SHS, three readings were taken using the Dinamap 8100 automated device,[46] with a mean of the second and third values featuring in our analyses. We defined hypertension according to existing guidelines as systolic/diastolic blood pressure of 140/90 mm Hg or above, self-reported use of antihypertensive drugs, or both.[47] Non-fasting venous blood was drawn in both studies.[48 49] Assaying took place at dedicated central laboratories for C reactive protein, glycated haemoglobin, and total cholesterol and high density lipoprotein cholesterol.[40 48]

Ascertainment of cause specific mortality

Participants in both studies were linked to mortality registries by using the procedures of the UK National Health Service Central Registry.[50] We extracted underlying cause of death, coded according to ICD-10 (international classification of disease, 10th revision), from death certificate data.[50] We generated the following mortality outcomes: cardiovascular disease, all cancers combined, lung cancer, smoking attributable cancers, obesity attributable cancers, and suicide. The ICD codes denoting these causes of death are given in supplemental table 1.

Statistical analyses

We calculated hazard ratios and accompanying 95% confidence intervals by using Cox regression models,[51]

doi: 10.1136/bmj.m131 | BMJ 2020;368:m131 | the bmj

adjusting effect estimates for age and sex. In these survival analyses, we censored individuals according to the date of death or the end of follow-up (14 February 2011 in HSE, 31 December 2009 in SHS, 22 February 2016 for UK Biobank), whichever came first. To quantify the difference between the hazard ratios in each of the two studies, we calculated a ratio of the hazard ratio as we have done in other contexts[50] (HSE-SHS was the referent). We used Stata version 15 for all analyses.

Patient involvement

These analyses are based on existing data of typically healthy populations, and we were not involved in their recruitment. Thus, to our knowledge, no patients were explicitly engaged in designing the present research question or the outcome measures, nor were they involved in developing plans for recruitment, design, or implementation of the study. No patients were asked to advise on interpretation or writing up of results. Results from UK Biobank are routinely disseminated to study participants via the study website and social media outlets.

Results

In table 1 (biomedical factors) and supplemental figure 1 (demographic, social, and behavioural factors plus drug use), we compare the baseline characteristics of participants in UK Biobank against those in the compilation of 18 cohort studies. UK Biobank study members were less likely to have had a sub-university level education, to be living alone or unmarried, to be sedentary, to have existing cardiovascular disease, or to be taking drug treatments for raised blood glucose, although the reverse was seen for lipid lowering and antihypertensive drugs. In analyses restricted to study members not reporting the use of such therapies, we essentially observed no marked difference between studies members for total and high density lipoprotein cholesterol or for glycated haemoglobin. Whereas values for C reactive protein were lower in UK Biobank, both systolic and diastolic blood pressure were somewhat higher. Taken together, UK Biobank participants had a generally more favourable risk factor profile.

In UK Biobank, 14 288 deaths from all causes occurred during an average of 7.0 years of follow-up in 499 701 people who consented to be linked to mortality registers. In the combined HSE-SHS databases, 10 years of mortality surveillance gave rise to 7861 deaths in 89 895 people with these consents. Of the five mortality categories examined in survival analyses, rates of cardiovascular disease, all cancers combined, and tobacco and obesity attributable cancers were markedly lower in UK Biobank, whereas the rate of suicide was higher (supplemental table 2).

In figure 1, for each study, we depict the association of known baseline demographic and behavioural risk factors with cardiovascular disease mortality. The expected direction of association was the same in both studies for the seven characteristics, whereby being male, being of higher age, being physically inactive, not drinking alcohol, not being married or cohabiting, being a current or former smoker, and not having a higher education degree were related to elevated rates of cardiovascular disease mortality. Some modest differences existed in the magnitude of these effects in four of the risk factors examined, such that hazard ratios were typically higher in UK Biobank. When we explored the links between biomedical factors and cardiovascular disease mortality (fig 2), all 10 of the biomarkers featured showed known associations with cardiovascular disease deaths in both studies. Although agreement between studies was again high, some heterogeneity was also apparent in the strength of these effects for higher levels of glycated haemoglobin, existing cardiovascular disease (stronger effects in UK Biobank than in HSE-SHS for both risk factors), and hypertension (the reverse). Taken together, a high degree of concordance existed for cardiovascular disease risk factor associations in UK Biobank and HSE-SHS.

Next, we examined the association of selected baseline factors with some non-cardiovascular disease mortality outcomes, including different presentations of cancer deaths and completed suicides (fig 3). Known risk factor associations were replicated across both studies. The magnitude of the association of cigarette smoking with lung cancer and malignancies

Table 1 | Summary of baseline biomedical characteristics in UK Biobank and Health Survey for England and Scottish Health Surveys (HSE-SHS) cohort studies

Characteristics	UK Biobank	HSE-SHS
No of studies	1	18
No of participants (women)	502 655 (273 472)	89 895 (48 364)
Mean (SD) age, years	56.5 (8.10)	53.5 (8.6)
Mean (SD) FEV_1, L	2.81 (0.80)	2.89 (0.89)
Mean (SD) total cholesterol, mmol/L	5.89 (1.07)	5.95 (1.14)
Median (IQR) high density lipoprotein cholesterol, mmol/L	1.43 (1.20-1.71)	1.40 (1.20-1.70)
Median (IQR) glycated haemoglobin, mmol/mol	35.0 (32.6-37.4)	36.6 (33.3-40.9)
Median (IQR) C reactive protein, mg/L	1.26 (0.63-2.49)	1.50 (0.70-3.10)
Mean (SD) systolic blood pressure, mm Hg	137.7 (19.3)	133.3 (18.4)
Mean (SD) diastolic blood pressure, mm Hg	81.6 (10.6)	76.6 (11.5)

FEV_1=forced expiratory volume in one second; IQR=interquartile range; SD=standard deviation.
Sample size given is for study members with age data only and is lower for all other characteristics. Analyses for total cholesterol and high density lipoprotein exclude participants taking lipid lowering drugs; analyses for glycated haemoglobin exclude people with self-reported diabetes and those taking blood glucose lowering drugs; analyses for C reactive protein exclude people with values >10 mg/L; and analyses for blood pressure exclude people taking antihypertensive drugs.

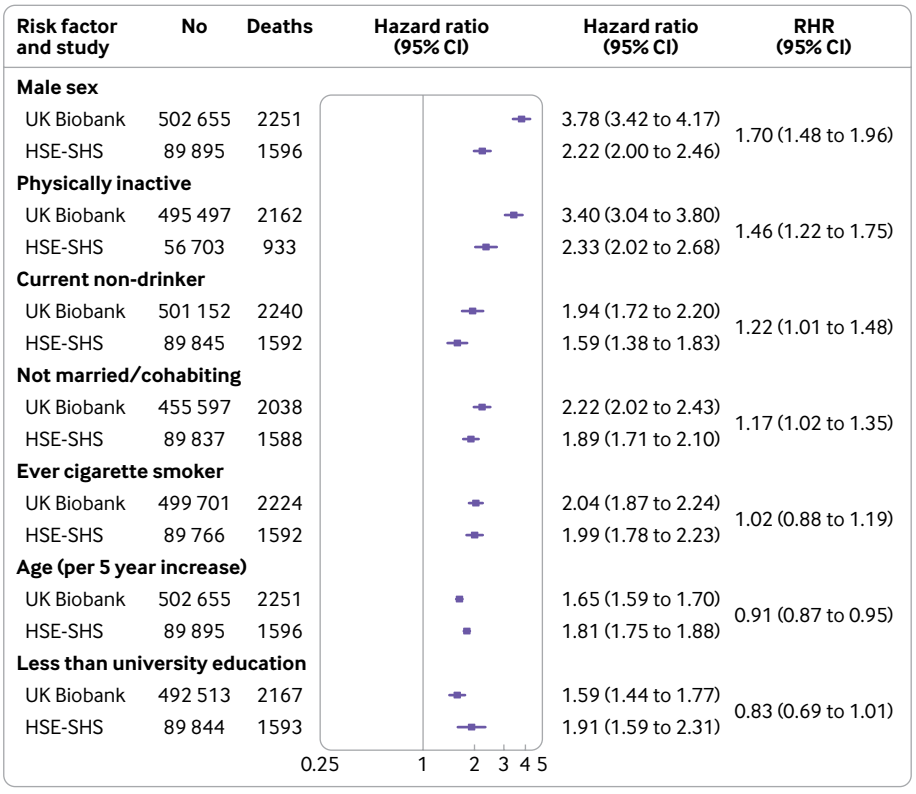

Fig 1 | Association of baseline demographic and behavioural characteristics with cardiovascular disease mortality in UK Biobank and Health Survey for England/Scottish Health Surveys (HSE-SHS) cohort studies. Hazard ratios are adjusted for age and sex, with the exception of individual effects for age and sex which are mutually adjusted. Squares indicate hazard ratios and error bars denote 95% CI for relation of each characteristic with risk of death outcome. Ratio of hazard ratios (RHR) summarises between study differences (HSE-SHS is reference group) for effect estimates for each outcome

causatively linked to tobacco intake were weaker for UK Biobank, whereas obesity and cancers attributed to it yielded similar effects in each study. Hazard ratios were also essentially the same for lower educational attainment and suicide, although statistical power was modest in these analyses, particularly for HSE-SHS, as evidenced by the wide confidence intervals. Physical stature showed the predicted opposing and shallow gradients for cardiovascular disease (negative) and cancer (positive); again, effect sizes were very similar in both studies.

Given well known secular changes in risk factors levels, as evidenced by repeat cross sectional surveys,[52] we used sensitivity analyses to explore the effect of having the same calendar period (2006-08) for recruitment of participants in HSE and UK Biobank (supplemental figure 2), and in another set of analyses we additionally aligned mortality surveillance by right censoring in UK Biobank (follow-up to 14 February 2011) (supplemental figure 3). Owing to a rarity of events, these analyses were restricted to death from cardiovascular disease. Risk factor associations were essentially the same as those apparent in the main analyses, the only exception being obesity. We also found that results held in sex specific analyses for demographic and behavioural characteristics (supplemental figure 4) and biomarkers (supplemental figure 5). Lastly, given that, as described, the self reported use of drugs for lowering blood pressure and lipids was higher in members of UK Biobank relative to our comparator cohorts, we tested whether this was also evident for other health seeking behaviours such as vitamin and mineral supplementation. The prevalence of such use was counter to expectations, being lower in UK Biobank (21.8%) than in HSE-SHS (33.1%).

Discussion

In a comparison of findings between UK Biobank and 18 studies from the HSE-SHS consortium, we found close agreement for a series of well established risk factors for cause specific mortality. These concordant results were apparent despite the response rate in UK Biobank being an order of magnitude lower than in the comparator cohorts and that study having a generally more favourable prevalence of sociodemographic, behavioural, and health related characteristics at baseline and lower rates of cause specific mortality during follow-up, as shown here and elsewhere.[24]

Findings from other studies

The only other analyses of risk factor relations in UK Biobank versus those in comparator studies of which we are aware are those for cardiometabolic

doi: 10.1136/bmj.m131 | *BMJ* 2020;368:m131 | the bmj

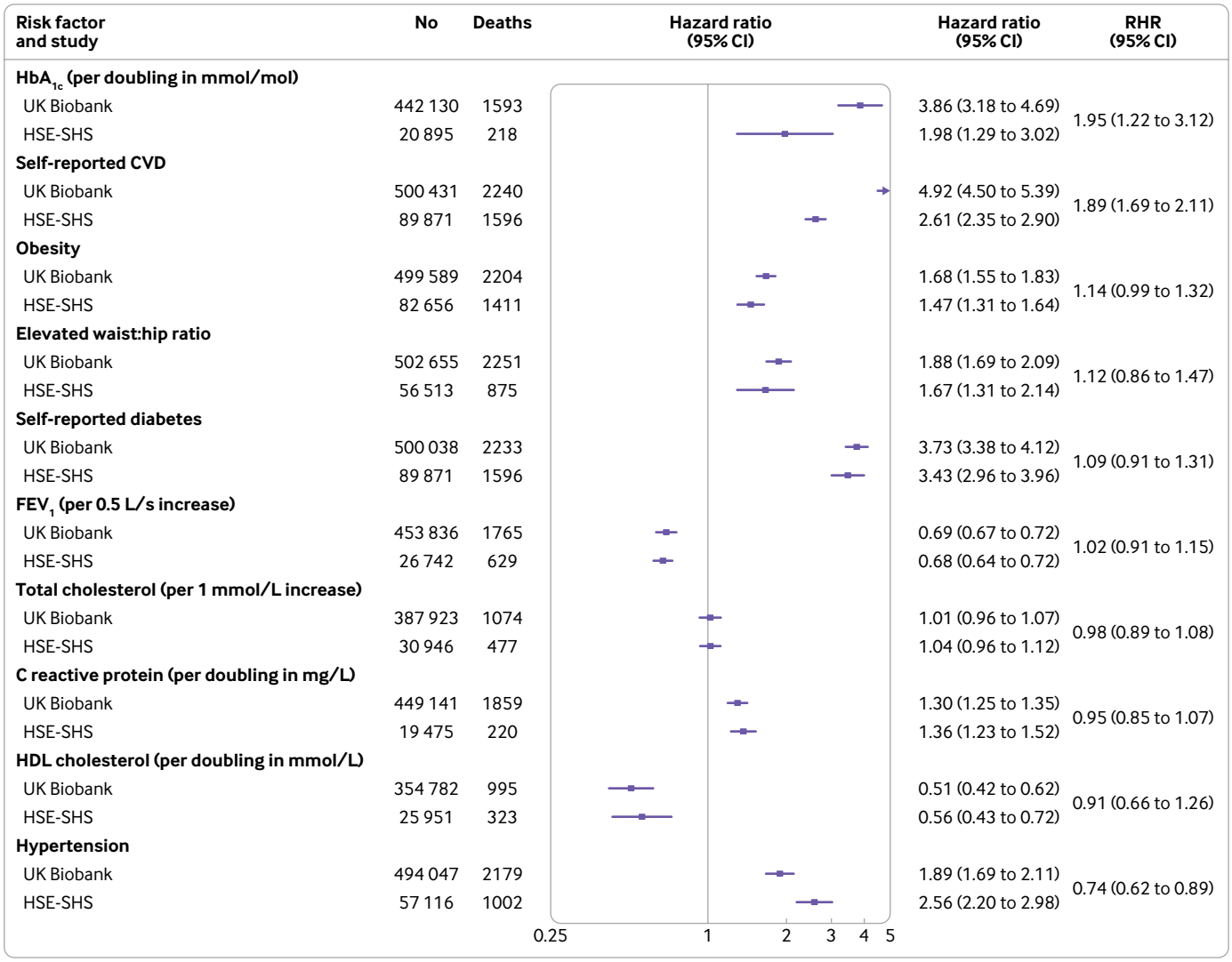

Fig 2 | Association of baseline biomedical characteristics with cardiovascular disease (CVD) mortality in UK Biobank and Health Survey for England/Scottish Health Surveys (HSE-SHS) cohort studies. Hazard ratios are adjusted for age and sex. Squares indicate hazard ratios and error bars denote 95% CI for relation of each characteristic with risk of death outcome. Ratio of hazard ratios (RHR) summarises between study differences (HSE-SHS is reference group) for effect estimates for each outcome. Distributions of glycated haemoglobin (HbA_{1c}), C reactive protein, and high density lipoprotein (HDL) cholesterol were skewed, so they were log_2 transformed and effect estimates reflect doubling for each biomarker. Elevated waist:hip ratio was denoted by ≥0.90 for men and ≥0.85 for women; obesity was indicated by body mass index ≥30. FEV_1=forced expiratory volume in one second

multimorbidity and venous thromboembolism in the Emerging Risk Factors Collaboration, a pooling of data from up to 91 cohort studies.[53 54] The goal of those papers, however, was discovery of risk factors rather than testing well established associations between risk factors and chronic disease. Blood based biomarkers in UK Biobank were also not available at the time of these analyses and, in the report featuring venous thromboembolism as the endpoint of interest,[54] inter-study comparison was hampered by differing approaches to disease ascertainment.

As described, UK Biobank principal investigators, while acknowledging that their study has little value in describing the prevalence of a risk factor or rates of mortality—never stated objectives—have attempted to minimise unease around the investigation of chronic disease aetiology—its primary purpose—by arguing that generalisable associations with risk factors can be obtained in non-representative samples provided sufficiently large numbers of people with a range of exposures are included.[11 25 26] They cite the circumstantial evidence of cohort studies drawing on selected populations that have markedly higher response rates than UK Biobank—Framingham residents,[27] British physicians,[55] US nurses[56]—all of which produced results that have subsequently been shown to be transportable to general population based studies and have contributed much to the prevention of cardiovascular disease and selected cancers. Similarly, our findings mirror those from analyses in which we have compared risk factors for coronary heart disease in another highly select group, a cohort of British

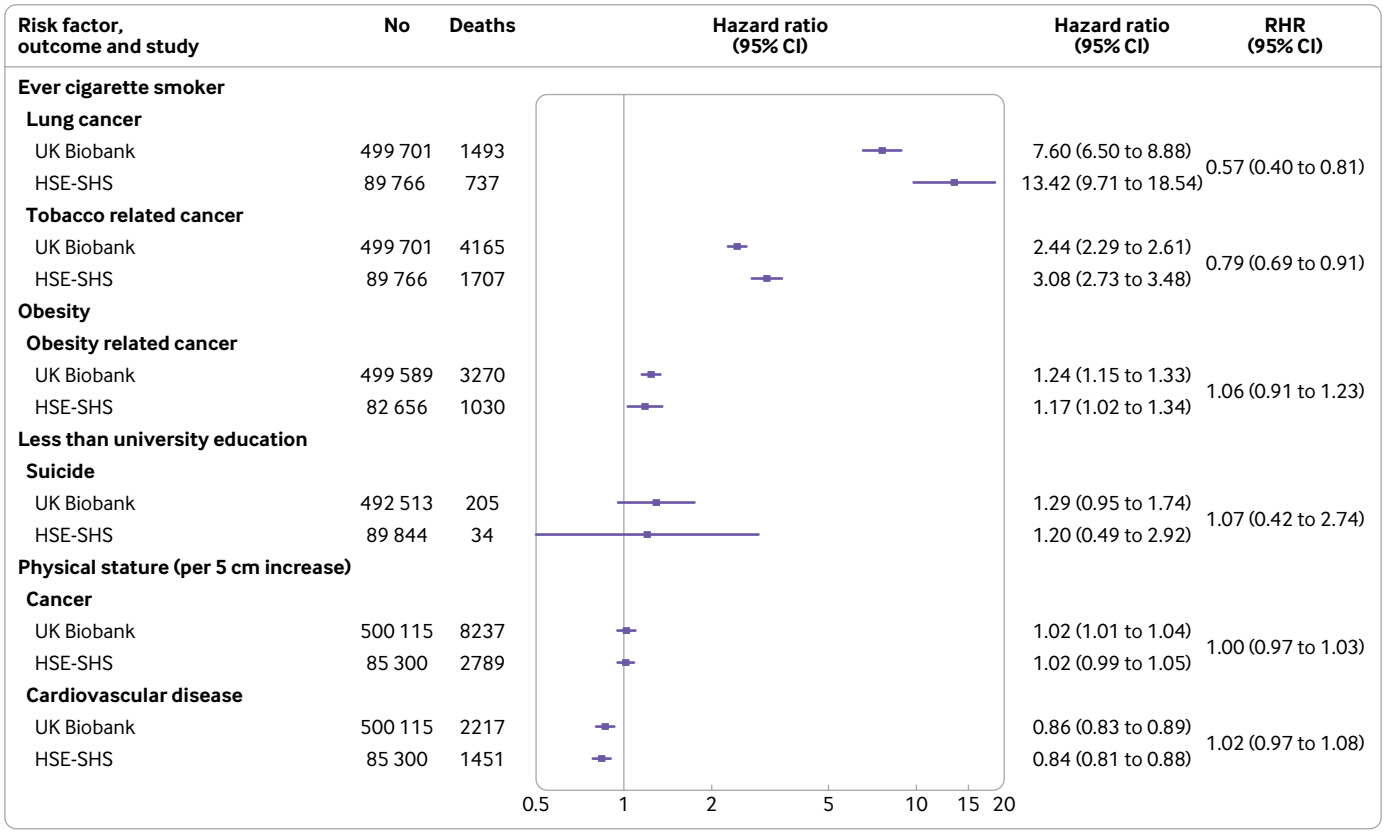

Fig 3 | Association of selected baseline characteristics with cause specific mortality in UK Biobank and Health Survey for England/Scottish Health Surveys (HSE-SHS) cohort studies. Hazard ratios are adjusted for age and sex. Squares indicate hazard ratios and error bars denote 95% CI for relation of each characteristic with risk of death outcome. Ratio of hazard ratios (RHR) summarises between study differences (HSE-SHS is reference group) for effect estimates for each outcome

civil servants (the Whitehall II prospective cohort study), with those from a cohort based on the general population (the British Regional Heart Study).[57] In those analyses, we also found near identical risk factor relations across studies.

Limitations of study

Our work inevitably has some shortcomings. Firstly, whereas UK Biobank includes people from the contiguous countries that comprise the UK, the comparator studies included no data from Wales. We have no reason to believe that the absence of these data would affect our results, however. Secondly, whereas core elements of data collection in the HSE-SHS consortium were essentially constant across studies, scientific themes for data collection differed from year to year.[40] As such, selected biomedical data were not collected in all survey years and the analytical sample size was diminished as a result. Thirdly, for two variables—physical activity and alcohol intake—baseline data were not directly comparable between studies, although we were able to harmonise data into binary groups. These represent two of 23 risk factor-outcome combinations, however, which means that exclusion of these data would have no effect on our overall conclusions of high agreement between studies. Fourthly, the mode of data collection differed between studies—data collection in UK Biobank took place in designated research centres, whereas it was home based in HSE-SHS—although we see no strong justification for this affecting our results. Fifthly, in the main analyses, the endpoint of the interest was cardiovascular disease mortality, which is an amalgam of both incidence of the condition and survival with it. This raises the question of whether risk factor effects differ for incidence, which is temporally closer to assessment of exposure than is death. However, comparison of risk factors for coronary heart disease and stroke, as ascertained from mortality records and hospital admissions (incidence), have shown no evidence of differential associations.[58 59] Lastly, although blood samples have been frozen in HSE-SHS, so offering the potential for later genome sequencing, comparison with genetic risk prediction of chronic disease in UK Biobank is currently not possible. From a purely gene-outcome association perspective, however, with genetic variants being unlikely to be associated with either self-selection into the study or confounding factors, UK Biobank is likely to produce generalisable estimates of genetic risk.[19]

Conclusions

Despite a low response rate, risk factor associations in UK Biobank seem to be generalisable. This suggests

that the cost and time saving features of recruitment of study members did not affect aetiological utility.

AUTHOR AFFILIATIONS

[1]Department of Epidemiology and Public Health, University College London, London WC1E 6BT, UK

[2]School of Biological and Population Health Sciences, Oregon State University, Corvallis, OR, USA

[3]MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton, UK

[4]Lothian Birth Cohorts, Department of Psychology, University of Edinburgh, Edinburgh, UK

[5]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

[6]National Institute for Health Research Blood and Transplant Unit in Donor Health and Genomics at the University of Cambridge, Cambridge, UK

[7]Stroke Research Group, Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK

We thank participants in the studies featured in this manuscript.

Contributions: GDB generated the idea for the study, formulated an analytical plan, and wrote the manuscript. CRG (UK Biobank) and SB (HSE-SHS) formulated an analytical plan and did all the data analyses. SB prepared the figures. All authors commented on an earlier version of the manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. CRG had full access to UK Biobank data, and SB had full access to HSE-SHS data. GDB takes responsibility for the decision to submit the manuscript for publication. GDB, CRG, and SB are the guarantors.

Funding: GDB is supported by the UK Medical Research Council (MR/P023444/1) and the US National Institute on Aging (1R56AG052519-01; 1R01AG052519-01A1). MK is supported by the UK Medical Research Council (MR/R024227), the US National Institute on Aging (NIH) (R01AG056477), NordForsk, and the Academy of Finland (311492). CRG is supported by the UK Medical Research Council (MRC_MC_UU_12011/2 and MRC_MC_UP_A620_1015). SB is supported by the NIHR Blood and Transplant Research Unit in Donor Health and Genomics (NIHR BTRU-2014-10024), UK Medical Research Council (MR/L003120/1), British Heart Foundation (RG/13/13/30194), and NIHR Cambridge Biomedical Research Centre at the Cambridge University Hospitals NHS Foundation Trust. There was no direct financial or material support for the work reported in the manuscript. The funders of the studies had no role in study design, data collection, data analysis, data interpretation, or report preparation.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organisation for the submitted work; the authors have published papers using data from the studies featured in this manuscript (these counts are not mutually exclusive, such that selected publications involve more than one author from the present group and more than one of the datasets: GDB (8 UK Biobank; 38 HSE/SHS), CRG (28; 2), MK (4; 13), IJD (30; 0), and SB (9; 9)); IJD was responsible for the design of some of the cognitive function tests in the revised battery used in the imaging sessions in UK Biobank and is also a study participant; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: In UK Biobank, ethical approval for data collection was received from the North-West Multi-centre Research Ethics Committee and the research was carried out in accordance with the Declaration of Helsinki of the World Medical Association. In HSE-SHS, ethical approval for data collection was granted by the London Research Ethics Council or local research ethics councils. No additional ethical approval was required for the analyses of the data. Participants in both studies gave informed consent.

Data sharing: Data from UK Biobank (https://www.ukbiobank.ac.uk/) and the Health Surveys for England and the Scottish Health Surveys (https://data-archive.ac.uk/) are available to bona fide researchers on application. Part of this research has been conducted using the UK Biobank Resource under Application 10279.

Transparency: GDB affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Dissemination to participants and related patient and public communities: Findings will be disseminated via the media departments of the authors' institutes. Results from UK Biobank are routinely disseminated to study participants via the study website and Twitter feed.

Pre-print deposition: medRxiv (https://www.medrxiv.org/content/10.1101/19004705v1).

1 Nagai A, Hirata M, Kamatani Y, et al, BioBank Japan Cooperative Hospital Group. Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol* 2017;27(3s):S2-8. doi:10.1016/j.je.2016.12.005

2 Chen Z, Lee L, Chen J, et al. Cohort profile: the Kadoorie Study of Chronic Disease in China (KSCDC). *Int J Epidemiol* 2005;34:1243-9. doi:10.1093/ije/dyi174

3 Leitsalu L, Haller T, Esko T, et al. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int J Epidemiol* 2015;44:1137-47. doi:10.1093/ije/dyt268

4 Cronin RM, Jerome RN, Mapes B, et al, Vanderbilt University Medical Center Pilot Team, and the Participant Provided Information Committee. Development of the Initial Surveys for the All of Us Research Program. *Epidemiology* 2019;30:597-608. doi:10.1097/EDE.0000000000001028

5 Gaziano JM, Concato J, Brophy M, et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* 2016;70:214-23. doi:10.1016/j.jclinepi.2015.09.016

6 Al Kuwari H, Al Thani A, Al Marri A, et al. The Qatar Biobank: background and methods. *BMC Public Health* 2015;15:1208. doi:10.1186/s12889-015-2522-7

7 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:e1001779. doi:10.1371/journal.pmed.1001779

8 Conroy M, Sellors J, Effingham M, et al. The advantages of UK Biobank's open-access strategy for health research. *J Intern Med* 2019;286:389-97. doi:10.1111/joim.12955

9 UK Biobank. Published papers. 2019. https://www.ukbiobank.ac.uk/published-papers/.

10 Allen N, Sudlow C, Downey P, et al. UK Biobank: Current status and what it means for epidemiology. *Health Policy Technol* 2012;1:123-6. doi:10.1016/j.hlpt.2012.07.003

11 Manolio TA, Collins R. Enhancing the feasibility of large cohort studies. *JAMA* 2010;304:2290-1. doi:10.1001/jama.2010.1686

12 Evans SJ. Good surveys guide. *BMJ* 1991;302:302-3. doi:10.1136/bmj.302.6772.302

13 Fincham JE. Response rates and responsiveness for surveys, standards, and the Journal. *Am J Pharm Educ* 2008;72:43. doi:10.5688/aj720243

14 Swanson JM. The UK Biobank and selection bias. *Lancet* 2012;380:110. doi:10.1016/S0140-6736(12)61179-9

15 Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. *Int J Epidemiol* 2013;42:1012-4. doi:10.1093/ije/dys223

16 Elwood JM. Commentary: On representativeness. *Int J Epidemiol* 2013;42:1014-5. doi:10.1093/ije/dyt101

17 Nohr EA, Olsen J. Commentary: Epidemiologists have debated representativeness for more than 40 years--has the time come to move on?*Int J Epidemiol* 2013;42:1016-7. doi:10.1093/ije/dyt102

18 Richiardi L, Pizzi C, Pearce N. Commentary: Representativeness is usually not necessary and often should be avoided. *Int J Epidemiol* 2013;42:1018-22. doi:10.1093/ije/dyt103

19 Ebrahim S, Davey Smith G. Commentary: Should we always deliberately be non-representative?*Int J Epidemiol* 2013;42:1022-6. doi:10.1093/ije/dyt105

20 Rothman KJ, Gallacher JE, Hatch EE. Rebuttal: When it comes to scientific inference, sometimes a cigar is just a cigar. *Int J Epidemiol* 2013;42:1026-8. doi:10.1093/ije/dyt124

21 Richiardi L, Pizzi C, Pearce N. Representativeness. *Int J Epidemiol* 2014;43:632-3. doi:10.1093/ije/dyt271

22 Rothman K, Hatch E, Gallacher J. Representativeness is not helpful in studying heterogeneity of effects across subgroups. *Int J Epidemiol* 2014;43:633-4. doi:10.1093/ije/dyt265

23 Keyes KM, Westreich D. UK Biobank, big data, and the consequences of non-representativeness. *Lancet* 2019;393:1297. doi:10.1016/S0140-6736(18)33067-8

24 Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* 2017;186:1026-34. doi:10.1093/aje/kwx246

25 Collins R. What makes UK Biobank special?*Lancet* 2012;379:1173-4. doi:10.1016/S0140-6736(12)60404-8
26 Littlejohns TJ, Sudlow C, Allen NE, Collins R. UK Biobank: opportunities for cardiovascular research. *Eur Heart J* 2019;40:1158-66. doi:10.1093/eurheartj/ehx254
27 Dawber TR, Moore FE, Mann GV. Coronary heart disease in the Framingham study. *Am J Public Health Nations Health* 1957;47:4-24. doi:10.2105/AJPH.47.4_Pt_2.4
28 Kivimäki M, Shipley MJ, Ferrie JE, et al. Best-practice interventions to reduce socioeconomic inequalities of coronary heart disease mortality in UK: a prospective occupational cohort study. *Lancet* 2008;372:1648-54. doi:10.1016/S0140-6736(08)61688-8
29 Blackburn H. The Origins and Early Evolution of Epidemiologic Research in Cardiovascular Diseases: A Tabular Record of Cohort and Case-Control Studies and Preventive Trials Initiated From 1946 to 1976. *Am J Epidemiol* 2019;188:1-8.
30 McCartney G, Russ TC, Walsh D, et al. Explaining the excess mortality in Scotland compared with England: pooling of 18 cohort studies. *J Epidemiol Community Health* 2015;69:20-7. doi:10.1136/jech-2014-204185
31 Yusuf S, Reddy S, Ounpuu S, Anand S. Global burden of cardiovascular diseases: part I: general considerations, the epidemiologic transition, risk factors, and impact of urbanization. *Circulation* 2001;104:2746-53. doi:10.1161/hc4601.099487
32 Yusuf S, Joseph P, Rangarajan S, et al. Modifiable risk factors, cardiovascular disease, and mortality in 155 722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study. *Lancet* 2019;S0140-6736(19)32008-2. doi:10.1016/S0140-6736(19)32008-2
33 Batty GD, Shipley MJ, Gunnell D, et al. Height, wealth, and health: an overview with new data from three longitudinal studies. *Econ Hum Biol* 2009;7:137-52. doi:10.1016/j.ehb.2009.06.004
34 Lee CM, Barzi F, Woodward M, et al, Asia Pacific Cohort Studies Collaboration. Adult height and the risks of cardiovascular disease and major causes of death in the Asia-Pacific region: 21,000 deaths in 510,000 men and women. *Int J Epidemiol* 2009;38:1060-71. doi:10.1093/ije/dyp150
35 Stefan N, Häring HU, Hu FB, Schulze MB. Divergent associations of height with cardiometabolic disease and cancer: epidemiology, pathophysiology, and global implications. *Lancet Diabetes Endocrinol* 2016;4:457-67. doi:10.1016/S2213-8587(15)00474-X
36 Batty GD, Kivimäki M, Bell S, et al. Psychosocial characteristics as potential predictors of suicide in adults: an overview of the evidence with new results from prospective cohort studies. *Transl Psychiatry* 2018;8:22. doi:10.1038/s41398-017-0072-8
37 Li Z, Page A, Martin G, Taylor R. Attributable risk of psychiatric and socio-economic factors for suicide from individual-level, population-based studies: a systematic review. *Soc Sci Med* 2011;72:608-16. doi:10.1016/j.socscimed.2010.11.008
38 Iemmi V, Bantjes J, Coast E, et al. Suicide and poverty in low-income and middle-income countries: a systematic review. *Lancet Psychiatry* 2016;3:774-83. doi:10.1016/S2215-0366(16)30066-9
39 Jee SH, Kivimaki M, Kang HC, Park IS, Samet JM, Batty GD. Cardiovascular disease risk factors in relation to suicide mortality in Asia: prospective cohort study of over one million Korean men and women. *Eur Heart J* 2011;32:2773-80. doi:10.1093/eurheartj/ehr229
40 Mindell J, Biddulph JP, Hirani V, et al. Cohort profile: the health survey for England. *Int J Epidemiol* 2012;41:1585-93. doi:10.1093/ije/dyr199
41 Gray L, Batty GD, Craig P, et al. Cohort profile: the Scottish health surveys cohort: linkage of study participants to routinely collected records for mortality, hospital discharge, cancer and offspring birth characteristics in three nationwide studies. *Int J Epidemiol* 2010;39:345-50. doi:10.1093/ije/dyp155
42 Batty GD, Russ TC, Stamatakis E, Kivimäki M. Psychological distress in relation to site specific cancer mortality: pooling of unpublished data from 16 prospective cohort studies. *BMJ* 2017;356:j108. doi:10.1136/bmj.j108
43 Russ TC, Stamatakis E, Hamer M, Starr JM, Kivimäki M, Batty GD. Association between psychological distress and mortality: individual participant pooled analysis of 10 prospective cohort studies. *BMJ* 2012;345:e4933. doi:10.1136/bmj.e4933
44 World Health Organization. *Waist circumference and waist-hip ratio - Report of a WHO Expert Consultation, Geneva, 8-11 December 2008*. WHO, 2008.
45 World Health Organization. *Physical status: the use and interpretation of anthropometry: report of a WHO expert committee. Who Tech. Rep. Ser.* WHO, 1995.
46 Bolling K. *The Dinamap 8100 Calibration Study.* Her Majesty's Stationery Office, 1994.
47 Chobanian AV, Bakris GL, Black HR, et al, National Heart, Lung, and Blood Institute Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure, National High Blood Pressure Education Program Coordinating Committee. The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure: the JNC 7 report. *JAMA* 2003;289:2560-72. doi:10.1001/jama.289.19.2560
48 Elliott P, Peakman TC, UK Biobank. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int J Epidemiol* 2008;37:234-44. doi:10.1093/ije/dym276
49 Russ TC, Hamer M, Stamatakis E, Starr JM, Batty GD, Kivimäki M. Does the Framingham cardiovascular disease risk score also have predictive utility for dementia death? An individual participant meta-analysis of 11,887 men and women. *Atherosclerosis* 2013;228:256-8. doi:10.1016/j.atherosclerosis.2013.02.020
50 Batty GD, Gale CR, Kivimäki M, Bell S. Assessment of Relative Utility of Underlying vs Contributory Causes of Death. *JAMA Netw Open* 2019;2:e198024. doi:10.1001/jamanetworkopen.2019.8024
51 Cox D. Regression models and life-tables.[Ser B]. *J R Stat Soc [Ser A]* 1972;34:187-220.
52 Huffman MD, Capewell S, Ning H, Shay CM, Ford ES, Lloyd-Jones DM. Cardiovascular health behavior and health factor changes (1988-2008) and projections to 2020: results from the National Health and Nutrition Examination Surveys. *Circulation* 2012;125:2595-602. doi:10.1161/CIRCULATIONAHA.111.070722
53 Di Angelantonio E, Kaptoge S, Wormser D, et al, Emerging Risk Factors Collaboration. Association of Cardiometabolic Multimorbidity With Mortality. *JAMA* 2015;314:52-60. doi:10.1001/jama.2015.7008
54 Gregson J, Kaptoge S, Bolton T, et al, Emerging Risk Factors Collaboration. Cardiovascular Risk Factors Associated With Venous Thromboembolism. *JAMA Cardiol* 2019;4:163-73. doi:10.1001/jamacardio.2018.4537
55 Doll R, Hill AB. The mortality of doctors in relation to their smoking habits: a preliminary report. 1954. *BMJ* 2004;328:1529-33, discussion 1533. doi:10.1136/bmj.328.7455.1529
56 Colditz GA. The nurses' health study: a cohort of US women followed since 1976. *J Am Med Womens Assoc (1972)* 1995;50:40-4.
57 Batty GD, Shipley M, Tabák A, et al. Generalizability of occupational cohort study findings. *Epidemiology* 2014;25:932-3. doi:10.1097/EDE.0000000000000184
58 Batty GD, Kivimäki M, Bell S. Comparison of risk factors for coronary heart disease morbidity versus mortality. *Eur J Prev Cardiol* 2019:2047487319882512. doi:10.1177/2047487319882512
59 Hart CL, Hole DJ, Davey Smith G. Comparison of risk factors for stroke incidence and stroke mortality in 20 years of follow-up in men and women in the Renfrew/Paisley Study in Scotland. *Stroke* 2000;31:1893-6. doi:10.1161/01.STR.31.8.1893

Supplementary materials