OPEN

Is Cohort Representativeness *Passé*? Poststratified Associations of Lifestyle Risk Factors with Mortality in the UK Biobank

Emmanuel Stamatakis,[a] *Katherine B. Owen,*[b] *Leah Shepherd,*[b] *Bradley Drayton,*[b] *Mark Hamer,*[c] *and Adrian E. Bauman*[b]

Background: The UK Biobank (UKB) has been used widely to examine associations between lifestyle risk factors and mortality outcomes. It is unknown whether the extremely low UKB response rate (5.5%) and lack of representativeness materially affects the magnitude and direction of effect estimates.

Methods: We used poststratification to match the UKB sample to the target population in terms of sociodemographic characteristics and prevalence of lifestyle risk factors (physical inactivity, alcohol intake, smoking, and poor diet). We compared unweighted and poststratified associations between each lifestyle risk factor and a lifestyle index score with all-cause, cardiovascular disease (CVD), and cancer mortality. We also calculated the unweighted to poststratified ratio of HR (RHR) and 95% confidence interval as a marker of effect-size difference.

Results: Of 371,974 UKB participants with no missing data, 302,009 had no history of CVD or cancer, corresponding to 3,298,958 person years of follow-up. Protective associations between alcohol use and CVD mortality observed in the unweighted UKB were substantially altered after poststratification, for example, from a hazard ratio (HR) of 0.63 (0.45–0.87) unweighted to 0.99 (0.65–1.50) poststratified for drinking ≥5 times/week versus never drinking. The magnitude of the poststratified all-cause mortality hazard ratio comparing least healthy with healthiest tertile of lifestyle risk factor index was 9% higher (95% confidence interval: 4%, 14%) than the unweighted estimates.

Conclusions: Lack of representativeness may distort the associations of alcohol with CVD mortality, and may underestimate health hazards among those with cumulatively the least healthy lifestyles.

Keywords: Alcohol; Cohort study; Diet; Health behaviors; Mortality; Physical activity; Representativeness; Smoking

(*Epidemiology* 2021;32: 179–188)

Lifestyle risk factors such as physical inactivity, poor diet, and smoking have established associations with chronic disease,[1] premature mortality,[2] and health-related quality of life. Because of the chronic nature of such behavioral exposures and the near absence of long-term randomized controlled trials due to ethical or feasibility hurdles, much of our current knowledge on how they affect health comes from observational studies. For example, almost the entirety of the evidence used to develop guidelines on alcohol drinking[3] and physical activity[4] comes from observational cohort studies with mortality outcomes. Such guidelines are often translated into clinical practice and policy, and are used in clinical trials that involve lifestyle modification.[5]

With rare exceptions, the samples of such observational studies are unrepresentative of the general population.[6] Unrepresentativeness is often rooted in the very low response rates these studies achieve, such as the Australian 45 and Up Study[7] (18% response rate) and the UK Biobank (UKB)[8] (5.5% response rate). The markedly low response rate of the influential UKB resource (>1000 peer-reviewed publications), in particular, has ignited debate on how (lack of) sample representativeness in observational cohorts affects the magnitude, direction, and generalizability of the associations between behavioral exposures and disease or mortality outcomes. Compared with the general UK population, UKB participants

Editor's Note: A related article is found on p. 189.

Submitted April 16, 2020; accepted December 3, 2020

From the [a]Charles Perkins Centre, School of Health Sciences, Faculty of Medicine and Health, University of Sydney, Sydney, NSW, Australia; [b]Charles Perkins Centre, Prevention Research Collaboration, School of Public Health, Faculty of Medicine and Health, University of Sydney, Sydney, NSW, Australia; and [c]Institute Sport Exercise & Health, Division Surgery Interventional Science, Faculty of Medical Sciences, University College London, London, United Kingdom.

E.S. is funded by a National Health and Medicine Research Council (NHMRC, Australia) through a Senior Research Fellowship (grant code: APP1110526), a Leadership 2 Investigator Grant (code: APP1180812).

The authors report no conflicts of interest.

The analytic code is provided in the manuscript's online appendix. The UK Biobank data are accessible upon application.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com). This content is not peer-reviewed or copy-edited; it is the sole responsibility of the authors.

Correspondence: Emmanuel Stamatakis, The Charles Perkins Centre, University of Sydney, Camperdown, NSW, Australia. E-mail: emmanuel.stamatakis@sydney.edu.au.

ISSN: 1044-3983/21/3202-0179

DOI: 10.1097/EDE.0000000000001316

were considerably more likely to be older[6] and less likely to be physically inactive[8] or obese, smoke, or drink alcohol on a daily basis.[6] They had fewer chronic conditions, and markedly lower mortality and cancer incidence rates.[6] The UKB investigators previously stated that "UK Biobank is not representative of the general population on a variety of sociodemographic, physical, lifestyle and health-related characteristics, ….. As a result, UK Biobank is not a suitable resource for deriving generalizable disease prevalence and incidence rates,"[9] but have also informally recalled a previous statement[10] claiming that valid measures of association of lifestyle exposures with disease and mortality outcomes can be safely generalized as they do "not require participants to be representative of the population at large." As both statements have been withdrawn from the UKB website, there is currently no UKB guidance on the role of representativeness and low response rates on interpreting environment (including lifestyle) – disease associations, a topic which has attracted substantial theoretical discussions before[11] and after the launch of the UKB data resource,[10,12] but surprisingly little empirical assessment.[12–15] A recent simulation by Keyes and Westreich[10] in the Lancet illustrated how the "healthy volunteer effect" present in the UKB[6] may grossly distort relative risk estimates of environmental and lifestyle exposures with chronic disease.

Among the very few attempts – to our knowledge – to empirically evaluate the role of unrepresentativeness, a recent study compared lifestyle risk factor and cardiovascular disease (CVD) mortality estimates in the UKB and a pooled series of health surveillance cohorts from 1993 to 2008 in England and Scotland that had high response rates (69% on average).[14] Results were inconsistent as, despite the authors' conclusions that estimates were comparable, the magnitude of the age and sex adjusted estimates of CVD mortality risk for physical inactivity (physically inactive vs. the rest) and alcohol drinking (noncurrent drinker vs. the rest) were markedly larger in the UKB than the pooled cohorts, for example, the HR for physical inactivity was 3.40 (95% confidence interval [CI]: 3.04, 3.80) versus 2.33 (95% CI: 2.02, 2.68). These results offer very limited insights on the influence of poor cohort representativeness on the associations between lifestyle risk factor and mortality risk. Among other reasons, pooling representative datasets from two different countries (e.g., England and Scotland) results in a dataset that is representative of neither country. Age and sex adjusted estimates are rarely (if ever) used in policy and guideline development; multivariable adjusted models are necessary to reduce or eliminate confounding by socioeconomic and other lifestyle risk factors. Cohort unrepresentativeness may also be affected by analytic decisions to exclude study participants with prevalent disease at baseline to minimize the possibility of reverse causation (i.e., spurious associations between abstinence from alcohol or low physical activity levels and mortality risk due to existing illness).

When the target population is well defined, weighting methods can be used to restore the results in a nonrepresentative study sample to reflect the target population.[16] To the best of our knowledge, no study has calculated the multivariable adjusted associations of lifestyle risk factors with mortality in the UKB study (or any other large cohort), after restoring the cohort's socioeconomic, demographic, and health behavior profiles to closely match the target population. The aim of this study was to examine how sample representativeness affects the multivariable adjusted associations of lifestyle risk factors with all-cause and cause-specific (CVD and cancer) mortality.

METHODS

UK Biobank

This research has been conducted using the UKB Resource under Application Number 25813.[8] The UKB is a prospective cohort study including 502,600 participants 40–69 years of age who were recruited in 22 centers across the United Kingdom between 2006 and 2010. This sample was drawn from over 9 million people initially approached (response rate 5.45%) who were registered with the UK NHS, were between 40 and 69 years of age, and lived within 40 km (25 miles) from an assessment center in England, Wales, and Scotland. Full details of the study methods have been published elsewhere.[17] All participants consented to the use of their de-identified data, including access to their health-related records, for research.[17] The UKB has been approved by the National Research Ethics Service (Ref 11/NW/0382).

The Health Survey for England

The Health Survey for England 2008 (HSE)[18] served as the poststratification reference for lifestyle risk factors' prevalence. We used HSE and the already available corresponding nonresponse weights to estimate the total number of people with combinations of lifestyle risk factors for adults aged 40-69 years in the general UK population. We chose 2008 as the approximate mid-point year of the UKB baseline data collection (2006–2010). HSE is a household-based population surveillance study in which a multistage, stratified probability design was used to select households representative of the target populations of England.[19,20] The overall response rate in HSE 2008 was 64%.[21] To obtain a truly representative dataset of the target population in terms of key characteristics (age, sex, household type, geographical region, and social class), the survey team developed nonresponse weights using methods that are described in detail elsewhere.[21] In brief, the HSE team fitted a logistic regression model for all adults in participating households, excluding single-adult households. They calculated adult nonresponse weights as the inverse of the predicted probabilities of response estimated from the regression model, and trimmed the nonresponse weights for adults at the 1% tails to remove extreme values.[21]

The HSE contained records on 7721 participants 40–69 years of age in 2008. We excluded HSE participants who were missing any poststratification variables (smoking n = 21, highest qualification achieved n = 22, body mass index [BMI] n = 1,048, total excluded n = 1,055), leaving the total of 6666 participants to be included in the calculation of poststratification weights.

Outcomes (UKB and HSE)

Date of death was obtained through linkage with national datasets from the National Health Service (NHS) Information Centre (England and Wales) and the NHS Central Register Scotland (Scotland). Participants were followed until April 2020. Primary cause of death was recorded using the International Classification of Diseases 10th revision (ICD10). CVD deaths included codes I01.0–I199. Cancer deaths included codes C00.0–C97.

Lifestyle Risk Factors

The choice and categorization of UKB lifestyle risk factors was determined by the availability of comparable information in HSE 2008[18] data. Alcohol consumption was categorized using the number of days alcohol was consumed per week[3]: (1) never drinker; (2) previous drinkers; (3) current, drinking less than 5 times/week); 4) current, drinking ≥5 times/week). Physical activity was assessed using the short-form International Physical Activity Questionnaire (IPAQ).[22] IPAQ assesses physical activity across leisure time, domestic activities, occupational activity, and transport-related activity.[23] Physical activity was quantified using the metabolic equivalent task (MET) hours of physical activity/week, calculated by multiplying the MET value of activity by the number of hours/week. We then classified participants' physical activity as low (no physical activity), medium (>0, <7.5 MET hours/week), high (roughly equivalent to current public health PA guidelines; ≥7.5 MET hours/week).[8,24] Smoking was grouped as: never smoker, ex-smoker, and current smoker. To classify diet, we calculated average daily fruit and vegetable consumption as the sum of servings of cooked vegetables (1 serving = 2 tablespoons), salad and raw vegetables (1 serving = 2 tablespoons), fresh fruit and dried fruit consumed (1 serving = 1 piece) per day.

Composite Lifestyle Index

Several recent publications from the UKB[25–27] use composite lifestyle risk factor scores instead of a single lifestyle health behavior exposure. To examine the influence of sample representativeness on the associations of overall lifestyle and mortality, we categorized the three lifestyle risk factors (physical inactivity, fruit and vegetable consumption, and smoking) into three groups each as described above. We then applied the following scoring: least healthy = score of 2, medium = 1, most healthy = 0) to derive a composite variable ("lifestyle index") with eight groups. Alcohol was scored as never drinker = 0; previous drinker = 1; current (<5 time and >5 times combined) = 2. We then grouped the resultant score into tertiles: 5 – 8 (least healthy lifestyle), 4 (middle tertile), 0–3 (healthiest lifestyle),

Poststratification

We used poststratification to weight underrepresented and overrepresented groups in the UKB to that of the HSE data, which is generalizable to the English population.[28] We chose a source from the largest UK constituent country (England, 84% of total UK population) because there are no nationally representative UK-wide data on lifestyle risk factors.

The HSE data were divided into mutually exclusive groups, according to the six-way cross tabulation of age group (40–49, 50–59, and 60–70 years); sex: (male and female); highest qualification (college or university degree, high school diploma, other/none), smoking (ever, never smoker), physical activity (≥7.5 MET hours/week, <7.5 MET hours/week), and BMI (not overweight, overweight, or obese). We then calculated the sampling weights for each cell such that the weighted totals from the UKB sum to the totals in the UK population. We also considered alcohol and fruit and vegetable consumption for inclusion; however, it was not possible due to lower cell counts. The variable groupings listed earlier were selected to preserve adequate unweighted frequencies in the mutually exclusive cells in both the UKB and the HSE.

Statistical Analyses

We present unweighted and poststratified UKB estimates and compared the latter with actual Census 2011 UK (age and sex) and HSE (lifestyle risk factors) data.

Consistent with current practice, we excluded those with a history of CVD and cancer at baseline from the main multivariable analyses. We used Cox proportional hazard regression models to estimate hazard ratios (HRs) for the association between lifestyle risk factors/unhealthy index and all-cause mortality both before and after poststratification. We measured survival using age as the time scale, from age of assessment to age at death or censoring date.[29] We mutually adjusted models adjusted for lifestyle risk factors and additionally adjusted for age, gender, and highest qualification. We used the ratio of the HRs (RHRs), calculated as the poststratified HR divided by the unweighted HR) ($\mathrm{RHR} = \frac{\mathrm{HR}_{ps}}{\mathrm{HR}_{unweighted}}$) to quantify the relative change in estimates following poststratification. Percentile CIs for the RHR were estimated using bootstrapping with 1000 iterations.[30]

We ran a set of sensitivity analyses to establish the robustness of the results. We repeated the analysis including those with history of major CVD (coronary heart disease or stroke) or cancer and adjusting for history of CVD and history of cancer. We did an additional series of sensitivity analyses to address the high proportion of missing data for poststratification variables. These data were assumed to be missing at random, where the probability of missing variables depends on the observed values of other variables, rather than the missing values.[31] We imputed missing data using the multiple imputation by chained equations approach to create five datasets.[32] We used logistic regression for binary variables and polytomous regression for categorical variables and included all covariates in the imputation models. Cox proportional hazard regression models were performed on all five datasets and estimates were combined using Rubin's rules.[31]

All analyses were performed using R software version 3.6.1 (R Foundation for Statistical Computing, Vienna, Austria).[33] All analytic code can be found in eAppendix 1; http://links.lww.com/EDE/B762.

RESULTS

Sample

From the full UKB sample of 502,600, we excluded participants who were missing any poststratification variables (missing: smoking n = 2,951, physical activity n = 121,167, highest qualification achieved n = 10,141, BMI n = 10,138, total excluded = 130,626), leaving a total of 371,974 participants to be considered, corresponding to 3,298,958 person years of follow-up. For the main analyses, we excluded participants with a history of major cardiovascular events (n = 56,345) or who had been diagnosed with cancer (n = 18,782) before baseline, leaving a total of n = 302,009 individuals with no CVD or cancer at baseline (n = 5162 participants had history of both CVD and cancer). Of the 302,009 individuals in the main analyses, there were 11,875 all-cause deaths (5.82%), of which we classified 3,580 as CVD and 7,217 as cancer deaths.

Fit of the Poststratified Dataset to the Target Population

eTable 1; http://links.lww.com/EDE/B763 presents the unweighted and poststratified distribution of key characteristics in the UKB Study. We noted considerable corrections in the distribution of the poststratified sample for age (e.g., percentage of participants in the 40- to 49-years age group increased from 24.9% in the unweighted to 38.0% in the poststratified), educational qualification (e.g., from 47.9 to 32.4% college/university educated), and physical activity (from 87.2% to 69.2% meeting the recommendations). Table 1 shows the key characteristics of the UKB Study after excluding those with a history of CVD or cancer (n after exclusion = 302,009), and eTable 2; http://links.lww.com/EDE/B763 shows the characteristics of the sample imputed exposures and covariates (n = 400,793). eTable 3; http://links.lww.com/EDE/B763 compares the distributions of the poststratified UKB and the HSE 2008 samples key characteristics. With the exception of alcohol intake, where some relatively modest differences were present, and fruit and vegetable consumption, the poststratification in the UKB achieved identical distributions across sociodemographic and lifestyle risk factors. For both men and women, poststratification normalized the age distribution toward to the actual UK population, especially in the groups of 40–44, 45–49, 60–64, and 65–69 years where the UKB sample was markedly unrepresentative (eFigure 1; http://links.lww.com/EDE/B763). We also calculated mortality rates per 1000 person–years of follow-up for the unweighted and poststratified UKB samples. These are compared to the UK annual mid-year mortality rates over the period of the UKB follow-up in eTable 4; http://links.lww.com/EDE/B763. Poststratified mortality rates were consistently higher than the unweighted rates, and closer to actual UK mortality rates.

Individual Lifestyle Risk Factors

All HRs presented in tables, figures, or text are multivariable adjusted. Figures 1–3 present the unweighted and poststratified HRs for each lifestyle risk factor against all-cause, CVD, and cancer mortality. Table 2 shows the unweighted to poststratified RHR and corresponding 95% CIs for each lifestyle risk factor. With the exception of the CVD mortality alcohol use estimates and all-cause mortality smoking estimates, the unweighted and poststratified estimates for the remaining lifestyle risk factors were similar across the three mortality outcomes. Specifically, the protective associations between increased alcohol use and CVD mortality in the unweighted dataset were not present following poststratification, for example, from an HR of 0.63 (0.45–0.87) unweighted to 0.99 (0.65–1.50) poststratified for drinking ≥5 times/week versus never drinking. These alcohol findings were also reflected by the RHR (poststratified to unweighted): 1.52, 95% CI: 1.23, 1.86 for drinking <5 times/week; and 1.55, 1.23–1.86 for drinking ≥5 times/week) (Figure 2; Table 2). The poststratified all-cause mortality risk for current smokers was higher than the unweighted estimates (ratio of HRs 1.13, 95% CI: 1.06, 1.20) (Figure 2; Table 2). The pattern of the cancer mortality estimates comparisons was broadly similar to all-cause mortality with less evidence for differences between unweighted and poststratified estimates (Figure 3; Table 2). Including participants who had a history of major CVD or cancer at baseline affected minimally the poststratified versus unweighted comparisons of lifestyle risk factor estimates across all three mortality outcomes (eFigure 2; http://links.lww.com/EDE/B763 shows the CVD mortality results as an example).

Lifestyle Index

No consistent differences existed between the unweighted and poststratified all-cause mortality HRs in the lower and middle range values of the lifestyle index scores (eFigure 3; http://links.lww.com/EDE/B763). Figure 4A corroborates this pattern as the unweighted and poststratified estimates for all-cause mortality were very similar in the middle lifestyle index tertile but the magnitude of the poststratified estimates was higher in the least healthy lifestyle index tertile (ratio of HRs in the top lifestyle risk factor index tertile: 1.09, 95% CI: 1.04, 1.14, Table 3). The unweighted and poststratified estimates for CVD and cancer mortality were similar across tertiles of the lifestyle index (Figure 4B and C; Table 3. None of these findings were affected materially by the inclusion of participants with history of CVD or cancer (see eFigure 4; http://links.lww.com/EDE/B763, as an example).

Imputation

eFigures 5–7 (http://links.lww.com/EDE/B763) present the unweighted and poststratified HRs for each lifestyle

TABLE 1. Frequencies of Participant Demographic and Health-related Variables in the Health Survey for England and UK Biobank, Excluding People in the UK Biobank with History of Cancer or Cardiovascular Disease (n = 302,009 After Exclusion)

Variable		UK Biobank N Unweighted n	UK Biobank % Unweighted	UK Biobank % Poststratified
Age group (years)	40–49	84,158	28	42
	50–59	105,311	35	32
	60–70	112,540	37	26
Sex	Female	159,827	53	51
	Male	142,182	47	49
Education qualification	College or university degree	147,786	49	34
	Highschool diploma	118,080	39	41
	Other/None	36,143	12	25
Physical activity	≥7.5 MET hours/week	264,601	88	71
	>0, <7.5 MET hours/week	33,703	11	26
	No physical activity	3,705	1	3
Fruit and vegetable consumption	At least 10 portions/day	25,655	8	7
	5–9 portions/day	133,012	44	39
	under 5 portions/day	140,396	46	52
	Unknown	2,946	1	1
Alcohol use frequency	Never	10,687	4	4
	Previous	8,895	3	3
	Current: < almost daily	216,405	72	73
	Current: ≥ almost daily	65,891	22	19
	Unknown	131	<1	<1
Smoking status	Never	170,417	56	52
	Previous	101,227	34	34
	Current	30,365	11	14
Lifestyle index	Lowest tertile 1 most healthy	100,834	33	26
	Middle tertile	111,473	37	32
	Highest tertile: least healthy	86,637	29	40
BMI category (kg/m^2)	Underweight (<18.5)	1,450	<1	<1
	Normal (18.5–24.9)	104,945	35	30
	Overweight (25.0–29.9)	130,888	43	44
	Obese I (30.0–34.9)	48,066	16	18
	Obese II (35.0–39.9)	12,355	4	5
	Obese III ≥ 40.0	4,305	1	2

BMI indicates body mass index; MET, metabolic equivalent task.

risk factor against all-cause, CVD, and cancer mortality in the imputed dataset (n = 400,793). There were no appreciable differences with the unimputed data presented in Figures 1–3.

DISCUSSION

We used poststratification based on a nationally representative sample to restore the UKB's socioeconomic and behavioral profiles to the target population; and we compared the associations of lifestyle risk factors and mortality in the original and poststratified UKB samples. We found that the associations of physical activity, smoking, and diet with all-cause, CVD, and cancer mortality were broadly similar in the two sets of analyses. Poststratification eliminated the protective associations between alcohol use and CVD mortality we observed in the original UKB analyses. We also found that the all-cause mortality risk of current smokers and those with cumulatively the least healthy lifestyles may be underestimated due to poor sample representativeness. Nevertheless, the absolute difference in these estimates was 13% (smoking) and 9% (least healthy lifestyle index score) respectively, and the practical importance of such risk underestimation is likely to be small.

Our results suggest that the protective associations of alcohol intake with CVD outcomes observed in previous studies[8,34] may be spurious. Although we only used weekly frequency of alcohol intake in the current analyses, recent UKB results[8] suggested that alcohol volume also shows protective associations, for example, drinking within guidelines was associated with lower risk for CVD mortality compared to never drinkers (HR: 0.73, 95% CI: 0.56, 0.95) while drinking even more than double the recommended amounts was not associated with elevated CVD risk (HR: 0.86, 95% CI: 0.63, 1.17). The link of low response rates to spurious cardioprotective effect estimates of alcohol intake[34] is further

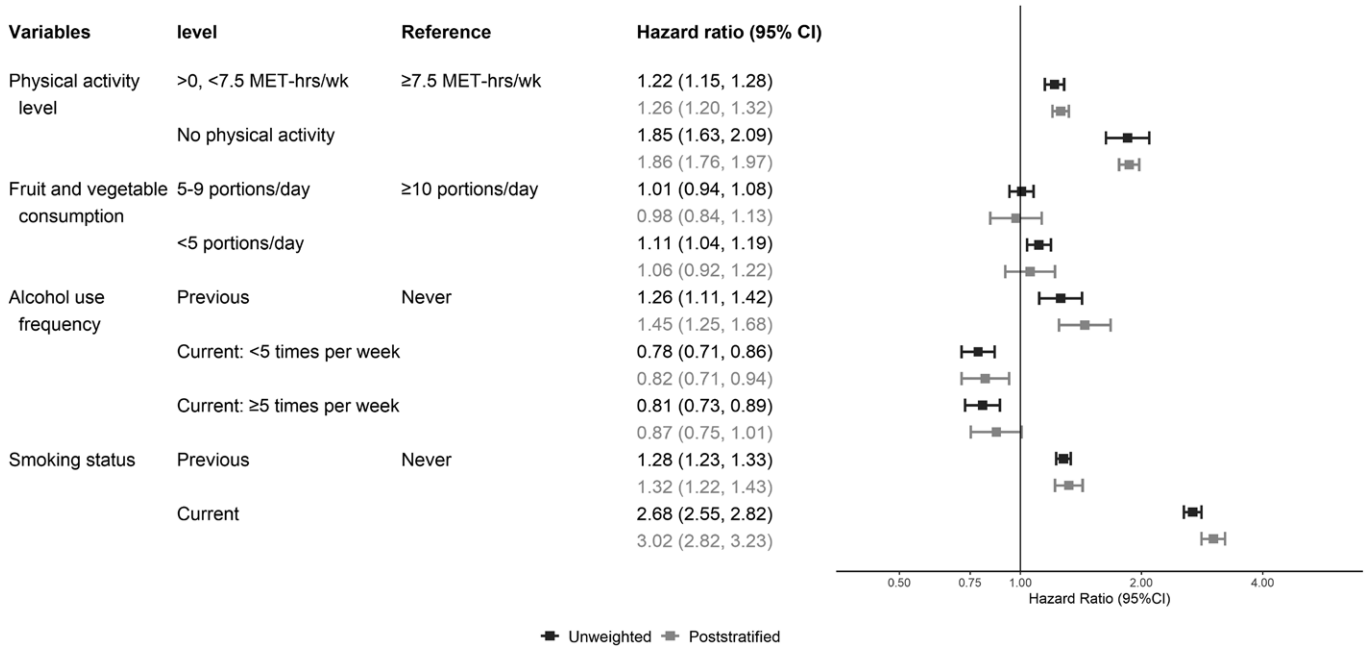

FIGURE 1. Adjusted (model mutually adjusted for all lifestyle risk factors, and additionally adjusted for age, sex, and highest educational qualification) hazard ratio of each lifestyle risk factor for all-cause mortality, excluding people with history of cancer or cardiovascular disease (n = 302,009 after exclusions). MET indicates metabolic equivalents.

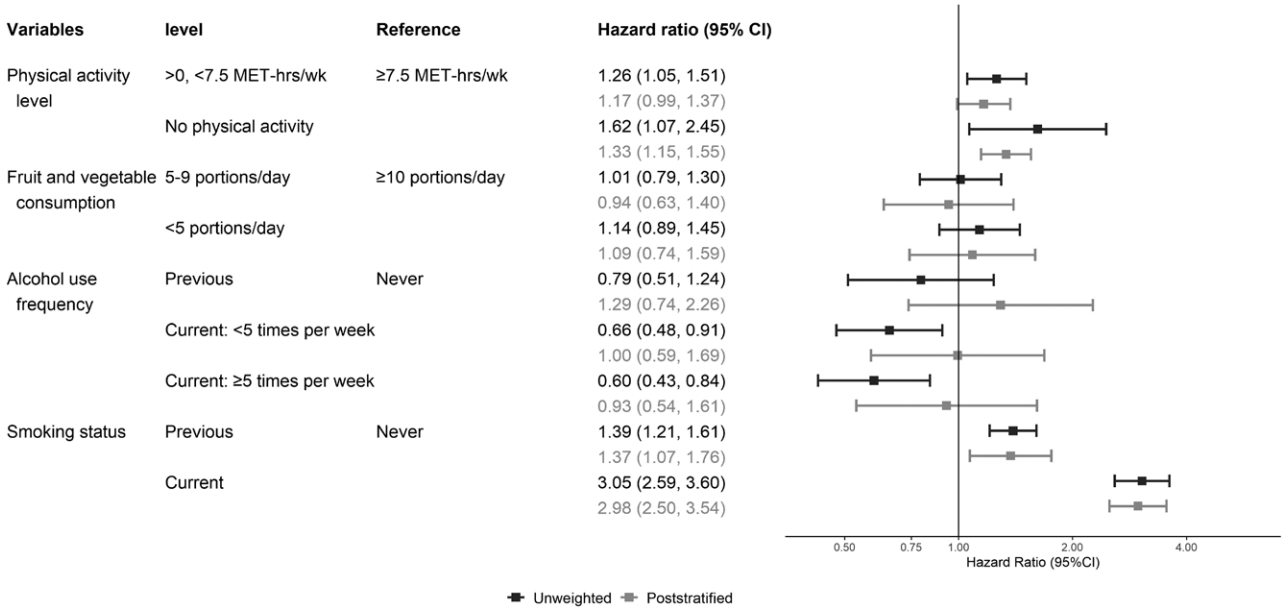

FIGURE 2. Adjusted (model mutually adjusted for all lifestyle risk factors, and additionally adjusted for age, sex, and highest educational qualification) hazard ratio of each lifestyle risk factor for all cardiovascular disease (CVD) mortality, excluding people with history of cancer or CVD (n = 302,009 after exclusion). MET indicates metabolic equivalents.

supported by the absence of such findings in studies involving nationally representative cohorts. For example, an analysis of eight pooled British (England and Scotland) cohorts with high response rates (68%–77%) found no association between moderate drinking volume and CVD mortality.[24] It is not possible to directly compare our alcohol use results to the recent study with similar aims to ours that compared the UKB to the pooled 1993–2008 dataset from England and Scotland.[14] Being noncurrent drinker (vs. the rest) showed a 22% (1%–48%) higher risk in the latter dataset compared to the Biobank. However,

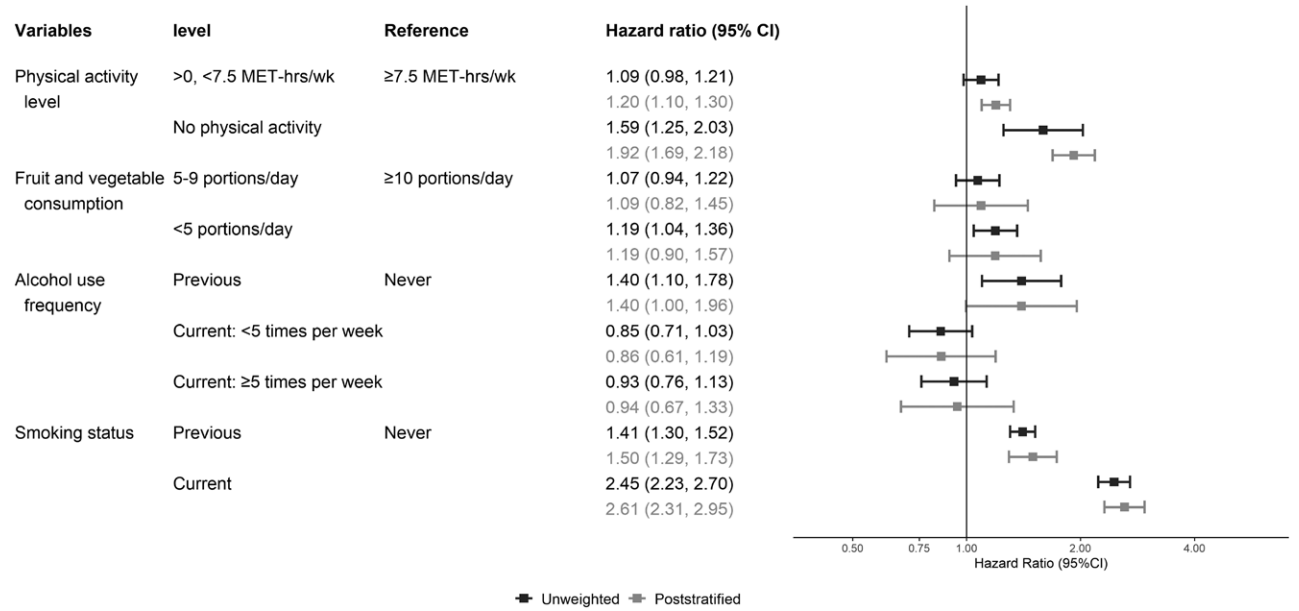

FIGURE 3. Adjusted (model mutually adjusted for all lifestyle risk factors, and additionally adjusted for age, sex, and highest educational qualification) hazard ratio of each lifestyle risk factor for cancer mortality, excluding people with history of cancer or CVD (n = 302,009 after exclusions). CVD indicates cardiovascular disease; MET, metabolic equivalents.

TABLE 2. Adjusted RHR[a] of Each Lifestyle Risk Factor for All-cause, CVD, and Cancer Mortality, Excluding People with History of Cancer or CVD (n = 302,009 After Exclusion)

		RHRs (Poststratified/Unweighted)		
Variable	**Level**	**All-cause Mortality**	**CVD Mortality**	**Cancer Mortality**
Physical activity level	≥7.5 MET hours/week	Reference	Reference	Reference
	>0, <7.5 MET hours/week	1.04 (0.99, 1.08)	0.92 (0.82, 1.03)	1.06 (0.99, 1.19)
	No PA	1.01 (0.90, 1.13)	0.83 (0.57, 1.13)	1.21 (0.98, 1.45)
Fruit and vegetable consumption	At least 10 portions/day	Reference	Reference	Reference
	5–9 portions/day	0.97 (0.89, 1.07)	0.93 (0.75, 1.19)	1.02 (0.86, 1.20)
	Under 5 portions/day	0.95 (0.87, 1.04)	0.96 (0.77, 1.22)	1.00 (0.84, 1.18)
Alcohol use frequency	Never	Reference	Reference	Reference
	Previous	1.15 (0.98, 1.35)	1.63 (1.07, 2.34)	1.00 (0.73, 1.36)
	Current: < almost daily	1.04 (0.92, 1.18)	1.52 (1.23, 1.86)	1.00 (0.79, 1.29)
	Current: ≥ almost daily	1.08 (0.95, 1.24)	1.55 (1.22, 1.99)	1.02 (0.79, 1.33)
Smoking status	Never	Reference	Reference	Reference
	Previous	1.03 (0.98, 1.08)	0.99 (0.87, 1.14)	1.06 (0.98, 1.16)
	Current	1.13 (1.06, 1.20)	0.98 (0.82, 1.14)	1.06 (0.94, 1.20)

[a]To be equivalent to Health Survey for England weighed estimates.
CVD indicates cardiovascular disease; MET, metabolic equivalent task; RHR, ratio of hazard ratio.

the "noncurrent drinker" group pooled lifetime abstainers and exdrinkers who might have quit due to health reasons and as such it offers little information on the risks of alcohol drinking. In the same study, physical inactivity (which was not clearly defined in the article) estimates were 46% (22%–75%) higher in the UKB than the pooled 1993–2008 cohorts. This contrasts our results where the poststratified and unweighted estimates for physical inactivity were closely aligned across all three mortality outcomes.

Compared to the sparse previous literature,[14] our study has a number of notable strengths. We calculated multivariable adjusted estimates that are usually used in policy and guideline development. We assessed differences between unweighted and poststratified estimates in the UKB using data handling methods commonly employed in the field of lifestyle risk factors, such as exclusion of participants with history of major chronic disease at baseline. Our HSE 2008 reference dataset was temporally consistent to the UKB

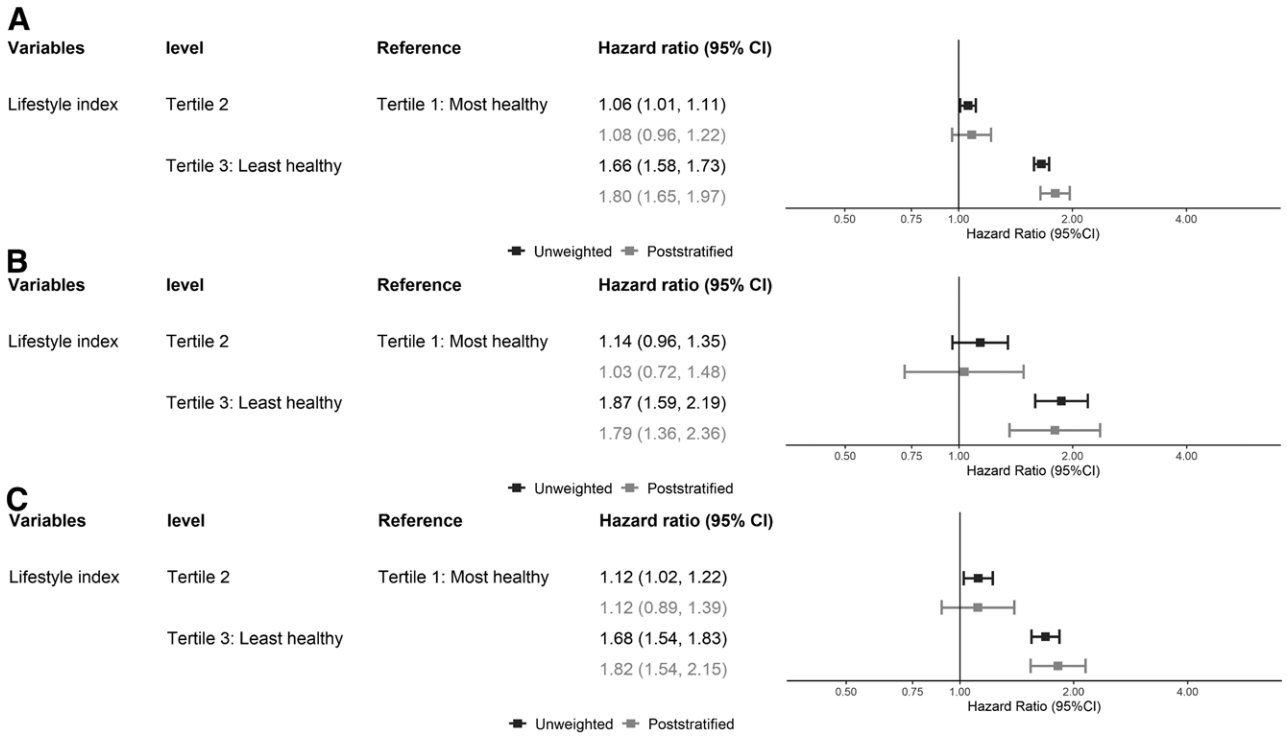

FIGURE 4. Adjusted (adjusted for age, sex, and highest educational qualification) hazard ratio of lifestyle index tertiles for all-cause mortality (A)/CVD mortality (B)/cancer mortality (C), excluding people with history of cancer or CVD (n = 302,009 after exclusions). CVD indicates cardiovascular disease.

TABLE 3. Adjusted RHRs[a] of Lifestyle Index for All-cause, CVD, and Cancer Mortality, Excluding People with History of Cancer or CVD (n = 302,009)

		RHRs (Poststratified/Unweighted)		
Variable	**Level**	**All-cause Mortality**	**CVD Mortality**	**Cancer Mortality**
Lifestyle index	Lowest tertile: most healthy	Reference	Reference	Reference
	Middle tertile	1.02 (0.98, 1.07)	0.91 (0.78, 1.05)	1.00 (0.90, 1.09)
	Highest tertile: least healthy	1.09 (1.04, 1.14)	0.96 (0.81, 1.13)	1.08 (0.97, 1.19)
	Missing	1.08 (0.90, 1.29)	0.81 (0.52, 1.29)	1.05 (0.67, 1.54)

[a]To be equivalent to Health Survey for England weighed estimates.
CVD indicates cardiovascular disease; RHR, ratio of hazard ratio.

baseline (2006–2010); and was weighed for nonresponse to give a reference that is truly representative of the population of adults living in the largest constituent country of the United Kingdom. Another unique strength of our study is that poststratification allowed us to correct the UKB's distribution not only for socioeconomic and demographic variables, but also for health behavior profiles. This is an innovative and more policy relevant method than previous approaches. Our study is relevant to both individual lifestyle risk factors and composite lifestyle risk factors indices that are used increasingly in large-scale observational research.[25–27]

Our study has some limitations. In the absence of a nationally representative UK-wide data resource we used an English reference. This is unlikely to have had a major impact on our results as England corresponds to 84% of the total UK population, although we acknowledge that there are some differences in lifestyle risk factor prevalence between UK contries.[35] As in previous UKB physical activity publications,[36] we were not able to make use of the entire dataset due to missing data on lifestyle risk factors and covariates. However, our sensitivity analyses where we imputed all such data show that our comparisons, conclusions, and underlying study principle were unlikely to be affected by missing data. We cannot eliminate entirely the possibility that the chosen categories of lifestyle risk factors, necessitated by the requirements of poststratification weights development, are not sufficient to correct for complex selection biases. Equally, the variables we used in the development of the poststratification

weight may not have captured differences between HSE and UKB participants in terms of unmeasured factors such as genetic and psychosocial characteristics. Our approach assumes that measurement errors of lifestyle risk factors in HSE and the UKB are comparable between studies. We did not seek to employ best-practice casual inference modeling strategies, such as providing a sound structural framework to identify confounders and mediators based on directed acyclic graphs, or taking into account competing risks in the cause-specific mortality analyses. These omissions were intentional because our primary goal was to replicate common practices in the published UKB (and analogous) literature of lifestyle risk factors and mortality. For example, such literature[25,37–39] does not typically take into account competing risks, while the use of a predefined structural framework for choice of confounders and evaluation of selection biases is still the exception rather than the rule. Accordingly, we acknowledge that both the unweighted and poststratified estimates we reported may be biased and/or imprecise.

In conclusion, lack of cohort representativeness in the UKB may lead to spurious cardioprotective effects estimates for alcohol intake and may underestimate health hazards among those with the least healthy lifestyles. Although physical inactivity, smoking, and dietary estimates appeared to be minimally affected, our findings suggest that the extremely low response rates in cohort studies may distort policy relevant research findings on the health effects of specific exposures. Our results suggest that future UKB (and analogous cohort) users should exercise caution when examining associations between established risk factors and mortality outcomes as poor cohort sample representativeness might influence materially some estimates. Further studies empirically evaluating the influence of unrepresentativeness across other categories of risk factors estimates are warranted.

REFERENCES

1. Lacombe J, Armstrong MEG, Wright FL, Foster C. The impact of physical activity and an additional behavioural risk factor on cardiovascular disease, cancer and all-cause mortality: a systematic review. *BMC Public Health.* 2019;19:900.
2. Ding D, Rogers K, van der Ploeg H, Stamatakis E, Bauman AE. Traditional and emerging lifestyle risk behaviors and all-cause mortality in middle-aged and older adults: evidence from a large population-based Australian cohort. *PLoS Med.* 2015;12:e1001917.
3. Department of Health. *Alcohol Guidelines Review – Report from the Guidelines Development Group to the UK Chief Medical Officers*. Department of Health London; 2016.
4. UK Chief Medical Officers. *UK Chief Medical Officers' Physical Activity Guidelines*. Crown Copyright; 2019.
5. Look AHEAD Research Group. Cardiovascular effects of intensive lifestyle intervention in type 2 diabetes. *N Engl J Med*. 2013;369:145–154.
6. 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;186:1026–1034.
7. Stamatakis E, Gale J, Bauman A, Ekelund U, Hamer M, Ding D. Sitting time, physical activity, and risk of mortality in adults. *J Am Coll Cardiol.* 2019;73:2062–2072.
8. Powell L, Feng Y, Duncan MJ, Hamer M, Stamatakis E. Does a physically active lifestyle attenuate the association between alcohol consumption and mortality risk? Findings from the UK biobank. *Prev Med.* 2020;130:105901.
9. UK Biobank. Researchers. Available at: http://www.ukbiobank.ac.uk/scientists-3. Accessed 10 September 2020.
10. Keyes KM, Westreich D. UK Biobank, big data, and the consequences of non-representativeness. *Lancet.* 2019;393:1297.
11. Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. *Int J Epidemiol.* 2013;42:1012–1014.
12. Ebrahim S, Davey Smith G. Commentary: should we always deliberately be non-representative? *Int J Epidemiol.* 2013;42:1022–1026.
13. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: a potential outcomes perspective. *Epidemiology.* 2017;28:553–561.
14. Batty GD, Gale CR, Kivimäki M, Deary IJ, Bell S. 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. *BMJ.* 2020;368:m131.
15. Mealing NM, Banks E, Jorm LR, Steel DG, Clements MS, Rogers KD. Investigation of relative risk estimates from studies of the same population with contrasting response rates and designs. *BMC Med Res Methodol.* 2010;10:26.
16. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. *Am J Epidemiol.* 2017;186:1010–1014.
17. 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.
18. Scholes S, Coombs N, Pedisic Z, et al. Age- and sex-specific criterion validity of the health survey for England physical activity and sedentary behavior assessment questionnaire as compared with accelerometry. *Am J Epidemiol.* 2014;179:1493–1502.
19. Mindell J, Biddulph JP, Hirani V, et al. Cohort profile: the health survey for England. *Int J Epidemiol.* 2012;41:1585–1593.
20. 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;39:345–350.
21. NHS Information Centre. *The Health Survey for Enland 2008. Vol 2, Methods and* Documentation. NHS Information Centre; 2009.
22. Craig CL, Marshall AL, Sjöström M, et al. International physical activity questionnaire: 12-country reliability and validity. *Med Sci Sports Exerc.* 2003;35:1381–1395.
23. IPAQ Research Committee. Guidelines for data processing and analysis of the International Physical Activity Questionnaire (IPAQ)-short and long forms. 2005. Available at: https://sites.google.com/site/theipaq/. Accessed 15 January 2021.
24. Perreault K, Bauman A, Johnson N, Britton A, Rangul V, Stamatakis E. Does physical activity moderate the association between alcohol drinking and all-cause, cancer and cardiovascular diseases mortality? A pooled analysis of eight British population cohorts. *Br J Sports Med.* 2017;51:651–657.
25. Said MA, Verweij N, van der Harst P. Associations of combined genetic and lifestyle risks with incident cardiovascular disease and diabetes in the UK Biobank Study. *JAMA Cardiol.* 2018;3:693–702.
26. Lourida I, Hannon E, Littlejohns TJ, et al. Association of lifestyle and genetic risk with incidence of dementia. *JAMA*. 2019;322:430–437.
27. Rutten-Jacobs LC, Larsson SC, Malik R, et al; MEGASTROKE consortium; International Stroke Genetics Consortium. Genetic risk, incident stroke, and the benefits of adhering to a healthy lifestyle: cohort study of 306 473 UK biobank participants. *BMJ.* 2018;363:k4168.
28. Levy PL, Lemeshow S. *Sampling of Populations: Methods and Applications*. John Wiley & Sons; 2008.
29. Thiébaut AC, Bénichou J. Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Stat Med.* 2004;23:3803–3820.
30. Davison A, Hinkley DV. *Bootstrap Methods and Their Applications*. Cambridge University Press; 1997.
31. Little RJ, Rubin DB. *Statistical Analysis With Missing Data (Vol. 793)*. John Wiley & Sons; 2019.
32. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res.* 2011;20:40–49.

33. R Foundation for Statistical Computing. R: A language and environment for statistical computing. 2019. Available at: https://www.R-project.org/. Accessed 15 January 2021.
34. Zhao J, Stockwell T, Roemer A, Naimi T, Chikritzhs T. Alcohol consumption and mortality from coronary heart disease: an updated meta-analysis of cohort studies. *J Stud Alcohol Drugs.* 2017;78: 375–386.
35. British Heart Foundation Health Promotion Research Group. Coronary heart disease statistics, 2010 edition. In: Foundation BH, ed. *Coronary Heart Disease Statistics*. British Heart Foundation; 2010.
36. Kunzmann AT, Mallon KP, Hunter RF, et al. Physical activity, sedentary behaviour and risk of oesophago-gastric cancer: a prospective cohort study within UK Biobank. *United European Gastroenterol J.* 2018;6:1144–1154.
37. Pearce M, Strain T, Kim Y, et al. Estimating physical activity from self-reported behaviours in large-scale population studies using network harmonisation: findings from UK Biobank and associations with disease outcomes. *Int J Behav Nutr Phys Act.* 2020;17:40.
38. Schutte R, Papageorgiou M, Najlah M, et al. Drink types unmask the health risks associated with alcohol intake - prospective evidence from the general population. *Clin Nutr.* 2020;39:3168–3174.
39. Celis-Morales CA, Lyall DM, Welsh P, et al. Association between active commuting and incident cardiovascular disease, cancer, and mortality: prospective cohort study. *BMJ.* 2017;357:j1456.