# Is cohort representativeness *passé*? Matching the UK Biobank sample to target population characteristics and recalculating the associations between behavioural lifestyle risk factors and mortality

## ABSTRACT

Background: The UK Biobank (UKB) has been used widely to examine the associations between lifestyle risk factors (LRF) and mortality outcomes. It is unknown if the extremely low UKB response rate (5·5%) and lack of representativeness materially affects the magnitude and direction of effects.

Methods: We used post-stratification to match the UKB sample to the target population in terms of sociodemographic characteristics and prevalence of LRFs (physical inactivity, alcohol intake, smoking, fruit and vegetable intake, obesity). We compared unweighted and post-stratified associations between each LRF and tertiles of a lifestyle index score with all-cause, CVD, and cancer mortality. We calculated the unweighted/post-stratified ratio of hazard ratios and 95%CI (RHR) as a marker of effect size difference.

Results: Out of 371,974 UKB participants with no missing data, 302,009 had no history of CVD or cancer, corresponding to 2,345,142 person years of follow-up. The protective associations between alcohol use and CVD mortality observed in the unweighted UKB were substantially altered after post-stratification as indicated by the higher point estimates and the reduced overlap of the 95% confidence intervals, e.g. from an HR of 0·63 (0·45 to 0·87) unweighted to 0·99 (0·65 to 1·50) post-stratified for drinking ≥5 times/week compared to never drinker. The magnitude of the post-stratified all-cause mortality HR comparing the least healthy with the healthiest tertiles of the lifestyle risk factor index was 9% higher (95%CI: 2% to 18%) than the unweighted estimates.

Conclusions: Lack of representativeness may distort the associations of alcohol with CVD mortality; and may under-estimate health hazards among those with cumulatively the least healthy lifestyles.

Keywords: cohort study, representativeness, health behaviours, alcohol, physical activity, diet, smoking, mortality

Wordcount: 3968

Abstract:  259

## INTRODUCTION

Lifestyle risk factors such as physical inactivity, poor diet, and smoking have established links with chronic disease[1], premature mortality[2], and health related quality of life. Because of the chronic nature of such behavioural exposures and the near absence of long term randomised 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[8] (5·5% response rate). The markedly low response rate of the influential UK Biobank resource (>1000 peer-reviewed publications), in particular, has ignited debate on how (lack of) sample representativeness in observational cohorts affects the magnitude, direction, and generalisability of the associations between behavioural exposures and disease or mortality outcomes. Compared with the general UK population, UK Biobank participants 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 UK Biobank investigators state 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 informally recalled a previous statement[10] claiming that valid measures of association of lifestyle exposures with disease and mortality outcomes can be safely generalised as they do "not require participants to be representative of the population at large". No UK Biobank materials currently offer 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 prior to[11] and after the launch of the UK Biobank data resource ,[10 12] but surprisingly little empirical testing [12-15] A recent simulation by Keyes and Westreich[10] in the Lancet illustrated how the "healthy volunteer effect" present in the UK Biobank[6] may grossly distort relative risk estimates of environmental and lifestyle exposures with chronic disease.

Among the very few attempts to empirically test the role of unrepresentativeness, a recent study compared lifestyle risk factor - CVD mortality estimates in the UK Biobank and a pooled series of health surveillance cohorts from 1993 – 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 (non-current drinker vs the rest) were markedly larger in the UK Biobank than the pooled cohorts, e.g. the HR for physical inactivity was 3·40 (95%CI: 3·04 to 3·80) vs 2·33 (95%CI: 2·02 to 2·68). These results offer very limited insights on the influence of poor sample 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 non-representativeness is often exacerbated by analytic decisions to exclude study participants with prevalent disease at baseline to minimise 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 non-representative study sample to reflect the target population.[16] To the best of our knowledge, no study has calculated the multivariable adjusted lifestyle risk factor associations with mortality in the UK Biobank study (or any other large cohort), after restoring the cohort's socioeconomic, demographic, and health behaviour 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 UK Biobank Resource under Application Number 25813.[8] The UK Biobank is a prospective cohort study including 502 600 participants aged 40-69 years who were recruited in 22 centres across the UK 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's National Health Service, were aged between 40–69 years and lived within 40 km (25 miles ) from an assessment centre 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 Health Survey for England

The Health Survey for England 2008 (HSE)[18] served as the post-stratification reference for lifestyle health behaviour prevalence. We used HSE and the already available corresponding non-response 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 UK Biobank baseline data collection (2006-10). The Health Survey for England 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, social class), the survey team developed non-response weights using methods that are described in detail elsewhere.[21] In brief, a logistic regression model was fitted for all adults in participating households, excluding single-adult households. The adult non-response weights were calculated as the inverse of the predicted probabilities of response estimated from the regression model. The non-response weights for adults were trimmed at the 1% tails to remove extreme values.[21]

The HSE contained records on 7721 participants aged 40-69 in 2008. We excluded HSE participants who were missing any post stratification variables (smoking n=21, highest qualification achieved n= 22, BMI n=1048, total excluded n = 1055), leaving total of 6,666 participants to be included in the calculation of post-stratification 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 to I199· Cancer deaths included codes C00·0 to C97.

*Lifestyle risk factors*

The choice and categorisation of UK Biobank lifestyle risk factors was determined by the availability of comparable information in HSE 2008[18] data. Alcohol consumption was categorised using number of days alcohol was consumed per week [22]: 1) never drinker; 2) previous-drinkers, 3) current, drinking less than 5 times/week); 4) current, drinking ≥5 times/week). PA was assessed using the short-form International Physical Activity Questionnaire (IPAQ).[23] IPAQ assesses PA across leisure time, domestic activities, occupational activity and transport-related activity.[24] Physical activity was quantified using the Metabolic Equivalent Task (MET)-hours of PA/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-hrs/week), high (roughly equivalent to current public health PA guidelines; ≥7·5 MET-hours/week).[8,25] 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 serve = 2 tablespoons), salad and raw vegetables (1 serve = 2 tablespoons), fresh fruit and dried fruit consumed (1 serve = 1 piece) per day..

*Composite lifestyle index*

Several recent publications from the UK Biobank[26,27,28] use composite lifestyle risk factor scores instead of a specific lifestyle health behaviour exposures. To examine the influence of sample representativeness on the associations of overall lifestyle and mortality, we categorised the three lifestyle risk factors (physical inactivity, fruit and vegetable consumption, 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.[29] Alcohol was scored as never drinker = 0; previous drinker = 1; current (< 5 time and > 5 times combined) = 2 . The resultant score was then grouped into tertiles: 5 – 8 (least healthy lifestyle), 4 (middle tertile), 0 – 3 (healthiest lifestyle),

*Post stratification*

Post stratification was used to weight underrepresented and overrepresented groups in the UK Biobank to that of the HSE data, which is generalisable to the English population.[30] We chose a source from the largest UK constituent country because there are no nationally representative UK-wide data on lifestyle health behaviours.

The HSE data was divided into mutually exclusive groups, according to the six-way cross tabulation of age group ( 40 – 49, 50 – 59, 60 – 70 years); sex: (male, female); highest qualification (college or university degree, high school diploma, other/none), smoking (ever, never smoker), physical activity (>=7·5 MET-hrs/week, <7·5 MET-hrs/week) and BMI (not overweight, overweight or obese). The sampling weights were then calculated for each cell such that the weighted totals from the UK Biobank sum to the totals in the UK population. Alcohol and fruit and vegetable consumption were also considered 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 UK Biobank and the HSE.

*Statistical analyses*

We present unweighted and post stratified UK Biobank estimates and we compared the latter with actual Census 2011 UK (age, 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 (HR) for the association between lifestyle risk factors/unhealthy index and all-cause mortality both before and after post stratification. Survival was measured using age as the time scale, from age of assessment to age at death or censoring date.[31] Models were mutually adjusted for lifestyle risk factors and additionally adjusted for age, gender, BMI, and highest qualification. The ratio of the hazard ratios (RHR), calculated as the post-stratified hazard ratio divided by the unweighted hazard ratio ($RHR = \frac{HR_{ps}}{HR_{unweighted}}$) was used to quantify the relative change in estimates following post stratification. Percentile confidence intervals for the ratio of the hazard ratio were estimated using bootstrapping with 1000 iterations.[32] We defined the effect size magnitude difference (unweighted vs post-stratified) as statistically significant when the 95% confidence intervals of RHR did not cross unity.

We ran a set of sensitivity analyses to establish the robustness of the results. The analysis was repeated 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 a  additional series of sensitivity analyses to address the high proportion of missing data for post stratification 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[33]. We imputed missing data  using the multiple imputation by chained equations approach to create 5 datasets[34]. Logistic regression was used for binary variables and polytomous regression for categorical variables and all covariates were included in the imputation models. Cox proportional hazard regression models were performed on all five datasets and estimates were combined using Rubin's rules[33].

All statistical tests were 2 sided. For all analyses, *P* value <.05 was considered statistically significant. All analyses were performed using R software version 3·6·1 (R Foundation for Statistical Computing, Vienna, Austria)[35].   All analytic code can be found in eAppendix1.

## RESULTS

### *Sample*

From the full UK Biobank sample of 502 600, we excluded participants who were missing any post stratification variables (missing Smoking n= 2951, physical activity n = 121 167, highest qualification achieved n = 10 141, BMI n = 10 138, total excluded=130 626), leaving total of 371 974 participants to be considered. 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) prior to 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).

### *Fit of the post-stratified dataset to the target population*

eTable 1 presents the unweighted and post-stratified distribution of key characteristics in the UK Biobank Study. We noted considerable corrections in the distribution of the post-stratified sample for age (e.g. percentage of participants in the 40-49 years age group increased from 24·9% in the unweighted to 38·0% in the post-stratified), 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 UK Biobank Study after excluding those with a history of CVD or cancer (n = 302 009) and eTable 2 shows the characteristics of the sample imputed exposures and covariates (n=400 793) . eTable 3 compares the distributions of the post-stratified UK Biobank 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 post-stratification in the UK Biobank achieved identical distributions across sociodemographic and lifestyle risk factors. For both men and women, post-stratification normalised the age distribution towards to the actual UK population, especially in the groups of 40-44, 45-49, 60-64, and 65-69 years where the UK Biobank sample was markedly un-representative (eFigure 1). We also calculated mortality rates per 1000 person-years of follow up for the unweighted and post-stratified UK Biobank samples. These are compared to the UK annual mid-year mortality rates over the period of the UK Biobank follow up in eTable4. Post-stratified mortality rates were consistently higher than the unweighted rates, converging towards the actual UK mortality rates.

*Individual lifestyle risk factors*

All hazard ratios presented in Tables, Figures, or text, are multivariable adjusted. Figures 1-3 present the unweighted and post-stratified hazard ratios for each lifestyle risk factor against all-cause, CVD, and cancer mortality. Table 2 shows the unweighted/post-stratified ratio of HR and corresponding 95% confidence intervals of the ratio for each lifestyle risk factor. With the exception of CVD mortality alcohol use estimates and all-cause mortality smoking estimates, the unweighted and post-stratified 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 post-stratification, a finding also reflected by the ratio of HRs (post-stratified/unweighted ): 1·52, 95%CI: 1·23 to 1·86 for drinking < 5 times/week; and 1·55, 1·23 to 1·86 for drinking ≥5 times/week)(Figure 2 and Table 2). The post-stratified all-cause mortality risk for current smokers was higher than the unweighted estimates (ratio of HRs 1·13, 95%CI: 1·06 to 1·20 (Figure 2 and 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 post-stratified estimates (Figures 3 & Table 2).Including participants who had a history of major CVD or cancer at baseline affected minimally the post-stratified vs unweighted comparisons of lifestyle risk factor estimates across all three mortality outcomes (eFigure 2 shows the CVD mortality results as an example).).

*Lifestyle Index*

No consistent differences existed between the unweighted and post-stratified all-cause mortality hazard ratios in the lower and middle range values of the lifestyle index scores (eFigure 3). Figure 4 corroborates this pattern as the unweighted and post-stratified estimates for all-cause mortality were very similar in the middle lifestyle index tertile but the magnitude of the post-stratified 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 to 1·14, Table 3). The unweighted and post-stratified estimates for CVD and cancer mortality were similar across tertiles of the lifestyle index (Figures 5- 6 & Table 3. None of these findings were affected materially by the inclusion of participants with history of CVD or cancer (see eFigure 4, as an example).

## DISCUSSION

We used post-stratification based on a nationally representative sample to restore the UK Biobank's socioeconomic and behavioural profiles to the target population; and we compared the associations of lifestyle risk factors and mortality in the original and post-stratified UK Biobank 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. Post-stratification eliminated the protective associations between alcohol use and CVD mortality that were observed in the original UK

Biobank analyses. We also found that the all-cause mortality risk of current smokers and those with cumulatively the least healthy lifestyles may be under-estimated due to poor sample representativeness. Nevertheless, the absolute difference in these estimates was  13% (smoking) and 9% (least healthy lifestyle score) respectively, and  the practical importance of such risk underestimation is likely small.

Our results suggest that the protective associations of alcohol intake with CVD outcomes observed in previous studies[36,37] may be spurious.  Although we only used weekly frequency of alcohol intake in the current analyses, recent UK Biobank results[36] suggested that alcohol volume also shows protective associations, e.g. drinking within guidelines was associated with lower risk for CVD mortality compared to never drinkers (HR: 0·73, 95%CI: 0·56 to 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 to 1·17). The link of low response rates and spurious cardioprotective effects of alcohol intake[37] is further supported by the absence of such effects in studies involving nationally representative cohorts. For example, an analysis of 8 pooled British (England-Scotland) cohorts with high response rates (68%-77%)[25]  found no association between moderate drinking volume and CVD mortality.[25]  It is not possible to directly compare our alcohol use results to the recent study with similar aims to ours that compared the UK Biobank to the pooled 1993 – 2008 dataset from England and Scotland[14]. Being non-current drinker (vs the rest) showed a 22% (1%-48%) higher risk in the latter dataset compared to the Biobank. However, the “non-current drinker” group pooled lifetime abstainers and ex-drinkers 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 manuscript) estimates were 46% (22%-75%) higher in the UK Biobank than the pooled 1993 – 2008 cohorts. This contrasts our results where the post-stratified 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 tested differences between unweighted and post-stratified estimates in the UK Biobank 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 UK Biobank baseline (2006-2010); and was weighed for non-response to give a reference that is truly representative of the population of adults living in the largest constituent country of the UK. Another unique strength of our study is that post-stratification allowed us to correct the UK Biobank's distribution not only for socioeconomic and demographic variables, but also for health behaviour 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.[26,27,28]

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 some between-UK countries differences in lifestyle risk factors prevalence exist[38]. As in previous UK Biobank physical activity publications[39], we were not able to make use of the entire dataset due to missing data on lifestyle risk factors data 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 post-stratification weights development, are not sufficient to correct for complex selection biases. Equally, the variables we used in the development of the post-stratification weight may not have captured differences between HSE and UK Biobank participants in terms of unmeasured factors such as genetics and psychosocial characteristics. Our approach assumes that measurement errors of lifestyle risk factors in HSE and the UK Biobank are comparable between studies.

In conclusion, lack of cohort representativeness in the UK Biobank may lead to spurious cardio-protective effects for alcohol intake; may under-estimate 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 UK Biobank (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 testing 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**(1):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 Medicine* 2015;**12**(12):e1001917.
3. Department of Health. Alcohol Guidelines Review – Report from the Guidelines development group to the UK Chief Medical Officers. In: Health Do, ed. London, 2016.
4. UK Chief Medical Officers. UK Chief Medical Officers' Physical Activity Guidelines. 2019.
5. Look AHEAD Research Group. Cardiovascular Effects of Intensive Lifestyle Intervention in Type 2 Diabetes. *New England Journal of Medicine* 2013;**369**(2):145-154.
6. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, Collins R, Allen NE. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am J Epidemiol* 2017;**186**(9):1026-1034.
7. **Stamatakis E**, Gale E, Bauman A, Ekelund U, Hamer M, Ding D. Sitting Time, Physical Activity, and Risk of Mortality in Adults. *Journal of the American College of Cardiology* 2019;**73**(16):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* 2019;**130**:105901.
9. UK Biobank. Researchers. http://www.ukbiobank.ac.uk/scientists-3.
10. Keyes KM, Westreich D. UK Biobank, big data, and the consequences of non-representativeness. *Lancet* 2019;**393**(10178):1297.

11. Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. *Int J Epidemiol* 2013;**42**(4):1012-4.
12. Ebrahim S, Davey Smith G. Commentary: Should we always deliberately be non-representative? *Int J Epidemiol* 2013;**42**(4):1022-6.
13. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing Study Results: A Potential Outcomes Perspective. *Epidemiology* 2017;**28**(4):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 (Clinical research ed.)* 2020;**368**:m131-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**(8):1010-1014.
17. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* 2015;**12**(3):e1001779.
18. Scholes S, Coombs N, Pedisic Z, Mindell JS, Bauman A, Rowlands AV, Stamatakis E. 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**(12):1493-502.
19. Mindell J, Biddulph JP, Hirani V, Stamatakis E, Craig R, Nunn S, Shelton N. Cohort profile: the health survey for England. *Int J Epidemiol* 2012;**41**(6):1585-93.
20. Gray L, Batty GD, Craig P, Stewart C, Whyte B, Finlayson A, Leyland AH. 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**(2):345-50.
21. NHS Information Centre. The Health Survey for Enland 2008. Vol 2, Methods and Documentaiton. London 2009.
22. Department of Health. Alcohol Guidelines Review–Report from the Guidelines development group to the UK Chief Medical Officers.  Department of Health London, 2016.
23. Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, Pratt M, Ekelund U, Yngve A, Sallis JF. International physical activity questionnaire: 12-country reliability and validity. *Medicine & science in sports & exercise* 2003;**35**(8):1381-1395.
24. IPAQ Research Committee. Guidelines for data processing and analysis of the International Physical Activity Questionnaire (IPAQ)-short and long forms. http://www.ipaq.ki.se/scoring.pdf 2005.
25. 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. *British Journal of Sports Medicine* 2017;**51**(8):651-657.
26. 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 cardiology* 2018;**3**(8):693-702.
27. Lourida I, Hannon E, Littlejohns TJ, Langa KM, Hyppönen E, Kuźma E, Llewellyn DJ. Association of Lifestyle and Genetic Risk With Incidence of Dementia. *JAMA* 2019;**322**(5):430-437.
28. Rutten-Jacobs LC, Larsson SC, Malik R, Rannikmäe K, Sudlow CL, Dichgans M, Markus HS, Traylor M. 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.

29. Stamatakis E, Gale J, Bauman A, Ekelund U, Hamer M, Ding D. Sitting time, physical activity, and risk of mortality in adults. *Journal of the American College of Cardiology* 2019;**73**(16):2062-2072.
30. Levy PL, S. . *Sampling of populations: methods and applications* John Wiley & Sons, 2008.
31. Thiébaut ACM, Bénichou J. Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Statistics in Medicine* 2004;**23**(24):3803-3820.
32. Davison A, Hinkley, DV. *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press, 1997.
33. Little RJ, Rubin DB. *Statistical analysis with missing data (Vol. 793)* John Wiley & Sons, 2019.
34. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research* 2011;**20**(1):40-49.
35. R Foundation for Statistical Computing. R: A language and environment for statistical computing. https://www.R-project.org/, 2019.
36. 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. *Preventive medicine* 2020;**130**:105901-105901.
37. 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. *Journal of studies on alcohol and drugs* 2017;**78**(3):375-386.
38. British Heart Foundation Health Promotion Research Group. Coronary Heart Disease Statistics, 2010 Edition In: Foundation BH, ed. *Coronary Heart Disease Statistics*. Oxford, 2010.
39. Kunzmann AT, Mallon KP, Hunter RF, Cardwell CR, McMenamin ÚC, Spence AD, Coleman HG. Physical activity, sedentary behaviour and risk of oesophago-gastric cancer: A prospective cohort study within UK Biobank. *United European Gastroenterology Journal* 2018;**6**(8):1144-1154.

**Table 1: Frequencies of patient demographic and health related variables in HSE and UKB, excluding people in UKB with history of cancer or CVD (n=302 009)**

| variable | | UK Biobank N Unweighted | UK Biobank % Unweighted | UK Biobank % Post Stratified |
|---|---|---|---|---|
| Age group | 40-49 | 84158 | 27·87 | 42·18 |
| | 50-59 | 105311 | 34·87 | 31·84 |
| | 60-70 | 112540 | 37·26 | 25·99 |
| Sex | Female | 159827 | 52·92 | 51·28 |
| | Male | 142182 | 47·08 | 48·72 |
| Education Qualification | College or university degree | 147786 | 48·93 | 34·06 |
| | Highschool diploma | 118080 | 39·10 | 41·11 |
| | Other/None | 36143 | 11·97 | 24·83 |
| Physical Activity | >=7·5 MET-hrs/wk | 264601 | 87·61 | 70·92 |
| | >0, <7·5 MET-hrs/wk | 33703 | 11·16 | 25·59 |
| | No PA | 3705 | 1·23 | 3·49 |
| Fruit and Vegetable Consumption | At least 10 portions/day | 25655 | 8·49 | 7·20 |
| | 5 to 9 portions/day | 133012 | 44·04 | 39·32 |
| | under 5 portions/day | 140396 | 46·49 | 52·18 |
| | Unknown | 2946 | 0·98 | 1·29 |
| Alcohol use frequency | Never | 10687 | 3·54 | 4·28 |
| | Previous | 8895 | 2·95 | 3·50 |
| | Current: < almost daily | 216405 | 71·66 | 72·78 |
| | Current: >=almost daily | 65891 | 21·82 | 19·38 |
| | Unknown | 131 | 0·04 | 0·07 |
| Smoking status | Never | 170417 | 56·43 | 52·29 |
| | Previous | 101227 | 33·52 | 33·61 |
| | Current | 30365 | 10·05 | 14·10 |
| Unhealthy index | Tertile 1: Most healthy | 100834 | 33·39 | 26·23 |
| | Tertile 2 | 111473 | 36·91 | 32·32 |
| | Tertile 3: Least healthy | 86637 | 28·69 | 40·11 |
| BMI category | Underweight | 1450 | 0·48 | 0·44 |
| | Normal | 104945 | 34·75 | 30·12 |
| | Overweight | 130888 | 43·34 | 43·95 |
| | Obese I | 48066 | 15·92 | 18·21 |
| | Obese II | 12355 | 4·09 | 5·22 |
| | Obese III | 4305 | 1·43 | 2·06 |

**Table 2· Adjusted ratio of hazard ratios (HRR) [a] of each lifestyle risk factor for all cause, CVD, and cancer mortality, excluding people with history of cancer or CVD (n=302 009)**

| | | | **Ratio of Hazard Ratios (Post Stratified/ Unweighted) *** | |
|---|---|---|---|---|
| **Variable** | **Level** | **All-Cause Mortality** | **CVD Mortality** | **Cancer Mortality** |
| Physical Activity Level | >=7·5 MET-hrs/wk | Reference | Reference | Reference |
| | >0, <7·5 MET-hrs/wk | 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 to 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) |

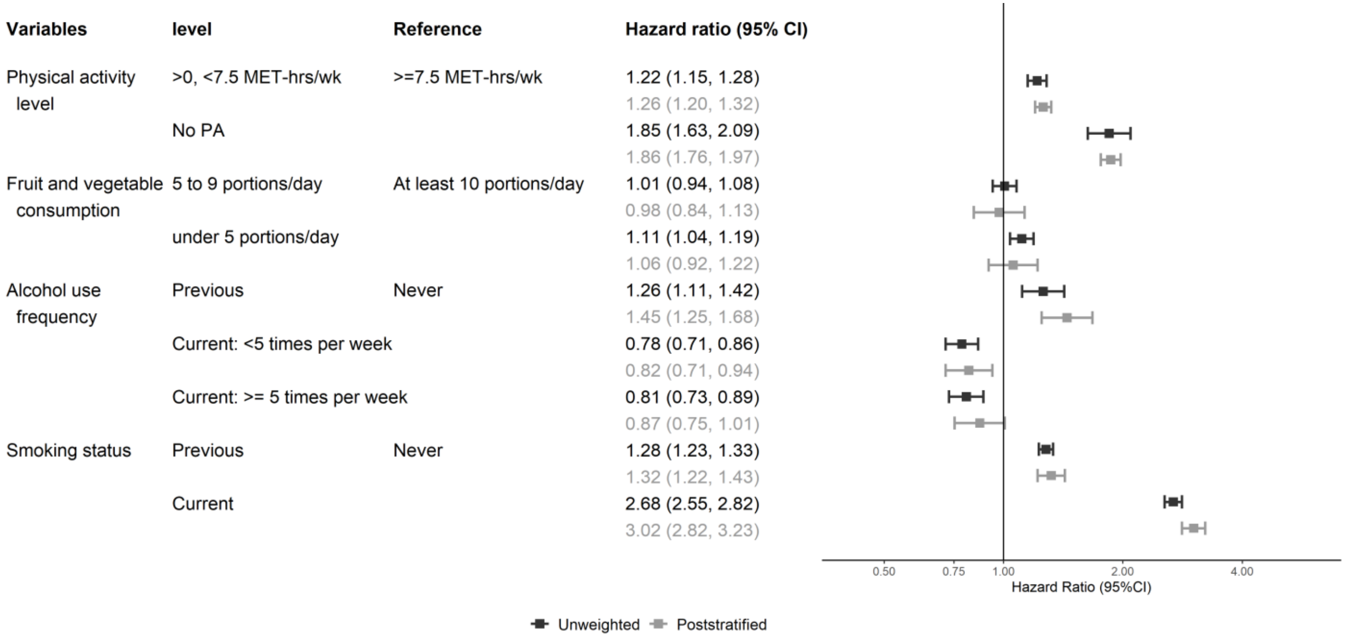
[a] To be equivalent to Health Survey for England weighed estimates

**Table 3·Adjusted ratio of hazard ratio ratios (RHR) [a] of lifestyle index for all cause, CVD, and cancer mortality, excluding people with history of cancer or CVD (n=302 009)**

|  |  |  | **Ratio of Hazard Ratios (Post Stratified/ Unweighted) *** |  |
|---|---|---|---|---|
| **Variable** | **Level** | **All-Cause Mortality** | **CVD Mortality** | **Cancer Mortality** |
| Lifestyle index | Tertile 1: Most healthy | Reference | Reference | Reference |
|  | Tertile 2 | 1.02 (0.98, 1.07) | 0.91 (0.78, 1.05) | 1.00 (0.90, 1.09) |
|  | Tertile 3: 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

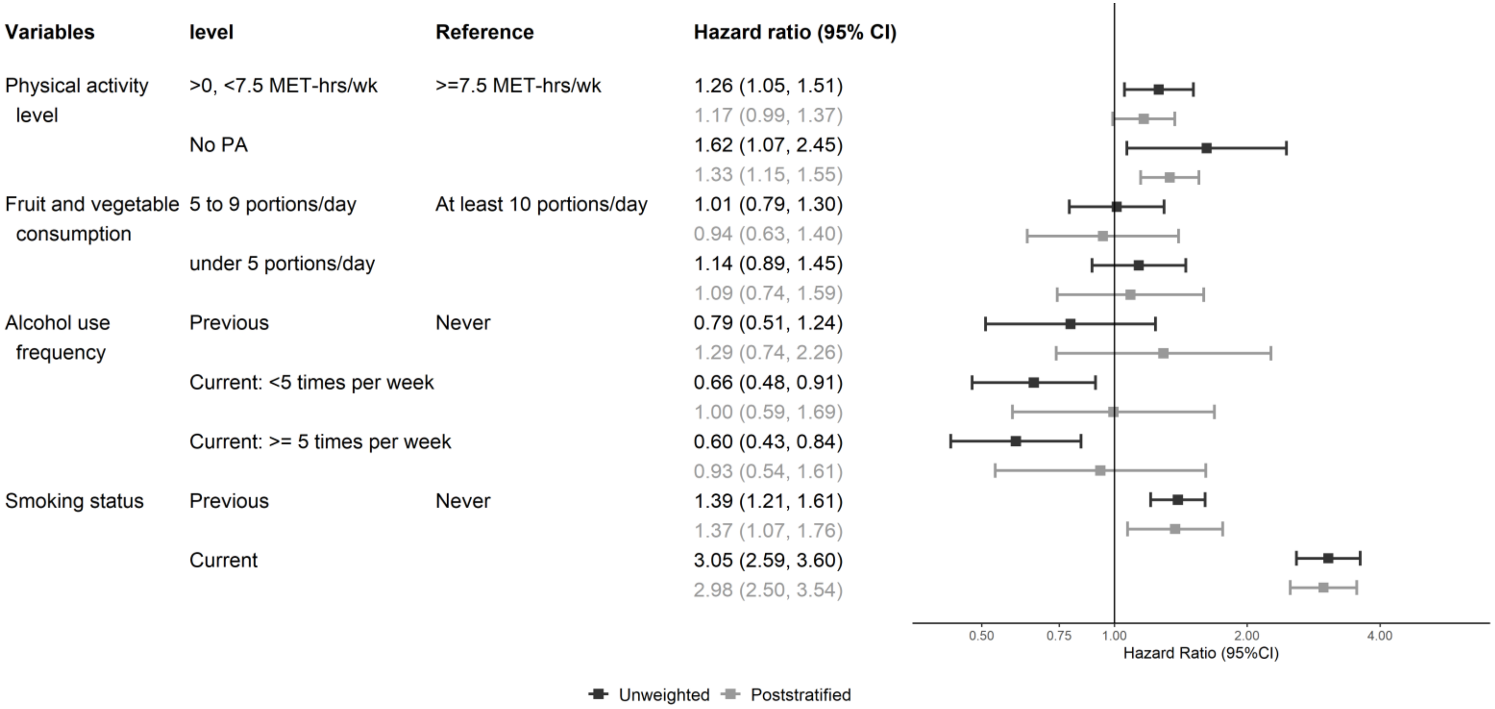
**Figure 1: Adjusted hazard ratio of each lifestyle risk factor for all-cause mortality, excluding people with history of cancer or CVD (n=302 009)**



| Variables | level | Reference | Hazard ratio (95% CI) |
|---|---|---|---|
| Physical activity level | >0, <7.5 MET-hrs/wk | >=7.5 MET-hrs/wk | 1.22 (1.15, 1.28) |
| | | | 1.26 (1.20, 1.32) |
| | No PA | | 1.85 (1.63, 2.09) |
| | | | 1.86 (1.76, 1.97) |
| Fruit and vegetable consumption | 5 to 9 portions/day | At least 10 portions/day | 1.01 (0.94, 1.08) |
| | | | 0.98 (0.84, 1.13) |
| | under 5 portions/day | | 1.11 (1.04, 1.19) |
| | | | 1.06 (0.92, 1.22) |
| Alcohol use frequency | Previous | Never | 1.26 (1.11, 1.42) |
| | | | 1.45 (1.25, 1.68) |
| | Current: <5 times per week | | 0.78 (0.71, 0.86) |
| | | | 0.82 (0.71, 0.94) |
| | Current: >= 5 times per week | | 0.81 (0.73, 0.89) |
| | | | 0.87 (0.75, 1.01) |
| Smoking status | Previous | Never | 1.28 (1.23, 1.33) |
| | | | 1.32 (1.22, 1.43) |
| | Current | | 2.68 (2.55, 2.82) |
| | | | 3.02 (2.82, 3.23) |

*Model additionally adjusted for age, sex, highest qualification.

**Figure 2: Adjusted hazard ratio of each lifestyle risk factor for all CVD mortality, excluding people with history of cancer or CVD (n=302 009)**

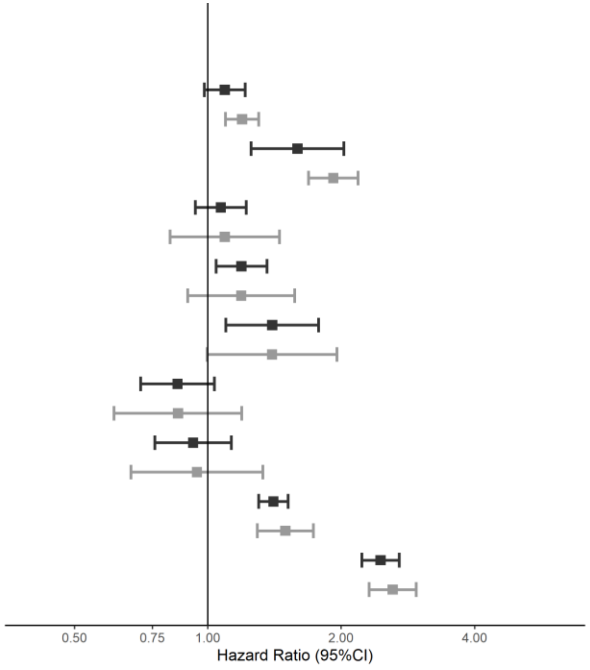


| Variables | level | Reference | Hazard ratio (95% CI) |
|---|---|---|---|
| Physical activity level | >0, <7.5 MET-hrs/wk | >=7.5 MET-hrs/wk | 1.26 (1.05, 1.51) |
| | | | 1.17 (0.99, 1.37) |
| | No PA | | 1.62 (1.07, 2.45) |
| | | | 1.33 (1.15, 1.55) |
| Fruit and vegetable consumption | 5 to 9 portions/day | At least 10 portions/day | 1.01 (0.79, 1.30) |
| | | | 0.94 (0.63, 1.40) |
| | under 5 portions/day | | 1.14 (0.89, 1.45) |
| | | | 1.09 (0.74, 1.59) |
| Alcohol use frequency | Previous | Never | 0.79 (0.51, 1.24) |
| | | | 1.29 (0.74, 2.26) |
| | Current: <5 times per week | | 0.66 (0.48, 0.91) |
| | | | 1.00 (0.59, 1.69) |
| | Current: >= 5 times per week | | 0.60 (0.43, 0.84) |
| | | | 0.93 (0.54, 1.61) |
| Smoking status | Previous | Never | 1.39 (1.21, 1.61) |
| | | | 1.37 (1.07, 1.76) |
| | Current | | 3.05 (2.59, 3.60) |
| | | | 2.98 (2.50, 3.54) |

*Model additionally adjusted for sex, highest qualification. Missing category for alcohol consumption and fruit and vegetable consumption not shown

**Figure 3: Adjusted hazard ratio of each lifestyle risk factor for cancer mortality, excluding people with history of cancer or CVD (n=302 009)**

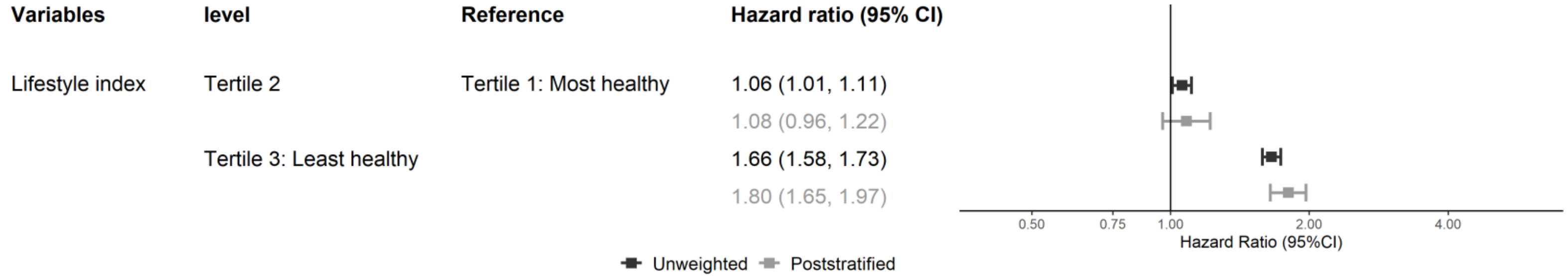
| Variables | level | Reference | Hazard ratio (95% CI) |
|---|---|---|---|
| Physical activity level | >0, <7.5 MET-hrs/wk | >=7.5 MET-hrs/wk | 1.09 (0.98, 1.21) |
| | | | 1.20 (1.10, 1.30) |
| | No PA | | 1.59 (1.25, 2.03) |
| | | | 1.92 (1.69, 2.18) |
| Fruit and vegetable consumption | 5 to 9 portions/day | At least 10 portions/day | 1.07 (0.94, 1.22) |
| | | | 1.09 (0.82, 1.45) |
| | under 5 portions/day | | 1.19 (1.04, 1.36) |
| | | | 1.19 (0.90, 1.57) |
| Alcohol use frequency | Previous | Never | 1.40 (1.10, 1.78) |
| | | | 1.40 (1.00, 1.96) |
| | Current: <5 times per week | | 0.85 (0.71, 1.03) |
| | | | 0.86 (0.61, 1.19) |
| | Current: >= 5 times per week | | 0.93 (0.76, 1.13) |
| | | | 0.94 (0.67, 1.33) |
| Smoking status | Previous | Never | 1.41 (1.30, 1.52) |
| | | | 1.50 (1.29, 1.73) |
| | Current | | 2.45 (2.23, 2.70) |
| | | | 2.61 (2.31, 2.95) |



■ Unweighted ■ Poststratified

*Model additionally adjusted for sex, highest qualification. Missing category for alcohol consumption and fruit and vegetable consumption not shown

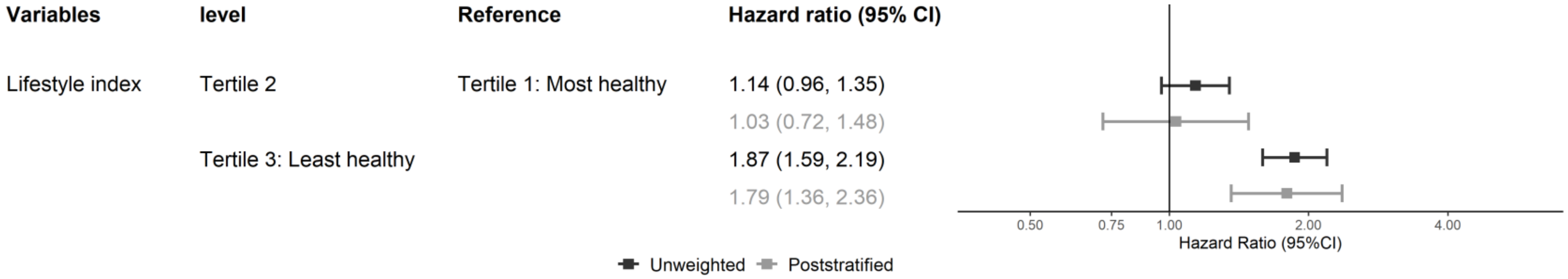
**Figure 4: Adjusted hazard ratio* of lifestyle index tertiles for all-cause mortality, excluding people with history of cancer or CVD (n=302 009)**



| Variables | level | Reference | Hazard ratio (95% CI) |
|---|---|---|---|
| Lifestyle index | Tertile 2 | Tertile 1: Most healthy | 1.06 (1.01, 1.11) |
| | | | 1.08 (0.96, 1.22) |
| | Tertile 3: Least healthy | | 1.66 (1.58, 1.73) |
| | | | 1.80 (1.65, 1.97) |

*Model adjusted for age, sex, highest qualification.

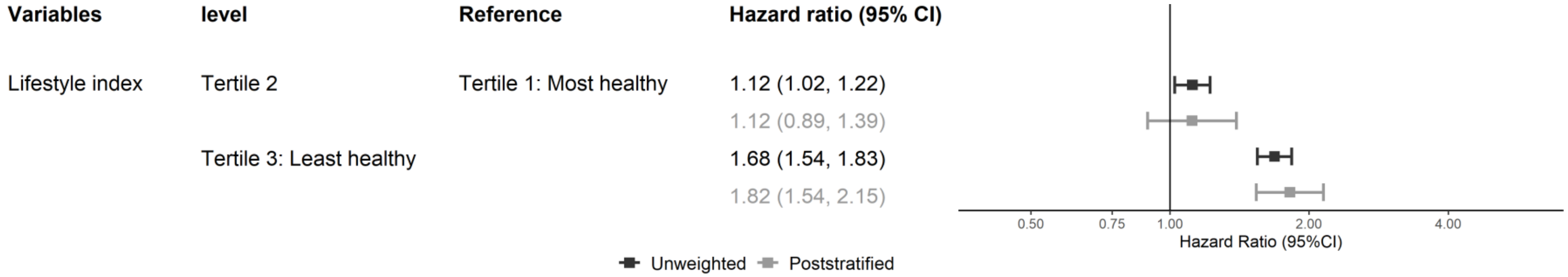
**Figure 5: Adjusted hazard ratio* of tertiles of lifestyle index for all CVD mortality, excluding people with history of cancer or CVD**



| Variables | level | Reference | Hazard ratio (95% CI) |
|---|---|---|---|
| Lifestyle index | Tertile 2 | Tertile 1: Most healthy | 1.14 (0.96, 1.35) |
| | | | 1.03 (0.72, 1.48) |
| | Tertile 3: Least healthy | | 1.87 (1.59, 2.19) |
| | | | 1.79 (1.36, 2.36) |

*Model adjusted for age, sex, highest qualification.

**Figure 6: Adjusted hazard ratio* of tertiles of lifestyle index for all cancer mortality, excluding people with history of cancer or CVD**



| Variables | level | Reference | Hazard ratio (95% CI) |
|---|---|---|---|
| Lifestyle index | Tertile 2 | Tertile 1: Most healthy | 1.12 (1.02, 1.22) |
| | | | 1.12 (0.89, 1.39) |
| | Tertile 3: Least healthy | | 1.68 (1.54, 1.83) |
| | | | 1.82 (1.54, 2.15) |

*Model adjusted for age, sex, highest qualification.