USC Dornsife Center for Economic and Social Research

USC Schaeffer Leonard D. Schaeffer Center for Health Policy & Economics

Should representativeness be avoided? Reweighting the UK Biobank corrects for pervasive selection bias due to volunteering

Sjoerd van Alten MSc, Benjamin W. Domingue PhD, Jessica Faul PhD, Titus Galama PhD, Andries T. Marees PhD

Paper No: 2023-002

CESR-SCHAEFFER

WORKING PAPER SERIES

Should representativeness be avoided? Reweighting the UK Biobank corrects for pervasive selection bias due to volunteering

Sjoerd van Alten MSc[1,2], Benjamin W. Domingue PhD[3], Jessica Faul PhD[4], Titus Galama PhD[1,2,5,6], and Andries T. Marees PhD[1]

[1]*School of Business and Economics, Vrije Universiteit Amsterdam, Amsterdam, Netherlands*
[2]*Tinbergen Institute, Amsterdam, Netherlands*
[3]*Graduate School of Education, Stanford University, Stanford, CA, USA*
[4]*University of Michigan*
[5]*Center for Economic and Social Research and Department of Economics, University of Southern California Dornsife, Los Angeles, CA, USA*
[6]*Erasmus School of Economics, Erasmus University Rotterdam, Rotterdam, Netherlands*

June 12, 2023

Correspondence to Sjoerd van Alten, MSc, School of Business and Economics, Vrije Universiteit Amsterdam, Amsterdam, Netherlands

s.j.d.van.alten@vu.nl

Abstract

We investigate to what extent volunteer-based sampling of large-scale biobanks biases associations and estimate inverse probability (IP) weights to correct for such bias. Using the UK Biobank (UKB) as an example of a large-scale volunteer-based cohort, and population-representative data from the UK Census as a reference, we compare 21 bivariate associations in both data sets. Volunteer bias in all associations as naively estimated in the UKB is substantial, and in some cases leads to estimates of the wrong sign. For example, older individuals in the UKB report being in better health. Correcting for volunteer bias using IP weights is therefore advised. Applying IP weights reduces 87% of volunteer bias on average and suggests volunteer-based sampling reduces the effective sample size of the UKB to ~32% of its original size. To aid the construction of the next generation of biobanks, we provide suggestions on how to best ensure representativeness in a volunteer-based design.

1 Introduction

In the last decade, large-scale biobanks (N>100,000) have become a key resource for medical, epidemiological, genetic, and social scientific research.[1–4] The UK Biobank (UKB) is a well-known example and has been used in over 3,200 peer-reviewed publications since its release in 2012.[5] Samples of such large sizes are vital for the identification of small effects with sufficient power and make, for example, genome-wide association studies feasible.[6] However, with a main focus on size, data collection has relied on respondents volunteering to participate, at the expense of representativeness.[3,7–11] As a result, these data sets exhibit "healthy volunteer bias": respondents tend to be healthier and of higher socioeconomic status than the population from which they were sampled.[9–11]

The degree to which selection bias challenges subsequent scientific investigations is contested. One view is that deviations from representativeness are appropriate if the goal is to uncover causal relationships (e.g., the effect of an exposure X on an outcome Y).[12,13] Rothman et al.[12] argue that well-designed studies will typically require construction of a highly artificial (i.e., non-representative) setting to elucidate the key causal mechanism. Hence, they claim that "representativeness should be avoided". They contrast such causal studies with descriptive studies (e.g., obtaining estimates of disease prevalence in a certain population), for which representative sampling is imperative.

We agree with this view that careful study design is essential. As long as the exposure of interest is unrelated to any other characteristics that might influence the outcome (including sample selection), internal validity is not at risk, and causal effects may be identified. In practice, however, exposures of interest are rarely unrelated to other participant characteristics in observational data. Understanding an outcome-exposure relationship then typically starts with estimating an association between two variables, possibly controlling for confounding factors. While some claim that generalisable associations can be estimated even in non-representative data,[8,14–16] in a volunteer-based sample, study participation in itself can serve as a "collider" on the path from the exposure to the outcome, resulting in bias whenever the association is estimated within the sample.[7,17–20]

Simulations confirm these intuitions (see Figure 1): volunteering into a data set leads to bias and, troublingly, the direction of bias is not known a priori as it depends on which variables influence selection into the data set and how these variables, in turn, relate to the exposure and outcome. Volunteer bias can lead to attenuation, underestimation, or overestimation of effect sizes. This may result in false positives *or* false negatives. It is even possible that volunteer bias results in estimated effect sizes that are of the opposite sign than in the underlying sampling population (see *scenario 2a* in Figure 1).

The inclusion of control variables in regression models is no solution and may even increase bias in the estimated associations when some of these control variables are related to study participation.[21] Understanding how volunteering biases estimates is vital to understanding the costs and benefits of large (but non-representative) vis à vis smaller (but more representative) data sets. In addition, methods to correct for such bias are needed.

In this study, we will first examine the degree of volunteer bias in association statistics estimated in one of the largest and most utilized biobanks, the UK Biobank (UKB). We find it to be pervasive. We then use inverse probability (IP) weights, constructed using external, representative data from the UK Census, to correct for volunteering. We show that it is possible to correct for up to 87% of volunteer bias using these weights. Our findings illustrate how weighting for underlying sampling populations is essential. We evaluate the costs and benefits of volunteer-based sampling and provide guidance on how to improve the design of the next generation of biobanks. The IP weights constructed in this study are available as a data field to researchers using the UKB.

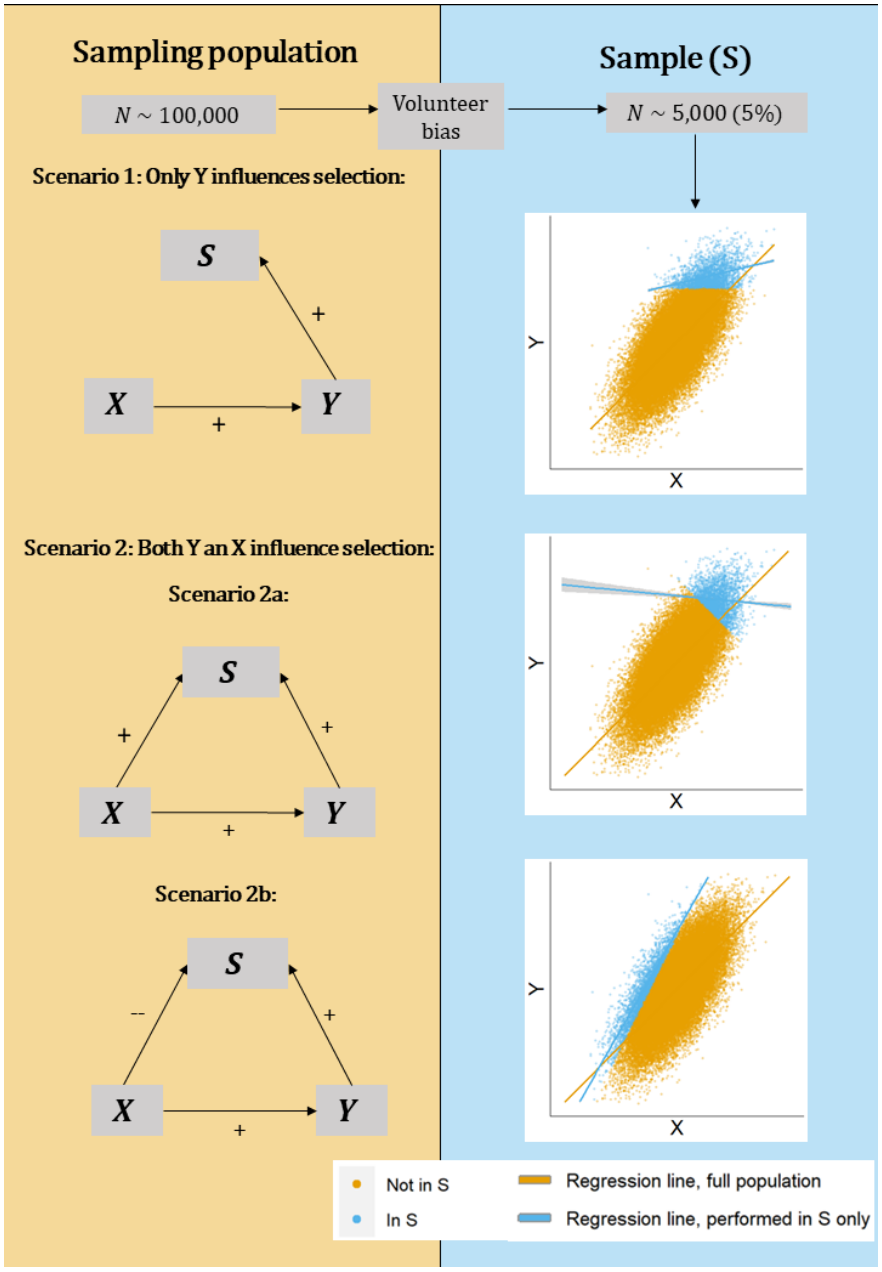


Figure 1: A simulated example of spurious associations due to volunteer bias in a selected sample S. In this example, we simulate an exposure $X \sim \mathcal{N}(0,1)$ and an outcome $Y = X + \epsilon$, $\epsilon \sim \mathcal{N}(0,1)$. The effect of X on Y is assessed using bivariate linear regression. X and Y are positively related in the population (the orange and blue dots combined) with slope 1. This is reflected by the orange regression lines in each of the three scatter plots. In scenario 1, individuals with higher values of Y, here modelled by a threshold $Y > Y^*$, select into the sample (S; the blue points) and there is no selection based on X. As a result, the regression line estimated within the selected sample S (the blue line) is attenuated towards the null. In scenario 2a, individuals with higher values of Y and with higher values of X, here modelled by a threshold $0.5Y + 0.5X > Z^*$, select into the sample S. As a result, the regression is downwards biased (and of incorrect sign). In scenario 2b, individuals with higher values of Y, but lower values of X, select into the sample S (here modelled by a threshold $Y - 2X > Z^*$). Now, the bias is upwards, and the effect of X on Y is overestimated. Note further that, in all scenarios, the standard deviations of X and Y estimated within the selected sample S are smaller, compared to the standard deviations in the full population, as a consequence of selection (this can be seen from the distributions of the blue dots in the simulations, which are more narrow).

2 Results

Data and analysis

We will refer to three main data sets: the *UKB*, the *UKB-eligible Census*, and the *Weighted UKB*. The goal of the inverse probability weighting (IP weighting) procedure is to make the UKB representative of the UKB-eligible population (i.e. all individuals who received an invitation to participate in the UKB). The UKB-eligible population differs from the full population of the United Kingdom in two important aspects. First, the age range is restricted as the UKB only sampled individuals aged between 40 and 69. Second, the geographic range is restricted as only individuals who lived close to any of 22 assessment centres were sampled. These assessment centres were mostly located in urban areas. As Figure 2 shows, this led to a highly uneven sampling of geographic areas and left out large swaths of Great Britain's land area and population.

To obtain a reference sample that is representative of this UKB-eligible population, we created the *UKB-eligible Census*. This is a subsample of microdata from the 2011 UK Census. 2011 is the Census year that is closest to the period of UKB data collection. The UK Census is highly representative, due to its high response rate of $> 95\%$[22]. To create the UKB-eligible Census, we restricted the UK Census microdata to be within the correct age range, and residing within the sampling radii around the 22 assessment centres from which UKB respondents were sampled, using the respondents' birth cohort and region of residence, measured by 285 distinct grouped local authority (GLA) regions (see Methods and Supplementary Note S1 for additional detail).

To assess whether volunteer bias affects the UKB, we compared, in the UKB and the UKB-eligible Census, means, standard deviations, and regression coefficients as estimated by bivariate linear models. To obtain IP weights for UKB respondents, we estimated a probit model that predicts the UKB participation decision on concatenated data from the UKB ($UKB = 1$) and the UKB-eligible Census ($UKB = 0$), using 4,820 possible predictors of selection and variable selection using LASSO. These predictors were based on year of birth, sex, ethnicity, educational attainment, employment status, GLA region of residence, tenure of dwelling, number of cars in the

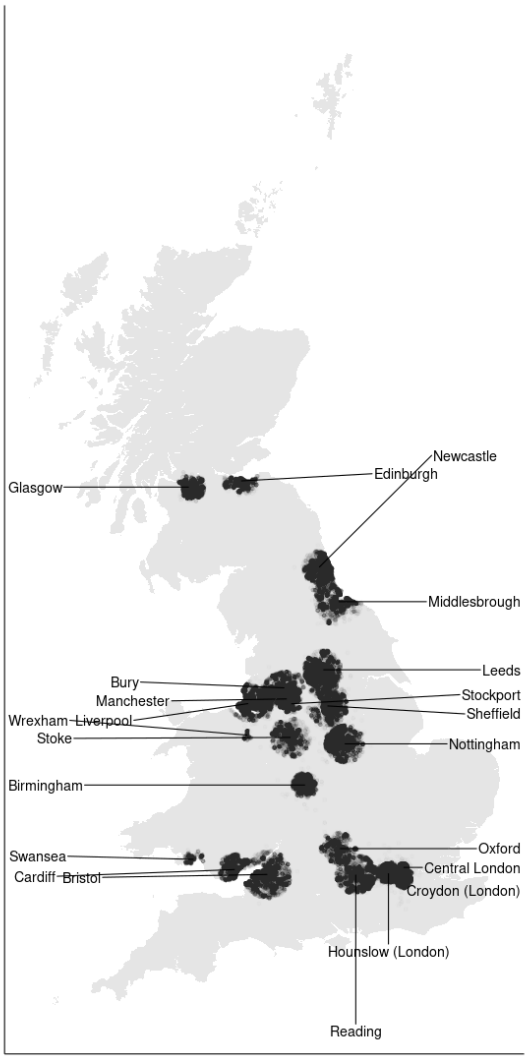


Figure 2: **UKB respondents' location of residence at assessment day.** Each black dot corresponds to the place of residence of a UKB respondent. Only respondents who lived near any of the UKB assessment centres (annotated), which were predominantly located in urban areas, received an invitation to participate in the UKB.

household, self-reported health, and whether the individual lived in a one-person household. To prevent overfitting, the data was randomly subdivided into five 20% holdout samples. A LASSO model was trained separately in each affiliated 80% training sample and then used to estimate UKB participation probabilities in its holdout sample only. These participation probabilities were then used to estimate IP weights for all UKB respondents that are inversely proportional to their estimated probability of participation

We assess the performance of these weights by estimating weighted counterparts of the aforementioned means, standard deviations, and regression coefficients in the UKB. We refer to such statistics *as if they were* obtained from a synthetic third data set, *the weighted UKB*. The methods section provides additional detail on weight construction and the various analyses used to assess the extent of volunteer bias in the UKB.

IP weights provide evidence for volunteer bias in the UKB

Our LASSO probit model adequately discriminates between UKB and UK Census observations, with an area under the curve (AUC) of 0.772[1] ($InterModelVigorish(IMV) = 0.006$)[2] when holding out the first holdout sample (these statistics were very similar for the other four holdout samples). This provides a first indication that the UKB and UK Census are distinct and that the predictors available to us adequately capture UKB participation. Figure 3 shows a variable importance plot. Variable importance is quantified by the relative increase in AUC after leaving each variable out in turn, relative to the full model's AUC (see Methods). All variables that we included are relevant predictors for UKB volunteering. The variables region and year of birth drive most of the performance of the LASSO model. These findings thus illustrate that (1) respondents in the UKB and UK Census differ substantially, consistent with the UKB being non-representative due to volunteering, and (2) that such volunteering is associated with many different factors such as geographic location, age, socioeconomic factors, and health-related factors. A rich model with many different predictors, such as our LASSO probit model, is thus essential to creating IP weights that are sufficiently well-measured to correct for volunteer bias.

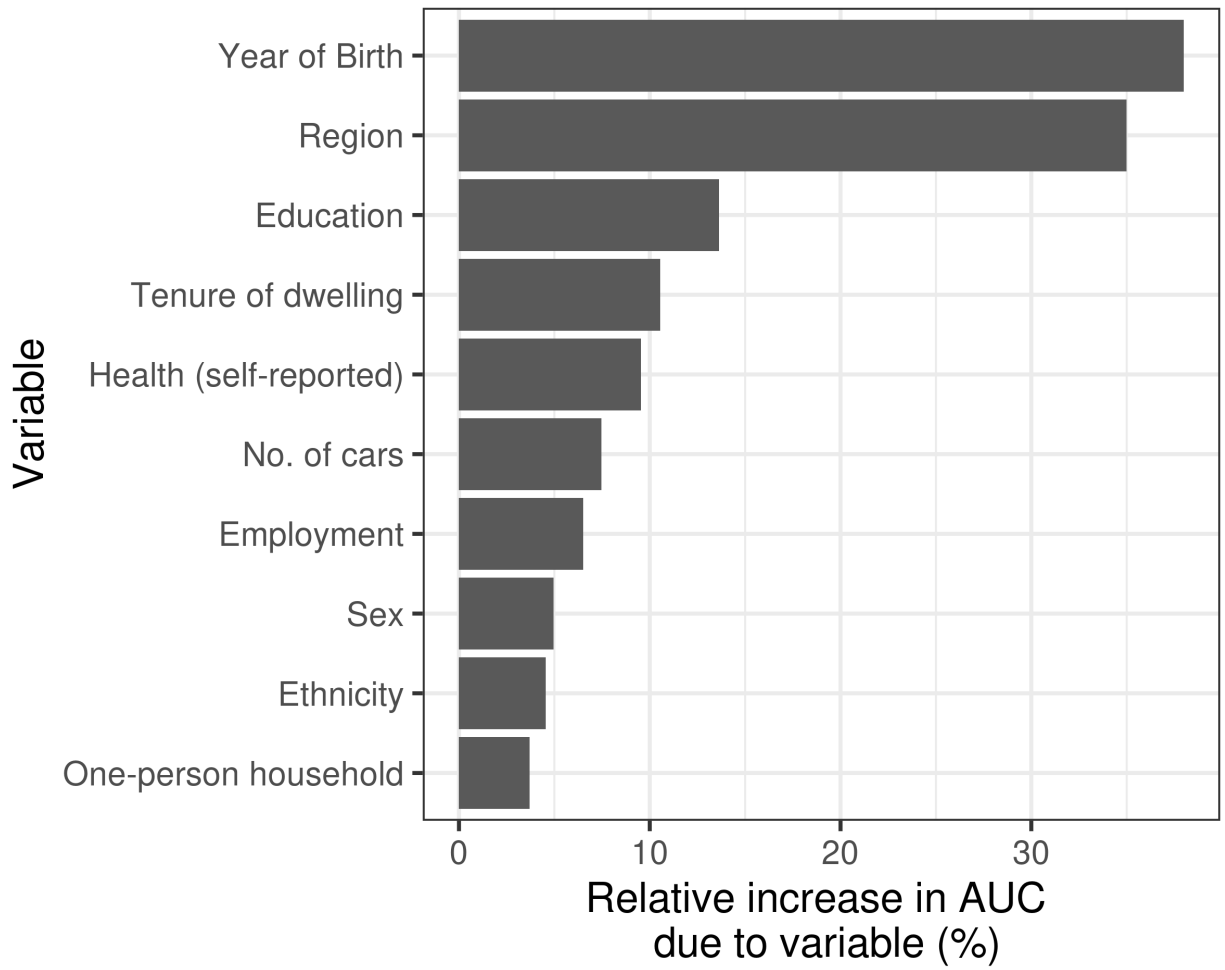


Figure 3: Variable importance plot showing the importance of each variable in predicting UKB volunteering on the first holdout sample. Variable importance is assessed by the relative increase in the AUC, relative to the full model's AUC. The variable importance plots for the LASSO model estimated on data that holds out the 4 other folds look very similar and are therefore not shown here.

Differences between the UKB and UK Census are consistent with healthy volunteer bias

A comparison of the UKB-eligible Census and the UKB reveals substantial non-random selection of UKB participants from the UKB-eligible population (Table 1). Compared to the UKB-eligible population, individuals who participated in the UKB were older, healthier, higher educated, of higher socioeconomic status, and more likely to be white. For all variables included in the table, means differ significantly between the UKB and the UKB-eligible population ($P < 10^{-8}$). Some differences are large. For example, individuals in the UKB-eligible population were over twice as likely to report being in poor health, compared to those who decided to participate in the UKB (9.3 versus 4.4 percent), despite the fact that UKB participants are ~ 3.5 years older on average.

For all four discrete and continuous variables in Table 1 we also observe smaller standard deviations in the UKB, compared to the UKB-eligible population, consistent with non-random (over)sampling of those more likely to volunteer (see Figure 1). Note that for binary variables, it is not possible to tell whether an increase or a decrease in the standard deviation is consistent with selective sampling (Supplementary Note S5).

After IP weighting (rows labelled "Weighted UKB"), the UKB becomes much more comparable to the UKB-eligible Census, both in terms of its means and its standard deviations. Hence, IP weighting is successful in correcting these statistics for volunteer bias.

Volunteer bias affects associations estimated in the UKB

We next test to what extent associations estimated in the UKB are affected by volunteer bias. Figure 4 plots coefficients of bivariate linear probability models estimated in the UKB-eligible Census data (yellow bars) and the UKB (blue bars). The width of these bars indicates the 95% confidence interval. P-values for the null hypothesis that the coefficients in the UKB-eligible Census and the UKB are the same (shown to the right of each association test) suggest that volunteering severely biases associations between basic demographic variables. *All* are statistically significantly different from one another at $P < 10^{-8}$, and the size of these differences is often large. For example, the association between being employed and reporting poor health among

Var	Dataset	N	Min	Mean	Max	SD
Discrete/Continuous variables						
Year of Birth	UKB-eligible Census	687 491	1938	1955.101	1968	9.039
	UKB	491 268	1938	1951.545	1968	8.261
	Weighted UKB	491 240	1938	1954.818	1968	8.905
Self-reported health	UKB-eligible Census	687 489	1	2.625	3	0.648
	UKB	488 956	1	2.701	3	0.546
	Weighted UKB	488 941	1	2.638	3	0.633
Years of education	UKB-eligible Census	687 489	7	12.608	20	5.071
	UKB	480 251	7	13.585	20	4.987
	Weighted UKB	480 233	7	12.892	20	5.042
No. of cars	UKB-eligible Census	683 138	0	1.364	4	0.966
	UKB	487 832	0	1.547	4	0.870
	Weighted UKB	487 820	0	1.392	4	0.962
Bivariate variables						
Female	UKB-eligible Census	687 491	0	0.508	1	0.500
	UKB	491 268	0	0.546	1	0.498
	Weighted UKB	491 240	0	0.512	1	0.500
University or equivalent	UKB-eligible Census	687 489	0	0.278	1	0.448
	UKB	480 251	0	0.336	1	0.472
	Weighted UKB	480 233	0	0.290	1	0.454
Reports “poor health"	UKB-eligible Census	687 489	0	0.093	1	0.290
	UKB	488 956	0	0.044	1	0.206
	Weighted UKB	488 941	0	0.085	1	0.279
Has paid work	UKB-eligible Census	687 491	0	0.609	1	0.488
	UKB	486 711	0	0.579	1	0.494
	Weighted UKB	486 693	0	0.614	1	0.487
Retired	UKB-eligible Census	687 491	0	0.249	1	0.432
	UKB	486 711	0	0.341	1	0.474
	Weighted UKB	486 693	0	0.250	1	0.433
Incapacitated	UKB-eligible Census	687 491	0	0.069	1	0.254
	UKB	486 711	0	0.033	1	0.179
	Weighted UKB	486 693	0	0.065	1	0.247
Unemployed	UKB-eligible Census	687 491	0	0.033	1	0.179
	UKB	486 711	0	0.016	1	0.127
	Weighted UKB	486 693	0	0.032	1	0.176
House owner	UKB-eligible Census	683 138	0	0.736	1	0.441
	UKB	484 157	0	0.899	1	0.302
	Weighted UKB	484 147	0	0.753	1	0.431
White ethnicity	UKB-eligible Census	687 491	0	0.888	1	0.315
	UKB	491 268	0	0.946	1	0.226
	Weighted UKB	491 240	0	0.893	1	0.309
One-person household	UKB-eligible Census	683 138	0	0.181	1	0.385
	UKB	487 922	0	0.185	1	0.388
	Weighted UKB	487 908	0	0.186	1	0.389

Table 1: **Summary statistics for the UKB-eligible Census, the UKB, and the Weighted UKB**. For all variables, mean values differ significantly between the UKB-eligible Census and the UKB, all with $P < 10^{-8}$ as obtained by a Z-test. After applying IP weighting, the means and standard deviations in the Weighted UKB are closer to those of the UKB-eligible Census.

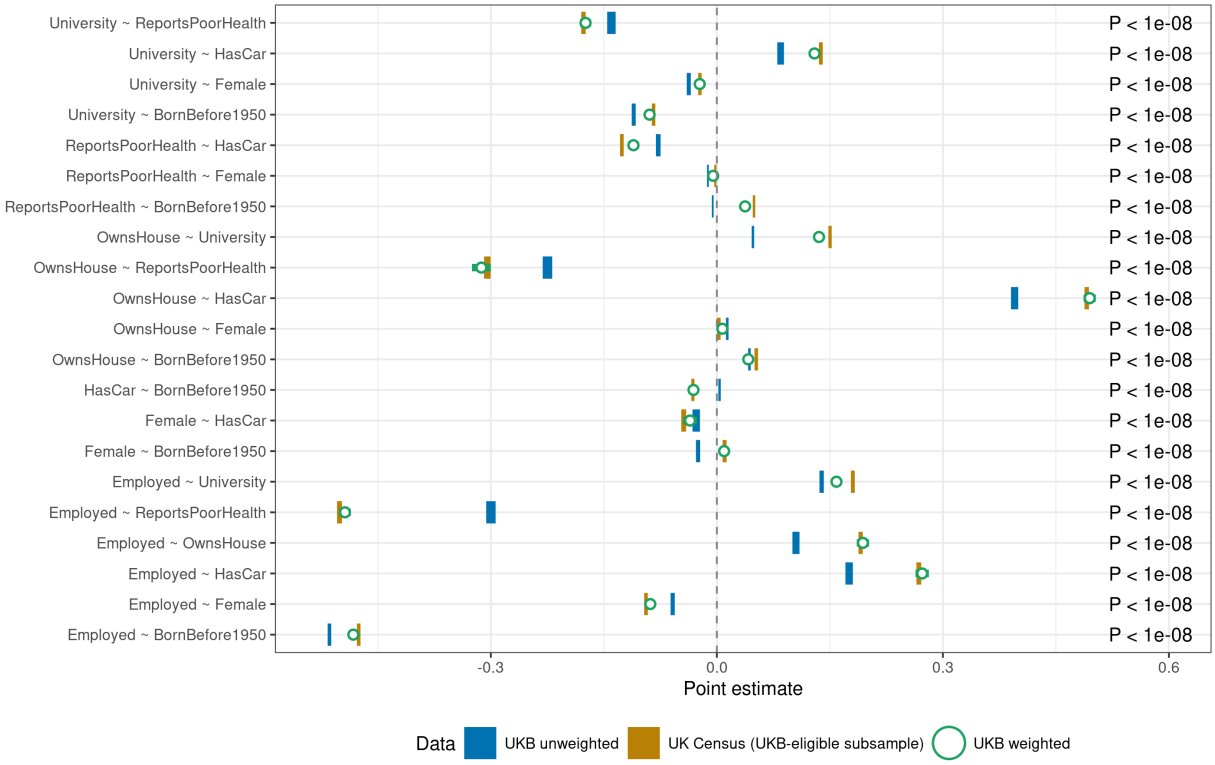


Figure 4: Estimated associations based on bivariate linear probability models in the UKB (solid blue bars), UKB-eligible Census (solid yellow bars), and Weighted UKB (open green circles). Bar widths indicate 95% confidence intervals (heteroskedasticity-robust standard errors). All blue and yellow bars are highly significantly different from one another ($P < 10^{-8}$). IP weighting leads to substantially improved associations: the open green circles are in all cases substantially closer to the yellow bars, than the blue bars are to the yellow bars.

UKB respondents (CI_{95}=[-0.306; -0.294]) is substantially weaker than in the broader UKB-eligible population, (CI_{95}=[-0.504;-0.498]). Estimating associations in the UKB can thus result in sizeable distortions of the actual associations in the underlying sampling population.

While most "UKB unweighted" estimates in Figure 4 are at least in the correct direction, volunteer bias can also lead to false positives or an incorrect sign. For example, in the UKB individuals born before 1950 were *less* likely ($CI_{95} = [-0.007; -0.004]$) to report being in poor health than younger individuals, contrary to the evidence that health deteriorates as we age. In the UKB-eligible Census, however, we observe the expected positive association ($CI_{95} = [0.0476; 0.0509]$). Another example is that women in the UKB were less likely to have been born before 1950 than men ($CI_{95} = [-0.0278; -0.0221]$), whereas the reverse holds in the UKB-eligible Census ($CI_{95} = [0.008; 0.0129]$), consistent with men living shorter lives than women.

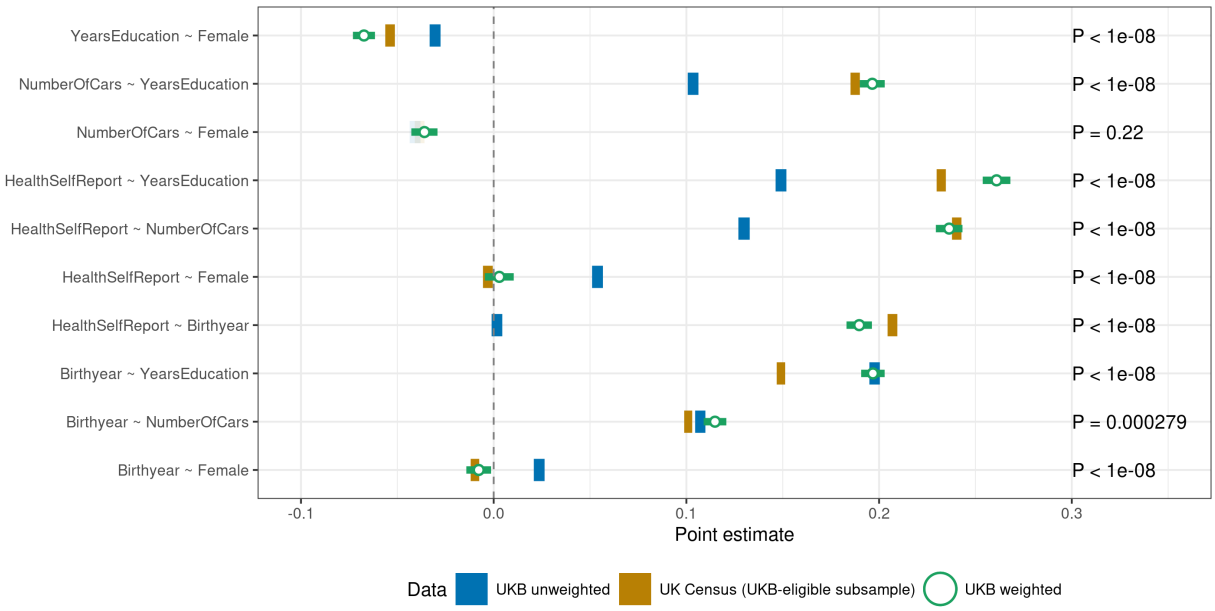


Figure 5: Estimated associations based on linear models in UKB (solid blue bars), UKB-eligible Census (solid yellow bars), and Weighted UKB (open green circles). All variables are standardised to have a mean of zero and a variance of one. P-values for the null hypothesis that the associations in the UKB and UKB-eligible Census are the same are shown. All solid blue and yellow bars are significantly different from one another. Bar widths indicate 95% confidence intervals (heteroskedasticity-robust standard errors). IP weighting leads to substantially improved associations: the open green circles are substantially closer to the yellow bars, than the blue bars are to the yellow bars. The standard errors around the weighted estimates of these models are used to estimate the equivalent sample size of the UKB (see Methods).

Last, women were more likely to own a house in the UKB than were men ($CI_{95} = [0.012; 0.016]$), whereas in the UKB-eligible Census, the corresponding point estimate is indistinguishable from zero, i.e. here we obtained a false positive result in the UKB.

We obtained similar results for bivariate linear models between several discrete and/or continuous variables (Figure 5). Here, once again the UKB fails to show that younger people were healthier, and incorrectly indicates that women were younger on average than men. As in Figure 4, most other associations in Figure 5 are of the same sign in both the UKB and the UKB-eligible Census population, but are nonetheless substantially different from one another.

IP weighting is consistent with reducing volunteer bias (variables shared with the UK Census)

IP weighted regressions correcting for volunteer bias (see green open circles in Figure 4, including 95% confidence intervals) show that the associations in the Weighted UKB are less biased. These

weighted estimates are substantially closer to the estimates in the UKB-eligible Census (yellow bars) compared to the original unweighted UKB associations (blue bars). The average bias reduction over all models shown in Figure 4 is 87%. For associations between discrete and/or continuous variables average bias reduction is 78% over all estimated associations (green open circles in Figure Figure 5).

Notably, including the variables that are found to influence UKB participation as control variables in the regression model does not mitigate, but rather *increases* volunteer bias in the estimated associations (see Supplementary Note S6). By contrast, IP weighting using our weights remains successful in reducing volunteer bias for these models, regardless of whether the same variables are included as control variables or not (see Supplementary Note S6). This should serves as a useful reminder that, although control variables are typically used to address various sources of confounding, they are no safeguard against volunteer bias. An IP weighting strategy should be used instead.

Volunteer bias reduces the effective sample size of the UKB to ~32% of its original size

The effective sample size provides an estimate of the size of a hypothetical population-representative sample with the same power as the weighted UKB. We used two methods to arrive at an effective sample size for the UKB (see Methods). The first uses the distribution of the weights and obtains an effective sample size of 200,810 (40.8% of the size of the unweighted UKB). The second is regression-specific. It results in effective sample sizes for each of the estimated association statistics that range between 118,370 and 202,999 (24.1%-41.3%), with an average of 156,698 (31.9%) across all models. Hence, after weighting, the information obtained from the full UKB sample of 491,268 is equivalent to that obtained from a hypothetical representative sample of between 118,370 and 202,999 individuals.

IP weights are robust to missing variables

Our IP weights may be less effective in correcting for volunteer bias for variables that were not available in the UK Census (and could therefore not be included in our UKB participation model). To test this, we re-estimated the weighted associations in Figure 4, with newly created weights based on UKB participation models that leave out the dependent and independent variable of the specific association tested (see Supplementary Note S7, Supplementary Figure S5). Reassuringly, using these "leave-two-variables-out" IP weights, we still capture 69% of the volunteer bias on average.

IP weighting consistent with reducing volunteer bias (variables <u>not</u> shared with UK Census).

Another test of the performance of our IP weights is to evaluate the effect of IP weighting on variables that were unique to the UKB, i.e. that were not measured in the Census, and to see if these corrections were consistent with volunteer bias. We estimated means of variables measured in the UKB *but not* in the UK Census with and without applying IP weighting (Supplementary Table S2). Overall, the UKB after weighting is younger, heavier, in worse (mental) health, more likely to smoke, and of lower socioeconomic status, compared to the unweighted UKB sample, which is consistent with healthy volunteer bias. For example, weighting the UKB increases the Townsend deprivation index from -1.317 to -0.414, consistent with oversampling of high socioeconomic status individuals. Weighting also increases the prevalence of substance use and the prevalence of various health conditions, such as reported chest pain or a disability, consistent with oversampling of healthier individuals. For other variables, for example, anthropometric ones, the application of IP weights results in negligible differences: these variables are likely unrelated to volunteering for the UKB. IP weighting for associations between these types of variables is likely of lesser importance.

3 Discussion

We uncovered substantial non-random selection of UKB participants comparing means and standard deviations of variables between the UKB and UKB-eligible population. Volunteer bias is large and highly statistically significant in all 21 associations we tested. In some cases, volunteer bias leads to false positive associations or associations that are of the incorrect sign. For example, older individuals in the UKB reported being in better health. Constructing IP weights to correct for volunteer bias we were able to correct for 87% of volunteer bias on average for associations tested between binary variables, and 78% for discrete and/or continuous variables. By contrast, naively controlling for variables that influence UKB participation increased, rather than decreased, volunteer bias.

Earlier studies of volunteer bias in the UKB exclusively focused on mortality as an outcome,[16,23] and concluded that volunteer bias is of little importance. These studies used the Health Survey of England (HSE) and the Scottish Health Survey (SHS). For example, Batty et al.[16] compared risk factors for mortality between the UKB and HSE/SHS, while Stamatakis et al.[23] estimated IP weights using HSE data to correct such risk factors for volunteer bias. For most associations, the weights of Stamatakis et al. made little difference, except for a protective association found between alcohol use and cardiovascular-related mortality that disappeared after correcting for volunteer bias[23].

We, however, have investigated an extensive and comprehensive set of associations between socio-economic and health-related variables, and find that volunteer bias matters substantially. Our study distinguishes itself in at least five ways. First, we compared the UKB to the UK Census, which, with a response rate of $> 95\%$, is highly representative of the UK population, compared to the HSE/SHS, which have a lower mean response rate ($\sim 68\%$)[16] and may therefore not be sufficiently representative due to, e.g., volunteering. Second, the use of rich UK Census microdata allowed us to include many more variables, and their interactions, improving the precision of the weights. Third, we used fine location information to restrict the UK Census data to the UKB-eligible population, which resides around 22 highly urbanised areas, thereby more precisely targeting the population that received an invitation to participate in the UKB. By contrast, HSE/SHS does not contain

detailed geographic information. This is of key methodological importance, as sufficient overlap between the UKB and the target population is essential to the validity of IP weights estimation.[24] Fourth, the large sample size of the UK Census aids more precise IP weight estimation (687,491 respondents in our final sample, compared to 6,666 in the HSE, used by Stamatakis et al.)[23]. Last, our weights are estimated using predictors of selection bias that were missing in previous analyses, most importantly, region of residence, which is one of the strongest predictors of selection into the UKB (see Figure 3). As a result, due to our precisely estimated weights, we find that weighting association statistics in the UKB substantially alters virtually every association statistic that we considered, unlike previous efforts. We highly recommend that researchers who use UKB data use our weights to address the robustness of their estimates to selection bias.

There are some limitations that researchers need to keep in mind. Our proposed method of IP-weighted regression reduces volunteer bias, but increases standard errors. However, this does not necessarily imply a decrease in power, as volunteer bias may take the form of attenuation bias, such that correcting for volunteer bias could result in larger effect sizes (see Figure 1). We were limited to the inclusion of ten variables that the UKB and UK Census have in common. Although these variables contain rich information (e.g., region of residence consists of 285 categories), these variables are, by and large, socio-demographic and health-related. Our method therefore mostly corrects for socioeconomic status-, health-, and demographic-related components of study participation. There may exist unobserved variables that nonetheless explain a substantial part of UKB participation, e.g., personality characteristics. Nonetheless, our weights reduce a substantial part (87%) of volunteer bias in UKB-estimated associations. Reassuringly, even when leaving variables out of IP weight estimation, we could correct for 69% of volunteer bias on average. Further, applying IP weights to variables in the UKB that were not available in the UK Census, and therefore not included in the selection model, we found that observed changes in their means are as one would expect from volunteer bias. This suggests the model is also able to correct for volunteer bias for variables not included in the selection model.

Our findings are relevant to the design of future biobanks. The effective sample size of the

UKB suggests it provides as much information as a representative data set of 118,370 to 202,999 respondents, or 24 to 41% of its actual size. Biobanks face a choice of whether to follow a volunteer-based sampling scheme that can be adjusted by providing well-estimated sampling weights, or to devote considerable resources to obtain a sample, possibly of smaller size, that is as close to representative as possible.

If future biobanks opt for volunteering, they should ensure that IP weight estimation is possible and provide these weights to their users. We recommend that during baseline data collection, biobanks collect all variables that might possibly relate to selection and that are available in an external data set that is representative of the population, such as a Census or administrative data. However, correcting for volunteer bias using IP weighting is no panacea as there could still be variables that are not measured and that are important to volunteering.

If future biobanks opt for representativeness, this may require increasing participation rates through methods such as telephone-based invitations rather than postal-based invitations,[25] or providing (monetary) incentives for study participation.[26] Another possible avenue might rely on volunteer-based sampling, combined with case prioritisation to ensure that the types of individuals that are unlikely to respond are prioritized.[27]

References

1 Wood AM, Kaptoge S, Butterworth AS, Willeit P, Warnakula S, Bolton T, et al. Risk thresholds for alcohol consumption: combined analysis of individual-participant data for 599 912 current drinkers in 83 prospective studies. Lancet. 2018;391(10129):1513-23.

2 Elliott LT, Sharp K, Alfaro-Almagro F, Shi S, Miller KL, Douaud G, et al. Genome-wide association studies of brain imaging phenotypes in UK Biobank. Nature. 2018;562(7726):210-6.

3 Beesley LJ, Salvatore M, Fritsche LG, Pandit A, Rao A, Brummett C, et al. The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. Statistics in medicine. 2020;39(6):773-800.

4 Douaud G, Lee S, Alfaro-Almagro F, Arthofer C, Wang C, McCarthy P, et al. SARS-CoV-2 is associated with changes in brain structure in UK Biobank. Nature. 2022;604(7907):697-707.

5 Publications [Internet]. [place unknown: publisher unknown]; June 7 2023. Available from: https://www.ukbiobank.ac.uk/enable-your-research/publications.

6 Duncan LE, Ostacher M, Ballon J. How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. Neuropsychopharmacology. 2019;44:1518-23.

7 Swanson JM. The UK Biobank and selection bias. Lancet. 2012;380:110.

8 Allen N, Sudlow C, Downey P, Peakman T, Danesh J, Elliott P, et al. UK Biobank: Current status and what it means for epidemiology. Health Policy and Technology. 2012;1(3):123-6.

9 Banks E, Herbert N, Mather T, Rogers K, Jorm L. Characteristics of Australian cohort study participants who do and do not take up an additional invitation to join a long-term biobank: The 45 and Up Study. BMC research notes. 2012;5:1-6.

10 Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. American journal of epidemiology. 2017;186(9):1026-34.

11 Klijs B, Scholtens S, Mandemakers JJ, Snieder H, Stolk RP, Smidt N. Representativeness of the LifeLines cohort study. PloS one. 2015;10:e0137203.

12 Rothman KJ, Gallacher JE, Hatch EE. Why representativeness should be avoided. International journal of epidemiology. 2013;42:1012-4.

13 Elwood JM. Commentary: on representativeness. International journal of epidemiology. 2013;42:1014-5.

14 Manolio TA, Collins R. Enhancing the feasibility of large cohort studies. JAMA. 2010;304:2290-1.

15 Collins R. What makes UK Biobank special? Lancet. 2012;379(9822):1173-4.

16 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.

17 Ebrahim S, Davey Smith G. Commentary: Should we always deliberately be non-representative? International journal of epidemiology. 2013;42:1022-6.

18 Solon G, Haider SJ, Wooldridge JM. What are we weighting for? Journal of Human resources. 2015;50:301-16.

19 Munafò MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. International journal of epidemiology. 2018;47:226-35.

20 Keyes KM, Westreich D. UK Biobank, big data, and the consequences of non-representativeness. Lancet. 2019;393:1297.

21 Pirastu N, Cordioli M, Nandakumar P, Mignogna G, Abdellaoui A, Hollis B, et al. Genetic analyses identify widespread sex-differential participation bias. Nature Genetics. 2021;53(5):663-71.

22 2011 Census England and Wales General Report; 2011.

23 Stamatakis E, Owen KB, Shepherd L, Drayton B, Hamer M, Bauman AE. Is cohort representativeness Passé? Poststratified associations of lifestyle risk factors with mortality in the UK Biobank. Epidemiology. 2021;32:179.

24 Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. American journal of epidemiology. 2008;168:656-64.

25 Sinclair M, O'Toole J, Malawaraarachchi M, Leder K. Comparison of response rates and cost-effectiveness for a community-based survey: postal, internet and telephone modes with generic or personalised recruitment approaches. BMC medical research methodology. 2012;12:1-8.

26 Smith MG, Witte M, Rocha S, Basner M. Effectiveness of incentives and follow-up on increasing survey response rates and participation in field studies. BMC medical research methodology. 2019;19:1-13.

27 West BT, Chang W, Zmich A. An Experimental Evaluation of Alternative Methods for Case Prioritization in Responsive Survey Design. Journal of Survey Statistics and Methodology. 2021.

28 Hastie T, Qian J. Glmnet vignette; 2014. http://www.web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf.

4 Methods

4.1 Data

UKB: Between 2006 and 2010, the UKB sent invites to ~9.2 million UK citizens aged 40 to 69 living in proximity to one of 22 assessment centres.[3] Roughly 500,000 participated, a sampling rate of 5.5%. UKB respondents are older, more likely to be female, and reside in less socioeconomically deprived areas, compared to the UKB's sampling population.[4,5] We dropped UKB respondents who lived outside any of the 22 assessment-centre sampling ranges or whose age at baseline fell outside of the ages that were sampled (see Supplementary Note S1, Supplementary Figure S1), who had missing values for sex, region, ethnicity, and/or year of birth, who resided in Census grouped local authority (GLA) districts with fewer than 70 UKB participants, or who died before Census day (March 27^{th} 2011). This resulted in a small loss of 11,237 UKB respondents (2.2%). The final UKB data set includes 491,268 UKB respondents. We study selection into this main sample.

UKB-eligible Census: The 2011 Census Microdata Individual Safeguarded Sample (Local Authority) for England and Wales,[6] and Scotland,[7] is a random 5% subsample of the 2011 UK Census ($N \approx 3.1$ million). We restricted the Census data to observations that would have been eligible for sampling into the UKB during its data collection period from 2006 to 2010 (the *UKB-eligible Census*), by restricting the UK Census microdata according to respondents' birth cohort and region of residence, as measured by 285 distinct grouped local authority (GLA) regions.

First, we only kept individuals aged 40 to 74 at the time of the Census, as these could have been aged between 40 and 69 between 2006 and 2010. We used the age of the individual on the day of the Census (5-year bins) to infer the year of birth bin of each individual. To ensure full overlap between our Census subsample and the UKB-eligible population, we restricted this age range further using adjustment factors (see Supplementary Note S2).

Second, we determined the regions of residence the UKB included in its sampling population, as follows. Location of residence in the UK Census data is reported at a higher level of aggregation

than in the UKB, namely by *grouped local authority* (GLA) *region* (grouped council authority; GCA, for Scotland, named GLA here for convenience). These regions consist of a single local authority when the population in these regions was larger than 120,000, and of aggregated groups of neighbouring local authorities otherwise. There are 285 distinct GLA regions covering all of Great Britain. We then restricted the UK Census data to only include those Census individuals that resided in a GLA region that falls *at least partially* within the sampling radius of one of the 22 assessment centres in England, Wales, or Scotland (see Figure S1 and Supplementary Note S1). We then use adjustment factors that account for the share of the area that falls within the UKB sampling regions, i.e. around the assessment centres, as well as the share of the population that is in the UKB-eligible age range (see Supplementary Note S2).

The final sample size of this "UKB-eligible Census" is 687,491.

Weighted UKB: The Weighted UKB is the UKB to which inverse probability (IP) weights, constructed as discussed below, are applied to correct for volunteer bias.

4.2 Inverse probability weighting

To obtain IP weights for UKB respondents, we estimated a probit model that predicts the UKB participation decision on concatenated data from the UKB (UKB=1) and the UKB-eligible Census (UKB=0). We used predictors based on year of birth, sex, ethnicity, educational attainment, employment status, region of residence, tenure of dwelling, number of cars in the household, self-reported health, and whether the individual lived in a one-person household. These variables were selected based on two inclusion criteria. First, they had to be assessed for all UKB baseline respondents and UK Census respondents. Second, they had to be assessed using the same (or very similar) wording in their respective questionnaires. We harmonised all responses into categories that are comparable in both data sets (Supplementary Note S3, and Supplementary Table S1 for a full summary of all variables in both data sets).

We used exact matching to impute missing variables (see Supplementary Note S4). All selected

variables were either binary or categorical. We entered them non-parametrically in the model by creating a dummy variable for each level the variable takes and included all possible two-way interactions between these dummy variables. As a result, our probit model used 4,820 predictors in its estimation.

LASSO estimation of UKB participation probabilities

We performed variable selection by LASSO estimation in *glmnet*.[28] Specifically, we model the likelihood of participating in the UKB for individual i, conditional on having received an invitation, $Pr(UKB = 1|Z_i')$, as

$$Pr(UKB = 1|Z_i') = \Phi(\alpha + Z_i'\delta + \nu_i), \tag{1}$$

with $\Phi(.)$ the standard normal cumulative distribution function (CDF), α a constant, ν_i a random error term, and Z_i' a vector of variables that influences one's individual propensity for participating in the UKB. Variables included in Z_i' are sex, year of birth (5-year cohort), education level, ethnicity, region of residence (Census GLA), tenure of dwelling, employment status, number of cars in the household, a dummy indicating whether the person lives in a single-person household, and self-reported health. These variables were included in a non-parametric manner (i.e., we used dummy variables for each category of the categorical variables under consideration). Furthermore, we included all possible two-way interactions between these dummy variables as predictors. In total, Z_i' contains 4,820 variables.

We estimated equation 1 by weighted probit regression on the training sample of stacked UKB and UKB-eligible UK Census data, where we assigned the outcome variable $UKB = 1$ to each UKB observation, and $UKB = 0$ to each UKB-eligible Census observation. Before we estimated our model, we divided our data set in folds that each hold 20% of the data. We then repeated the estimation procedure 5 times. At each step, one of these folds was not used in training the model, and the probability of selection was estimated in this holdout fold. As such, we made sure that the model used in constructing the weights was not overfitted on the data. To further prevent overfitting, we estimated the model using a LASSO variable selection procedure.[8] The LASSO

model maximises the log-likelihood function of the regular probit, subject to the absolute value of the sum of the coefficients being smaller than a certain constant (as determined by a penalisation parameter λ). This additional constraint in the optimisation problem prevents overfitting of the data when many regressors are included, as it ensures that coefficients of variables that are insufficiently predictive of selection are being shrunk to zero. We estimated our LASSO probit model using *glmnet*,[28] which solves the optimisation problem

$$(\hat{\alpha}, \hat{\delta}) = \arg\min\Bigg\{ \sum_{i=1}^{N} w_i \Bigg(\mathrm{UKB}_i \ln\big(\Phi(\alpha + \sum_j \delta_j z_{ij})\big) + (1 - \mathrm{UKB}_i) \ln\big(1 - \Phi(\alpha + \sum_j \delta_j z_{ij})\big) \Bigg) + \lambda \sum_j |\delta_j| \Bigg\},$$

where w_i is the weight (1 for UKB observations, and the adjustment factor constructed as described in section S2 for UK Census observations), and λ is the penalisation parameter.

The penalisation parameter was chosen through cross-validation using k-folding with 5 folds. The k-folding procedure ensures that λ is chosen as to yield an optimally predictive model on a holdout sample not used in the estimation of the LASSO model. For the model used on the first holdout sample, this results in a penalisation parameter of 0.000026. 568 out of the 4,680 variables we include had their coefficients shrunk to zero. This low penalisation parameter implies that the solutions to our model lie very close to those of a similarly specified regular probit model in which the same variables (including the two-way interactions) are included. These statistics were very similar for the models that held out the other 4 holdout samples of our data.

Variable importance of predictors included in the model for UKB participation

We constructed measures of variable importance as follows. For each variable in turn, we permuted each of their categories (including their interactions) in the holdout sample. We then predicted the participation probabilities from the LASSO probit model on this permuted sample. Through each permutation, the respective variable becomes unrelated to the model's outcome. Hence, the decrease in the AUC on the permuted holdout sample, relative to the original holdout sample, can

be taken as a measure of the predictive power of the permuted variable: the larger the reduction in AUC after permuting the variable, the more important that variable is to the model's performance.

Construction of Inverse Probability Weights

Inverse probability weighting (IPW) is a method to correct for volunteer bias in observational data.[10,11,24] We constructed inverse probability weights as

$$\mathrm{IPW}_i = \frac{\widehat{\Pr(\mathrm{UKB} = 1)}}{\widehat{\Pr(\mathrm{UKB} = 1|Z_i')}} \tag{2}$$

.

where $\widehat{\Pr(\mathrm{UKB} = 1)}$ is the average probability of being sampled in the UKB as estimated on the full weighted stacked UKB and UKB-eligible Census, and $\widehat{\Pr(\mathrm{UKB} = 1|Z_i')}$ is the probability of UKB participation for UKB participant i as predicted by the LASSO probit model.

A known issue with the estimation of inverse probability weights using a rich set of predictors is that such rich models may yield some values of $\widehat{\Pr(\mathrm{UKB} = 1_i|Z_i')}$ that are very close to zero, with excessively large values of IPW_i as a result. Such excessive weights can result in noisy estimates of weighted regression coefficients and hence dilute power.[13] We dealt with this issue by winsorising our distribution of estimated weights, setting any values of IPW_i lower than the 1st percentile equal to the value at the first percentile, and any values of IPW_i higher than the 99th percentile equal to the value at the 99th percentile. Supplementary Figure S2 visualises the distribution of these weights.

4.3 Unweighted and weighted means, standard deviations, and associations in the UKB

To assess whether volunteer bias affects the UKB, we compared, in the UKB and UKB-eligible Census, means, standard deviations, and regression coefficients as estimated by bivariate linear models. We tested the null hypothesis that these coefficients are the same using the following

Z-statistic: $\frac{\hat{\beta}_{\text{UK Census}} - \hat{\beta}_{\text{UKB}}}{\sqrt{\hat{se}^2_{\text{UK Census}} - \hat{se}^2_{\text{UKB}}}}$.

Next, we used our IP weights (IPWs) to correct these various statistics in the UKB for volunteer bias. Weighted means were constructed as $\bar{x} = (\sum IPW_i \cdot x_i)/(\sum IPW_i)$ and their 95% confidence interval as $CI = \bar{x} \pm \frac{\bar{sd}}{\sqrt{N}} \times z_{0.95/2}$. Weighted standard deviations were constructed as $\bar{sd} = \sqrt{\left[\sum IPW_i \cdot (x - \bar{x})^2\right] / [\sum IPW_i]}$. IP weighted regressions were estimated by weighted least squares.

To obtain a measure of the bias reduction achieved through IP weighting across all associations, we first defined absolute bias as the absolute difference between the point estimates estimated in the UKB-eligible Census, the UKB and the Weighted UKB. Next, we took the mean absolute bias of unweighted UKB estimates, minus the mean absolute bias of the weighted UKB estimates, and divided it by the mean absolute bias of unweighted UKB estimates.

4.4 UKB effective sample size

Our IP-weighting procedure is necessary to reduce volunteer bias in estimated means, variances, and associations in the UKB, but it increases confidence intervals, and therefore reduces power. The effective sample size reflects the power of the weighted UKB as equivalent to a hypothetical random sample of size $\hat{n}$. We used two methods to estimate $\hat{n}$. Both give similar results.

The first method estimates $\hat{n}$ using the distribution of the IP weights. This measure of effective sample size reflects the weighted sample's ability to uncover the sampling population's mean and variance.[14]

$$\hat{n} = \sum_{i=1}^{n} \lambda IPW_i,\ \lambda = \frac{\sum_{i=1}^{n} IPW_i}{\sum_{i=1}^{n} IPW_i^2} \tag{3}$$

The attractiveness of this measure is that it is not dependent on any particular regression model. However, it is designed to summarise the amount of information that the data reveals for estimation of means and variances only, and not for regression coefficients. Hence, we next used an additional regression-based measure of the equivalent sample size.

A regression-based equation for $\hat{n}$ can be obtained by rewriting the regular OLS formula for the coefficient's standard errors.[15] Because weighted regression results in wider standard errors than OLS, this measure of $\hat{n}$ represents the size of a representative sample that one would need in order to identify the same coefficient with the same amount of precision using OLS. Hence, we measured effective sample size using the following formula:

$$\hat{n} = \frac{var(\epsilon)}{se(\hat{\beta})^2 * var(x)} \tag{4}$$

This measure of $\hat{n}$ is regression-specific (i.e., it depends on which variables are included), as this influences the value of $se(\hat{\beta})$, $var(\epsilon)$, and $var(x)$. Our models with binary outcomes are not well-suited to calibrate an effective sample size, since the hypothetical case of homoskedastic errors (on which this equation for $\hat{n}$ relies) does not hold for linear models with binary outcomes. Thus, we estimated $\hat{n}$ only on models with discrete outcomes of three levels or more, or models with continuous outcomes (i.e., all models in Figure 5).

Some caveats are in order. First, formula 4 assumes that the error term in the model is homoskedastic, and that any heteroskedasticity in the error term is introduced because of weighting the data, and not because of characteristics of the true underlying data generating process. It could be that some of the models included in Figure 5 exhibit heteroskedastic errors, even in representative data. However, heteroskedasticity does not seem to drive the sharp reduction in effective sample size that we observe here. When applying the formula for $\hat{n}$ on the *unweighted* UKB coefficients, it matters little whether we use robust ($\bar{\hat{n}}$=484,468) or non-robust ($\bar{\hat{n}}$ =487,304) standard errors. In both cases, the formula produces estimates that are close to the true sample size of the unweighted UKB (491,268).

Another caveat is that our estimate for $\hat{n}$ can only be interpreted as the sample size in equally informative representative data if our weights are able to capture *all* of the volunteer bias. This caveat pertains to the first method of estimating $\hat{n}$ as well. As we discuss in the Discussion, our weights may suffer from omitted variables that also influence UKB volunteering. As a result, some volunteer bias may remain even for association estimates that use our IP weights. Because of this

potential of missing variables, a representative sample of size $\hat{n}$ is likely desirable to the weighted UKB. In other words, missing variables in the construction of our weights result in an overestimate of $\hat{n}$: the true effective sample size of the UKB is likely lower.

Data & Code availability

UK Biobank data is accessible upon request and approval by the UK Biobank committee (`https://www.ukbiobank.ac.uk/`). UK Census safeguarded microdata is available from UK Data Service (`https://ukdataservice.ac.uk/`), upon request and approval. All code used for generating the results is available at `https://github.com/sjoerdvanalten/UKBWeightsFinal`. The IP weights developed here have been returned to the UKB and are available as a data field to UKB-approved researchers.

Acknowledgements

Research reported in this publication was supported by the National Institute On Aging of the National Institutes of Health (RF1055654, R56AG058726 and R01AG078522), the Dutch National Science Foundation (016.VIDI.185.044), and the Jacobs Foundation. This research has been conducted using the UK Biobank Resource under Application Number 55154. We thank participants at the 2021 and 2022 BGA annual meetings, 2021 and 2022 ASHG conferences, the 2022 European Social Science Genomics Network Conference, the 2021 Integrating Genetics and Social Science Conference, and the 2022 European Health Economics Association PhD Conference for their feedback and comments.

Author Contributions

SA was responsible for all data analysis. ATM checked the coding and data analysis process. SA was responsible for the first draft of the manuscript. SA, BWD, JF, TJG, and ATM were jointly responsible for designing the study, drafting the final manuscript, and revising its contents.

References

1 Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics. 2011;12:1-8.

2 Domingue B, Rahal C, Faul J, Freese J, Kanopka K, Rigos A, et al. InterModel Vigorish (IMV): A novel approach for quantifying predictive accuracy with binary outcomes. SocArXiv. 2021.

3 Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, 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 medicine. 2015;12(3):e1001779.

4 Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. American journal of epidemiology. 2017;186(9):1026-34.

5 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.

6 2011 Census England and Wales General Report; 2011.

7 National records of Scotland. Scotland's Census 2011 General Report. 2015.

8 Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58:267-88.

9 Hastie T, Qian J. Glmnet vignette; 2014. http://www.web.stanford.edu/~hastie/Papers/Glmnet_Vignette.pdf.

10 Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000.

11 Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004:615-25.

12 Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. American journal of epidemiology. 2008;168:656-64.

13 Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. American journal of epidemiology. 2008;168:656-64.

14 Potthoff RF, Woodbury MA, Manton KG. "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. Journal of the American Statistical Association. 1992;87:383-96.

15 Howe LJ, Nivard MG, Morris TT, Hansen AF, Rasheed H, Cho Y, et al. Within-sibship GWAS improve estimates of direct genetic effects. bioRxiv. 2021.

Supplementary Material to: Should representativeness be avoided? Reweighting the UK Biobank corrects for pervasive selection bias due to volunteering

Contents

S1 UKB: Geographic sampling and restrictions to the main sample

We removed individuals in the UKB who were sampled by the UKB, but nonetheless did not meet the criteria to be eligible for the UKB-eligible population (those eligible to receive an invite). First, we dropped all UKB participants who were not aged between 40 and 69 at the start of their assessment centre's sampling period. That is, for assessment centres that started sampling in 2007, we kept all individuals born between 1937 and 1967, for assessment centres that started sampling in 2008, we kept all individuals born between 1938 and 1968, etc. However, we kept individuals born in 1969 assessed at the Bristol centre (which started assessment in 2008), and kept individuals born in 1970 assessed at the Birmingham centre (which started assessment in 2009), as these centres both sampled a significant number of individuals born in these years. This resulted in dropping 98 UKB respondents.

The UKB sampled respondents residing close enough around one of the 22 assessment centres. Figure S1 shows the place of residency of all UKB participants on the day they visited the assessment centre. Some have claimed that UKB assessment centres sent out invites to all individuals in the targeted age range that were living within a radius of 40 km of the assessment centre.[1] Therefore, we dropped 87 UKB participants who lived further than 40 km away from any assessment centre. However, a closer inspection of the data revealed that, for virtually all assessment centres, the sampling radius is considerably smaller, and the size of the sampling area varies per assessment centre. For example, Figure S1 shows that the assessment centre in Edinburgh only sampled respondents in relative proximity (max 22.4 km), whereas the assessment centre in Middlesbrough sent out invitations to a wider area (max 39.9 km).

We obtained the sampling radius of each assessment centre from the UKB data as follows: for each assessment centre, we assumed that the UKB participant who lived furthest away was also the furthest living person to receive an invite for that centre, and we defined the assessment centre's

sampling radius accordingly.[1] However, this method is sensitive to outliers. For 14 assessment centres, we obtained sampling radii that were unrealistically large, as the difference between the furthest participant and the participant in the 99.9th percentile of the distance-to-assessment centre distribution was more than 2 km. For these centres, we defined the sampling radius as the 99.7th percentile of distances to the assessment centre, plus 2 km. The resulting sampling radii are visualised as the circles in Figure S1. A small number of UKB respondents (108) fell outside the sampling radii of any assessment centre, and were excluded from our data set. Last, to ensure that our UKB participation probabilities were robustly estimated, we dropped UKB participants residing in Census grouped local authority (GLA) districts with fewer than 70 UKB participants. This resulted in a small loss of another 263 UKB respondents.

S2 Adjustment factors to restrict and adjust UK Census data to the UKB-eligible subsample

Restricting the UK Census based on age and region as described in Methods introduces some individuals in our sample that may not have received a UKB invite during the period 2006 to 2010, because they were not in the relevant age range (40 to 69) at the time that their nearest assessment centre started sampling, or because they lived too far from a UKB assessment centre. We therefore calculated adjustment factors to ensure that our UK Census sample is as representative as possible of the UKB-eligible population in 2006 and 2010, conditional on survival up until March 2011. To start, we assigned to each individual in our UK Census data set an initial adjustment factor of 20, to upweight the 5% random subsample to the full UK population. We next adjusted these adjustment factors in the following two ways.

First, individuals who resided in GLA regions that did not fully fall within the UKB sampling radius of any assessment centre have their adjustment factor multiplied by the proportion of the

[1]We reassigned individuals who lived within 40 km of an assessment centre, but visited an assessment centre further away, to their nearest assessment centre.

population living within this GLA region that is within the sampling radius (see Supplementary Note S1 for additional details on the estimation of the assessment-centre specific sampling radii). These GLA-specific sampling population proportions are calculated using 2011 UK Census population counts reported for the much less aggregated *lower layer super output areas* (LSOAs) for England and Wales. For Scotland, even less aggregated *output areas* (OAs) are used because LSOAs are unavailable. For England and Wales, there are 34,753 distinct LSOAs and for Scotland, there are 46,351 OAs. Figure S1 shows the GLA regions that are included in the final sample, and illustrates how adjustment factors are assigned to each Census respondent living in these regions (the adjustment factors range from adjustment factors of 0 [white] to 20 [black] and anything in between [grey]).

Second, the year of birth distribution of the UKB is assessment-centre specific, as some assessment centres started sampling sooner than others (Supplementary Note S1). For example, the assessment centre in Manchester sampled all its respondents in 2007, and hence only sampled respondents born between 1937 and 1967, as these were aged 40 to 69 at the time. By contrast, the centre in Swansea sampled all its respondents in 2010, and hence only sampled respondents born between 1940 and 1970. We faced the extra challenge that the UK Census only reports year of birth bins of 5 years. As a result, the year of birth distribution of our UKB-eligible UK Census sample and the UKB do not necessarily overlap, i.e., not everyone in the year of birth bins 1936-1940 or 1966-1970 (at the edges of the UKB-eligible distribution) was UKB-eligible. For Census individuals in these year of birth bins, we assigned them to an assessment centre based on the GLA in which they live. We multiplied their adjustment factor by the proportion of year of birth values (out of a maximum of five) in this year of birth bin that were sampled by the assessment centre. For example, Census respondents residing in or around Manchester that were born between 1936-1940 had their sampling weight multiplied by 0.8, since Manchester only sampled those with year of birth 1937, 1938, 1939, and 1940, but not 1936. When a GLA region overlapped with the sampling radius of multiple assessment centres, we took the largest possible adjustment factor.

The final UKB-eligible census subsample consists of 687,491 observations. Our adjustment

factors for this sample range between 0.22 and 20 with a mean of 15.8 and a median of 18.6. This implies a UKB-eligible population of 10,836,059 individuals (summing all sampling weights), slightly larger than the true number of invitations that were sent out by the UKB (9,238,452). Throughout this paper, all statistics reported on the *UKB-eligible population* (Table 1, Figure 4, Table S1, Table S3, Figure 5, Figure S3, Figure S4, Figure S5) are estimated on this subsample and weighted using the adjustment factors.

In Table S3, we compare the UKB-eligible subsample of the UK Census to the full UK population of this age range (i.e. all Census data). Compared to the full population, the UKB-eligible population is more ethnically diverse, younger, of lower socioeconomic status (as measured by an overall deprivation indicator), more urbanised (as measured by various proxies for urbanicity), and in worse health. These differences are relatively small compared to those induced by participation in the UKB.

S3 Harmonisation of variable categories across the UKB and UK Census

To estimate our IP weights, estimate our summary statistics on the UKB and the UKB-eligible Census, and estimate regression models in both these data sets, we selected variables that were assessed in a similar fashion in the UKB and the UK Census, allowing for comparisons. When needed, we altered the categorisation of several variables in the UKB and UK Census to ensure that the variables were comparable across both data sets.

S3.1 Year of Birth

- **UK Census:** Year of birth was derived from 5-year age bins that described the age of the individual at the day of the UK Census (40-44; 45-49; 50-54; 55-59; 60-64; 65-69; and 70-74). These were recoded into the following year of birth values: (1966-1970; 1961-1965;

1956-1960; 1951-1955; 1946-1950; 1941-1945; 1936-1940).[2]

- **UKB:** We recoded year of birth into the 5-year of birth bins as mentioned above.

S3.2 Sex

- **UK Census:** Respondents were asked "What is your sex?", and could answer male or female.

- **UKB:** Sex as recorded by the NHS registry, but possibly updated by the participant.

S3.3 Region of residence

- **UK Census:** Region of residence was inferred from the respondent's address. The respondent's address was prefilled by the Census data collectors, and corrected by the respondent when necessary. Region of residence is reported as 265 distinct grouped local authorities (GLA; England & Wales), and 20 grouped council areas (GCA; Scotland). GLAs or GCAs consist of local authorities and council areas, or groups of adjacent local authorities and council areas to ensure that each GLA/GCA has at least 120,000 inhabitants.

- **UKB:** Coordinates of home location at assessment, inferred from NHS registry data on the postcode level, rounded to the nearest kilometer. These coordinates were aggregated into the same GLAs and GCAs mentioned above, using .shp files that describe the borders of these GLAs and GCAs, obtained from the office for national statistics.

S3.4 Ethnicity

- **UK Census:** Respondents were asked "What is your ethnic group?", and could respond "White", "Mixed", "Asian/Asian Birtish", "Black/African/Caribbean/Black British", or "Other"

[2]The 2011 UK Census was conducted at 27th of March. We classify each respondent in their 5 year of birth bin assuming that they had *not yet* had their birthday in 2011. This approach inevitably results in some small classification error (e.g., someone who turned 65 on February 1st of 2011 has year of birth 1946, is classified in age bin 65-69 by the UK Census, and is next, erroneously, classified in year-of-birth-bin 1941-1945 by us).

- **UKB:** Wording of the question was the same as in the UK Census. The available categories were also the same, except that "Chinese" was recognised as a separate ethnic group. We merged this group with "Asian/Asian British". Additionally, respondents could answer "None of the above" or "Prefer not to answer". We coded these responses as missing.

S3.5 Economic status

- **UK Census:** UK Census respondents were asked what their main economic activity was last week through various questions. Answers were coded by the Census bureau into the following categories:

1 Economically Active (excluding Full-time students), In Employment, Employee, Part-time

2 Economically Active (excluding Full-time students), In Employment, Employee, Full-time

3 Economically Active (excluding Full-time students), In Employment, Self employed with employees, Part-time

4 Economically Active (excluding Full-time students), In Employment, Self employed with employees, Full-time

5 Economically Active (excluding Full-time students), In Employment, Self employed without employees, Part-time

6 Economically Active (excluding Full-time students), In Employment, Self employed without employees, Full-time

7 Economically Active (excluding Full-time students), Seeking work and ready to start within 2 weeks, and Waiting to start a job already obtained and available to start within 2 weeks

8 Economically Active Full-time students, In employment

9 Economically Active Full-time students, unemployed, seeking work and ready to start within 2 weeks, and waiting to start a job already obtained and available to start within 2 weeks

10 Economically Inactive, Retired

11 Economically Inactive, Student

12 Economically Inactive, Looking after home/family

13 Economically Inactive, Permanently sick/disabled

14 Economically Inactive, Other

We recoded these levels into a sparser number of categories, namely "Employed" (1-6, 8), "Retired" (10), "Stay-at-home" (12), "Incapacitated" (13), "Unemployed" (7), and "Student" (9, 11)

- **UKB:** At baseline, UKB respondents were asked in the touchscreen questionnaire: "Which of the following describes your current situation?". Participants could respond to this question as follows:

1 In paid employment or self-employed

2 Retired

3 Looking after home and/or family

4 Unable to work because of sickness or disability

5 Unemployed

6 Doing unpaid or voluntary work

7 Full or part-time student

-7 None of the above

-3 Prefer not to answer

We classified those doing unpaid or voluntary work as "Stay-at-home", and coded those answering "None of the above" or "Prefer not to answer" as missing.

S3.6 Tenure of household

- **UK Census:** Respondents answered the question: "Does your household own or rent this accommodation?" They could answer that they own it outright, own it with a mortgage or loan, part own and part rent (shared ownership), rent, or live rent-free.

- **UKB:** Respondents answered the question: "Do you own or rent the accommodation that you live in?" If respondents tapped the "Help" button they were shown the following:

  "Please select: - Own outright if you or someone in your household owns the accommodation that you live in. - Own with mortgage if you or someone in your household has a mortgage on the accommodation that you live in."

  Answer categories were the same as in the UK Census, except that respondents could also answer "None of the above", or "Prefer not to answer". We coded these as missing.

S3.7 Number of vehicles in the household

- **UK Census:** Respondents answered the question: "In total, how many cars or vans are owned, or available for use, by members of this household?" (Please include company vehicles if available for private use). They could answer none, 1, 2, 3, or 4 or more.

- **UKB:** The wording of this question was identical to the wording in the UK Census, as were the answering categories. In addition, we coded those answering "None of the above" or "Prefer not to answer" as missing.

S3.8 Household size

- **UK Census:** The only variable regarding household size available to us in the UK Census data was a dummy that indicated whether the person resided in a one-person household or not.

- **UKB:** Respondents answered the question;"Including yourself, how many people are living together in your household? (Include those who usually live in the house such as students living away from home during term, partners in the armed forces or professions such as pilots". We recoded the answers into a dummy variable that described whether the answer was 1, or some higher number. In addition, we coded those answering "None of the above" or "Prefer not to answer" as missing.

S3.9 Self-reported health

In the UK Census, respondents answered the question "How is your health in general?". In the UKB, respondents answered the question "In general how would you rate your overall health?". Self-reported health in the UKB was assessed through a 4-level Likert scale (Poor, Fair, Good, Excellent), whereas self-reported health in the 2011 UK Census was assessed through a 5-level scale (Very Bad, Bad, Fair, Good, Very Good). To minimise classification error, we harmonise these values in both data sets to a three-level scale (Bad, Fair, Good/Excellent), by combining the categories "Very Bad" and "Bad" in the UK Census, and lumping Good/Excellent/Very good into a single category. In addition, respondents in the UKB could answer "Do not know" or "Prefer not to answer", these answers were coded as missing.

S3.10 Education

In the UK Census, respondents were asked "Which qualifications do you have?" and were instructed to tick every box that applied. These answers were recoded by the Census bureau into the highest level of education obtained for each respondent. In the UKB, respondents were asked the question "Which of the following qualifications do you have? (You can select more than one)". However, the potential answers to each question differed between the UK Census and the UKB. We harmonised the level of education across both data sets as follows:

- **UK Census:** The UK Census assigned International Standard Classification of Education (ISCED) levels to the variable "highest degree obtained". We assigned years of education based on these levels: 7 for no degree (assuming primary school), 10 for a level 1 or level 2 degree, 13 for a level 3 degree, and 20 for a level 4+ degree. This assignment follows previous work in the UKB.[2] Two other categories in the UK Census were not assigned any ISCED level in the UK Census data. These were "apprenticeship" and "Other: Vocational/Work-related qualifications, etc.". We assigned 12 years to the "apprenticeship" (reflecting the fact that it requires continuing one's education after a GCSE-degree, but does not require an A/AS-level

degree).[3] For the "vocational" category, we again followed previous work,[2] and assigned 15 years of education.

- **UKB:** For the UKB, we similarly assigned years of education to degree categories that clearly fall within the categories recognised by the UK Census. These categories are: "No degree" (ISCED1), "College or university" (ISCED4+), "A levels/AS levels or equivalent" (ISCED3), "O levels/GCSEs or equivalent" (ISCED2) and "CSEs or equivalent" (ISCED2). However, for those who reported having a "National Vocational Qualification (NVQ) or Higher National Diploma (HND) or Higher National Certificate (HNC) or equivalent" or "Other professional qualifications, e.g. nursing, teaching", assigning years of education was not as straightforward. For those holding an NVQ, we do not know which level of NVQ certificate they hold (In the UK Census, an NVQ of level 1 is considered an ISCED1 degree, whereas an NVQ of level 4 or higher is considered ISCED4+). Accordingly, substantial heterogeneity in the variable "age at which left full-time education" can be seen for the group of respondents holding a degree in this category, with a substantial group of respondents having left full-time education before the age of 16 (Figure S6a). To solve this issue, for those holding an NVQ (or HND or HNC) we assigned years of education by taking the age at which the respondent reported to have left full-time education, minus 5, and capped the value at 19 years of education.[4] For those holding a professional degree, we similarly observed substantial heterogeneity in the age at which these respondents left full-time education (Figure S6b). Hence, for this group, we estimated years of education in similar fashion, but capped the variable at 15 years. When UKB respondents reported multiple degrees, we took the maximum of years of education associated with each degree.

These continuous years of education measures make educational attainment comparable across the UKB and the UK Census. For estimating the selection model, we discretized the variable. Those less than 8.5 years of education get level 1 (no degree), those between 8.5 and 11 get level 2 (O-levels or equivalent), those between 11 and 17.5 level 3 (A-levels, vocational, or equivalent), and those above 17.5 get level 4 (college/university, or equivalent).

S4 Missing data in the UK Census and UKB, and imputation procedure

Both the UK Census and UKB had missing data on the variables we use to predict UKB participation. Table S4 provides an overview for each variable we use. In the UKB, respondents explicitly had the option to not share information on all variables measured through self-reporting. They could either tick the options "prefer not to answer" or "do not know". We coded such values as missing. As a result, 4.9% of our included UKB respondents had missing data on at least one predictor included in the selection model. In the UK Census, data on the variables we use is typically not missing, but for some variables regarding the household in which the individuals live (i.e., tenure of dwelling, number of cars owned, and household size), no information is available for 0.63% of the observations, as these were people living in communal establishments.

Our model uses a large number of regressors to predict UKB selection status. 24,380 UKB and 4,355 UK Census observations had at least one regressor missing. These missing values are imputed using an exact matching procedure. We conduct exact matching by converting the data to a frame that holds the following variables: whether it was a UK Census or UKB data point, region, year of birth, sex, education, self-reported health, employment status, and sex. In step one, we fill in missing values by sampling from observations with the exact same values on all these variables. For observations for which an exact match could not be found, we attempted to match again using the same variables, but dropping region and ethnicity from the data frame. For 60 observations, this procedure did not yield an exact match, such that IP weights could not be estimated.

S5 Effects of sample selection bias on the standard deviation of bivariate variables

Note that for bivariate variables, it is not possible to tell whether an increase or a decrease in the standard deviation in the UKB, vis à vis the standard deviation in the UKB-eligible population, is

consistent with selective sampling. This is because, for bivariate variables, the standard deviation is $\sqrt{p(1-p)}$, with p the mean of the variable. Hence, the standard deviation is largest for $p = 0.5$. When selection into the UKB is such that the mean of the variable becomes closer to 0.5 (from above or from below), the standard deviation will be larger in the UKB than in the UKB-eligible population. For example, this is the case in Table 1 for "University or equivalent": in the UKB-eligible population, 27.8% holds such a degree, whereas in the UKB, this is 33.6%. This change in the mean of the variable is consistent with selective sampling in the UKB (where healthy and higher educated citizens are more likely to participate in scientific studies), but nonetheless results in a larger standard deviation of this variable in the UKB.

S6 Addressing selection bias by adding control variables in the regression

Typically, researchers do not explore outcome-exposure relationships through bivariate models, but include control variables to adjust for possible confounding factors. However, introducing control variables that are also correlated with participation into the data set can exacerbate rather than mitigate bias, as these variables are potential colliders.[5] We illustrate this within our models. First, we re-estimate all the models in Figure 4, including sex and year of birth (linearly, measured as a five-year birth bin) as control variables (the models in Figure 4 that had the variable "born before 1950" or "female" as the dependent or independent variable were now excluded). The new models are shown in Figure S3. We estimate the average volunteer bias after introducing these linear controls, and compare it to the average volunteer bias of the same models in Figure 4. By introducing these controls, volunteer bias *increases* on average by 20%. By contrast, IP weighting these models still performs well and reduces volunteer bias by 88% on average.

Further, we re-estimate the model that regresses the probability of reporting poor health on being born before 1950, after including a much wider range of possible control variables. In Figure S4 we report these coefficients without controls (as in Figure 4) and after controlling for sex, years

of education, number of cars owned, a single household indicator, tenure of household (4 dummy variables), employment status (5 dummy variables), ethnicity (4 dummy variables), and region (142 dummy variables). As can be seen from the figure, adding these richer controls does *not* succeed in changing the negative association of age on reporting poor health towards a positive one in the UKB. By contrast, IP weighting does succeed in flipping the sign, and getting the point estimate closer to the one obtained in the UKB-eligible population, both in the model with and without additional control variables.

S7 Robustness of IP weighted regressions to missing variables

The models that we show in Figure 2 are based on variables that we also include in our LASSO model that underlies the estimation of the IP weights. Hence, our approach potentially overstates the extent to which these IP weights can be expected to mitigate volunteer bias across various types of association models: missing variable bias may be introduced when one tries to use these weights to weight models based on variables that were not included in the IP weighting procedure, reducing the potency of the IP weights.

To confirm that our weights are robust to such missing variable bias, we re-estimate the IP weights for each of the models presented in Figure 2. For each association model, we re-estimate our LASSO model, but leaving out *both* the outcome and the input variable. Using this approach, we create "leave-variables-out" IP weights to weight each linear association model that we consider. As such, we can assess the performance of our IP weighting procedure when relevant model variables are not included in the weighting scheme. We only perform this sensitivity analysis on the first fold of our data (i.e., a 20% subsample), to reduce the computational burden of re-estimating our LASSO model 21 times.

Figure S5 shows the same results as in Figure 2, but now also includes weighted UKB estimates that use the newly created leave-variables-out weights. In general, the point estimates that are estimated using the leave-variables-out weights (red open circles) are very similar to those that are

based on the weights with all variables (green open circles). We confirm that, even when using the leave-variable-out weights, our IP-weighting procedure reduces volunteer bias as averaged over all models included in the figure by 69%.

The finding that weights estimated on different sets of variables result in very similar point estimates, and also result in substantial bias reduction, is encouraging. This means that our IP-weights are robust to missing variable bias, but are also robust to other forms of bias that may arise, for example, when there are subtle differences in assessment of variables between UKB and UK Census.

S8 Supplementary Tables & Figures

Table S1: Full summary statistics for the UKB-eligible Census and the UKB. P_{χ^2} in the 7th column reports the p-value of a χ^2-test for equal distributions for each variable as measured in the UKB-eligible Census and the UKB.

		UKB-eligible Census		UKB		
Variable	**Levels**	**N**	**%**	**N**	**%**	P_{χ^2}
Sex	Male	5 333 695.06	49.22	222 971	45.39	$< 10^{-8}$
	Female	5 502 364.27	50.78	268 297	54.61	
	Total	10 836 059.32	100.00	491 268	100.00	
Birthyear	1936-1940	674 012.26	6.22	33 586	6.84	$< 10^{-8}$
	1941-1945	1 289 775.68	11.90	101 081	20.58	
	1946-1950	1 671 907.91	15.43	114 996	23.41	
	1951-1955	1 624 169.90	14.99	82 433	16.78	
	1956-1960	1 874 898.22	17.30	70 732	14.40	
	1961-1965	2 151 457.40	19.85	61 117	12.44	
	1966-1970	1 549 837.96	14.30	27 323	5.56	
	Total	10 836 059.32	100.00	491 268	100.00	
Education	None	2 920 933.24	26.96	83 553	17.01	$< 10^{-8}$
	Lower secondary	2 832 405.03	26.14	141 037	28.71	
	A-levels/vocational	2 069 303.74	19.10	94,150	19.16	
	University	3 013 416.20	27.81	161 511	32.88	
	Total	10 836 058.20	100.00	480 251	100.00	
Ethnicity	White	9 625 885.91	88.83	464 757	94.60	$< 10^{-8}$
	Mixed	105 552.61	0.97	2 891	0.59	
	Asian/Asian British	656 703.51	6.06	11 273	2.29	
	Black/Black British	356 016.68	3.29	7 879	1.60	
	Other	91 900.62	0.85	4 468	0.91	
	Total	10 836 059.32	100.00	491 268	100.00	
Health (self-reported)	Bad	1 003 746.72	9.26	21 711	4.42	$< 10^{-8}$
	Fair	2 054 220.11	18.96	102 875	20.94	
	Good/Very Good	7 778 091.38	71.78	364 370	74.17	
	Total	10 836 058.20	100.00	488 956	100.00	
Employment status	Paid employment	6 602 507.40	60.93	281 812	57.36	$< 10^{-8}$
	Retired	2 693 524.53	24.86	165 944	33.78	
	Stay-at-home	385 636.64	3.56	13 607	2.77	
	Incapacitated	751 258.09	6.93	16 066	3.27	
	Unemployed	359 497.74	3.32	7 998	1.63	
	Student	43 634.92	0.40	1 284	0.26	
	Total	10 836 059.32	100.00	486 711	100.00	
No. of vehicles in household	0	1 992 849.34	18.50	42 717	8.70	$< 10^{-8}$

	1	4 350 708.42	40.40	204 076	41.54	
	2	3 243 366.45	30.12	185 911	37.84	
	3	881 116.19	8.18	41 968	8.54	
	4 or more	301 482.86	2.80	13 160	2.68	
	Total	10 769 523.25	100.00	487 832	100.00	
Tenure of dwelling	Owns house (no mortgage)	3 812 493.95	35.40	254 890	51.88	$< 10^{-8}$
	Owns house w/ mortage	4 109 126.28	38.16	180 172	36.67	
	Shared ownership	49 412.74	0.46	1 428	0.29	
	Rent	2 707 334.15	25.14	44 185	8.99	
	Rent-free	91 156.12	0.85	3 482	0.71	
	Total	10 769 523.25	100.00	484 157	100.00	
One-person household	No	8 822 644.06	81.92	397 792	80.97	$< 10^{-8}$
	Yes	1 946 879.19	18.08	90 130	18.35	
	Total	10 769 523.25	100.00	487 922	100.00	

Table S2: Mean of various variables in the UKB before and after IP weighting: These variables are available in UKB data, but not in the UK Census. The weighted mean gives an indication of what the mean of each variable looks like in the UKB's sampling population. 95% confidence intervals around each mean included

Variable	Mean [95% CI]	Weighted Mean [95% CI]	% Change
Anthropometric			
Height	**168.713** [168.684;168.741]	**169.123** [169.094; 169.153]	0.2%
BMI	**27.415** [27.4;27.429]	**27.67** [27.654; 27.685]	0.9%
Waist Circumference	**90.319** [90.277;90.361]	**90.916** [90.873; 90.959]	0.7%
Waist Hip Ratio	**0.871** [0.871;0.872]	**0.876** [0.875; 0.876]	0.5%
Hip Circumference	**103.451** [103.423;103.479]	**103.614** [103.584; 103.644]	0.2%
Demographic			
Age at Recruitment	**56.54** [56.518;56.563]	**53.517** [53.493; 53.541]	5.3%
Urban	**0.862** [0.861;0.863]	**0.875** [0.874; 0.876]	1.5%
Died after Census Day	**0.035** [0.034;0.035]	**0.036** [0.035; 0.036]	2.5%
Early lifetime			
Breastfed	**0.723** [0.722;0.725]	**0.682** [0.68; 0.683]	5.8%
Multiple Birth	**0.023** [0.022;0.023]	**0.024** [0.024; 0.025]	6.3%
Birth Weight	**3.319** [3.316;3.321]	**3.311** [3.309; 3.314]	0.2%
Adopted as a child	**0.015** [0.014;0.015]	**0.016** [0.016; 0.017]	12.4%
Maternal smoking	**0.292** [0.291;0.294]	**0.298** [0.297; 0.3]	2%
Food/Beverage consumption			
Tea	**3.411** [3.403;3.419]	**3.406** [3.397; 3.414]	0.1%
Bread Consumed	**0.849** [0.846;0.852]	**0.835** [0.833; 0.838]	1.6%
Cooked Veg. Consumption	**2.723** [2.718;2.729]	**2.685** [2.679; 2.691]	1.4%
Cheese Consumption	**2.523** [2.52;2.526]	**2.451** [2.448; 2.454]	2.8%
Health			
Disability	**0.289** [0.288;0.29]	**0.303** [0.301; 0.304]	4.8%
Asthma	**0.127** [0.125;0.129]	**0.134** [0.132; 0.136]	5.6%
DBP	**82.259** [82.225;82.293]	**82.023** [81.989; 82.058]	0.3%
SBP	**140.21** [140.148;140.273]	**138.281** [138.22; 138.343]	1.4%
Vitamin D	**48.695** [48.632;48.757]	**46.149** [46.086; 46.213]	5.2%
Calcium	**2.38** [2.38;2.38]	**2.378** [2.378; 2.379]	0.1%
Cholesterol	**5.696** [5.693;5.699]	**5.597** [5.594; 5.6]	1.7%
White Blood Cell Count	**6.882** [6.876;6.888]	**7.014** [7.008; 7.02]	1.9%
Red Blood Cell Count	**4.517** [4.516;4.518]	**4.544** [4.543; 4.546]	0.6%
Chest Pain	**0.162** [0.161;0.164]	**0.188** [0.187; 0.189]	15.6%
Hand grip strength (left)	**29.535** [29.503;29.567]	**30.053** [30.02; 30.086]	1.8%
Hand grip strength (right)	**31.672** [31.64;31.703]	**32.172** [32.139; 32.205]	1.6%
Health behavior			
Ever Smoked	**0.602** [0.6;0.603]	**0.608** [0.607; 0.61]	1.1%
Alcohol Freq.	**4.143** [4.139;4.148]	**3.972** [3.967; 3.976]	4.1%

Number of Cigarettes	**18.331** [18.274;18.387]	**18.723** [18.665; 18.782]	2.1%
Ever Addicated	**0.06** [0.059;0.061]	**0.074** [0.073; 0.075]	23.2%
Ever Smoked Cannabis	**0.442** [0.437;0.447]	**0.542** [0.537; 0.547]	22.5%
Max Freq. Cannabis use	**1.656** [1.645;1.666]	**1.733** [1.722; 1.744]	4.7%
Mental Health			
Depression	**0.277** [0.274;0.28]	**0.295** [0.292; 0.298]	6.6%
Other			
Left Handed	**0.093** [0.093;0.094]	**0.095** [0.094; 0.096]	1.8%
Ambidextrous	**0.017** [0.017;0.018]	**0.019** [0.019; 0.019]	10.4%
Happiness	**4.575** [4.571;4.579]	**4.519** [4.515; 4.523]	1.2%
Socioeconomic			
Townsend Cont.	**−1.317** [−1.326;−1.308]	**−0.417** [−0.427; −0.407]	68.3%
No. of people in household	**2.437** [2.433;2.441]	**2.607** [2.603; 2.611]	7%
Age completed full-time educ.	**16.718** [16.71;16.726]	**16.725** [16.716; 16.733]	0%
Time employed current job	**12.933** [12.894;12.973]	**11.911** [11.873; 11.949]	7.9%
Length working week	**35.256** [35.209;35.304]	**36.135** [36.088; 36.181]	2.5%
Heavy manual work	**1.552** [1.548;1.555]	**1.645** [1.641; 1.648]	6%
Household income	**44830.202** [44731.179;44929.225]	**43496.523** [43396.867; 43596.18]	3%

Table S3: Mean of various variables in the full UK Census 5% safeguarded subsample (ages 40-74), and UKB-eligible subsample of this Census. 95% confidence intervals around each mean included.

Variable	**Mean in full pop. [95% CI]**	**Mean in UKB-eligible pop. [95% CI]**
Demographic		
Age	**55.098** [55.082;55.115]	**54.77** [54.747;54.794]
Living in a couple	**0.719** [0.718;0.72]	**0.689** [0.688;0.69]
Born outside UK	**0.119** [0.119;0.12]	**0.164** [0.163;0.165]
White	**0.916** [0.916;0.917]	**0.87** [0.869;0.871]
Sex	**0.509** [0.508;0.51]	**0.507** [0.506;0.509]
Household size	**2.61** [2.608;2.613]	**2.649** [2.646;2.653]
Socioeconomic status		
Deprived in education dimension	**0.264** [0.264;0.265]	**0.264** [0.263;0.265]
Deprived in employment dimension	**0.154** [0.153;0.155]	**0.179** [0.178;0.18]
Deprived in health and disability dimension	**0.357** [0.356;0.358]	**0.371** [0.369;0.372]
Deprived in housing dimension	**0.082** [0.082;0.083]	**0.098** [0.098;0.099]
Deprivation indicator (total)	**1.857** [1.856;1.859]	**1.912** [1.91;1.915]
Years of education	**12.652** [12.643;12.66]	**12.643** [12.629;12.656]
Owns house	**0.407** [0.406;0.408]	**0.361** [0.36;0.362]
No. of cars	**1.463** [1.462;1.465]	**1.366** [1.363;1.368]
Health		
Self-reported health	**3.96** [3.958;3.962]	**3.91** [3.908;3.913]
Disability	**0.235** [0.234;0.236]	**0.247** [0.246;0.248]
Number of housecarers in household	**0.335** [0.334;0.336]	**0.347** [0.345;0.348]
No. in household with illness/disability	**0.449** [0.448;0.45]	**0.468** [0.466;0.47]
Employment		
Employed	**0.612** [0.611;0.613]	**0.613** [0.612;0.614]
Retired	**0.222** [0.222;0.223]	**0.216** [0.215;0.217]
Unemployed	**0.053** [0.053;0.053]	**0.04** [0.039;0.04]
Ever worked	**0.81** [0.809;0.811]	**0.866** [0.864;0.867]
Urbanicity		
Persons per room	**0.407** [0.407;0.407]	**0.424** [0.424;0.425]
Goes to work by public transport	**0.113** [0.112;0.114]	**0.167** [0.166;0.168]
Goes to work by car/motorcycle	**0.642** [0.641;0.643]	**0.635** [0.633;0.637]
Lives in flat or appartment	**0.135** [0.134;0.135]	**0.163** [0.162;0.164]
Religion		
Has no religion	**0.2** [0.2;0.201]	**0.182** [0.181;0.183]
Christian	**0.674** [0.673;0.674]	**0.667** [0.666;0.668]
Other religion	**0.126** [0.125;0.127]	**0.151** [0.15;0.152]
Observations	1 277 785	565 994

Variable	UK Census	UKB
Sex	0%	0%
Education	0%	2.24%
Region	0%	0%
Year of Birth	0%	0%
Health (Self-reported)	0%	0.47%
Tenure of dwelling	0.63%	1.45%
Employment status	0%	0.93%
Number of cars	0.63%	0.7%
One-person household	0.63%	0.68%
Ethnicity	0%	0%

Table S4: Prevalence of missing observations in UKB-eligible Census and UKB

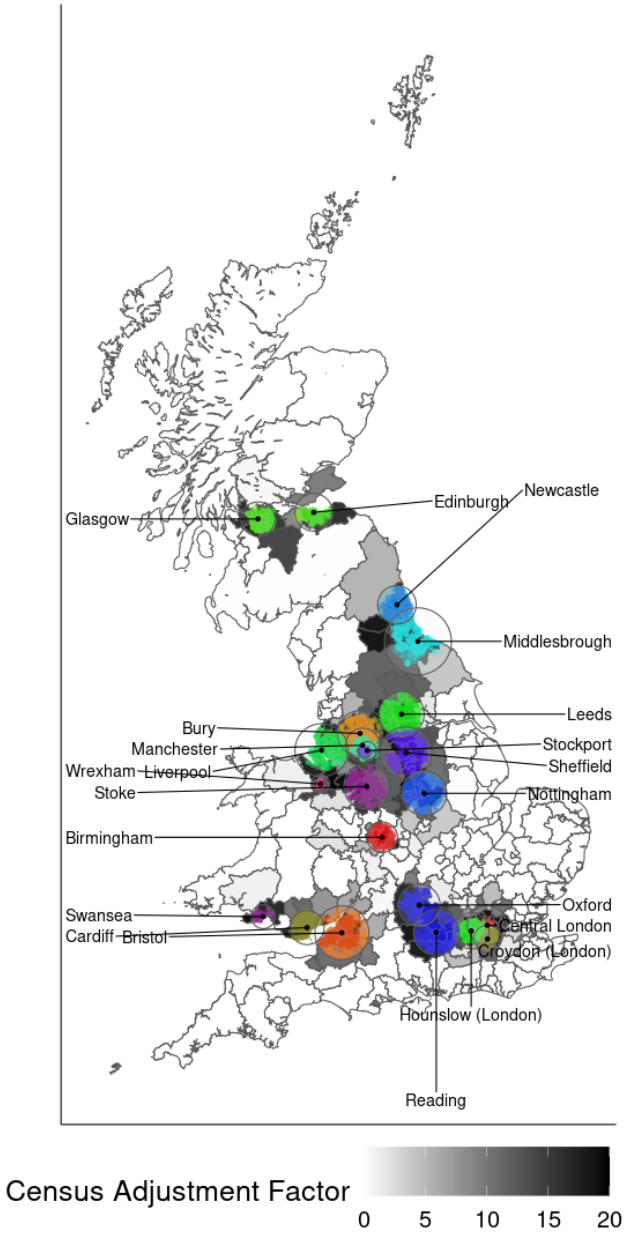


Figure S1: Geographic distribution of UKB participants and the UKB-eligible Census. Each dot on the map represents the geographic location of a UKB participant's residence, coloured by the assessment centre that they visited. Each black dot shows the location of an assessment centre. Each circle visualizes the inferred sampling radius around each assessment centre. Census Grouped Local Authority (GLA) regions that are included in the UKB-eligible Census are coloured in grey. For Census observations from GLA regions that are fully within any assessment centre's sampling region, we assign a sampling weight of 20 (darkest grey shades in the map). For UKB-eligible Census observations from GLA regions that fall only partially within an assessment centre's sampling region, we use adjustment factors of 20 times the share of this region's population that lives within the assessment centre's radius (lighter grey shades in the map).

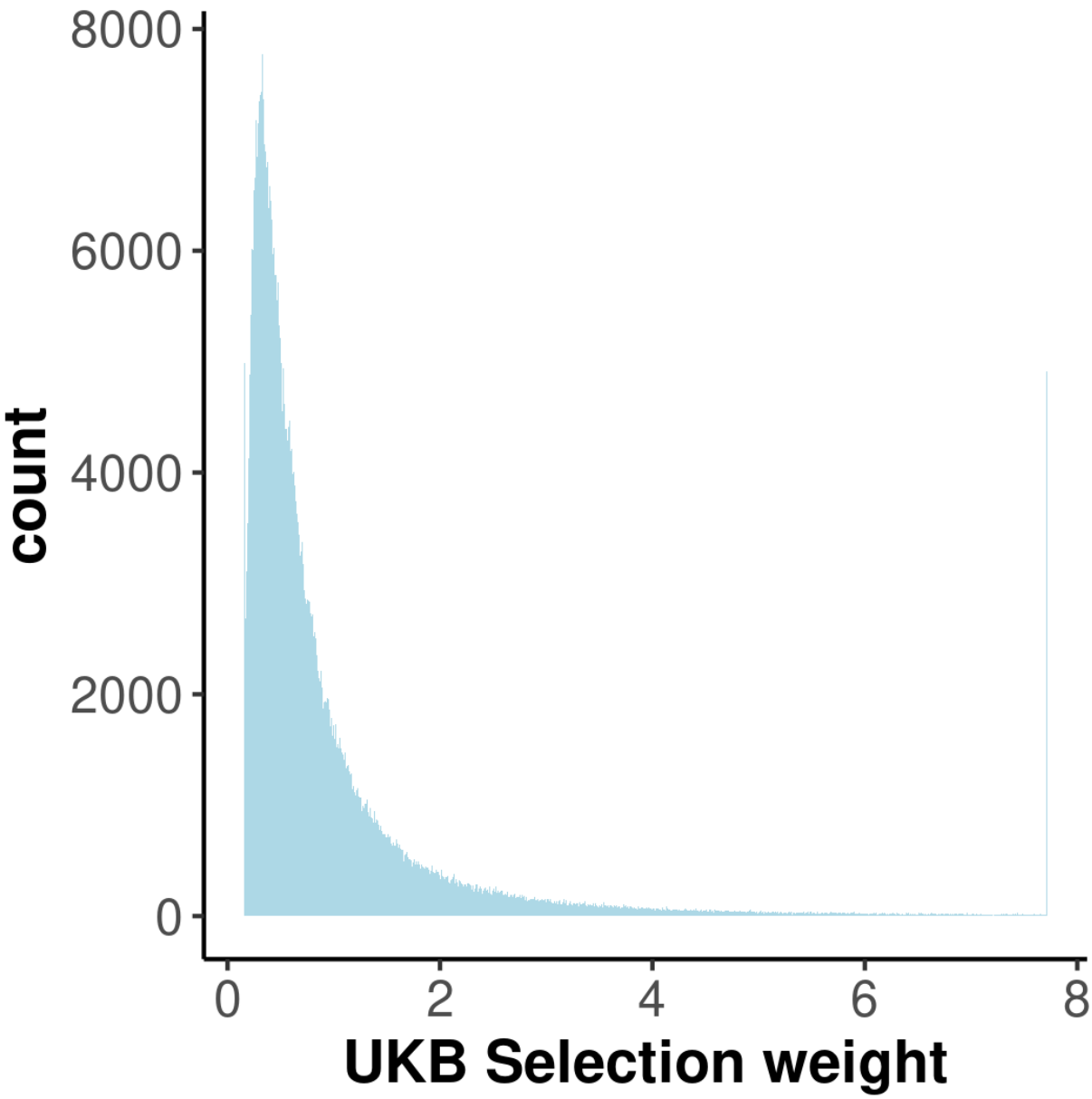


Figure S2: Histogram of the distribution of UKB IP weights after winsorising (setting values below the 1st percentile equal to the value at the 1st percentile, and values above the 99th percentile equal to the 99th percentile.)

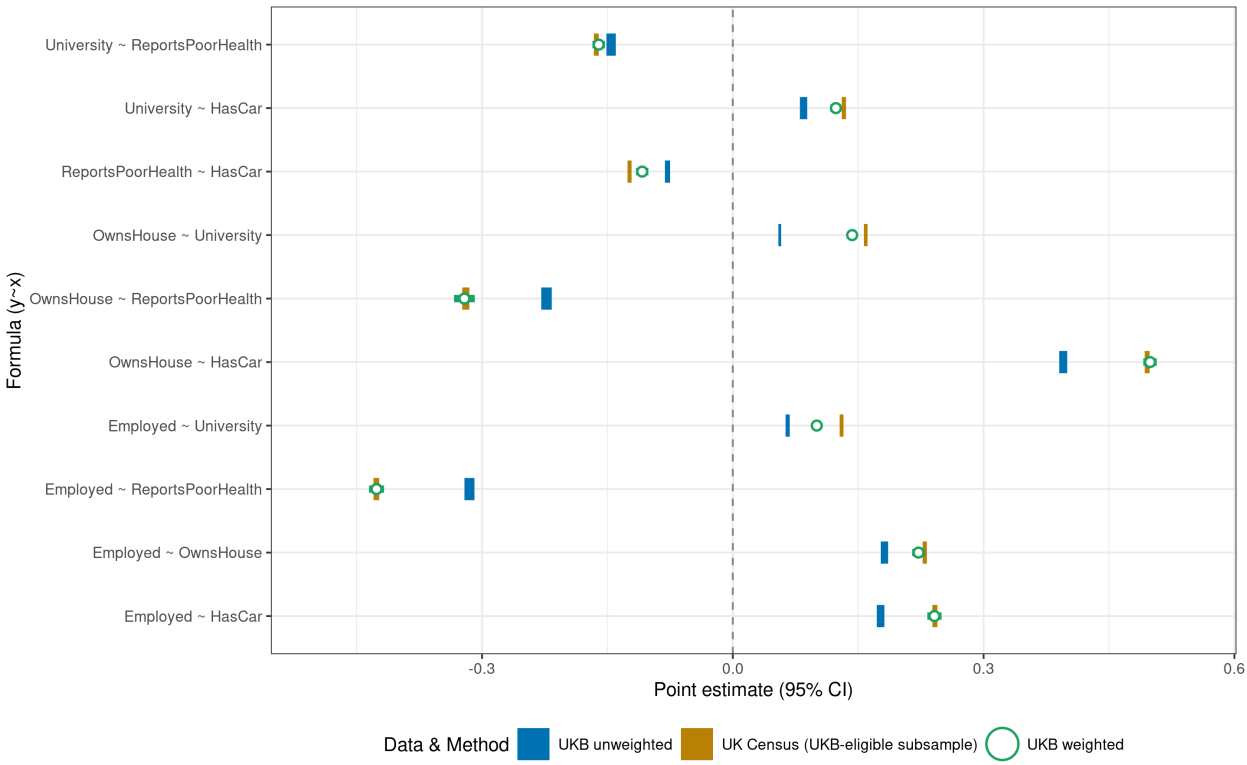


Figure S3: Estimated associations based on bivariate linear models in UKB (solid blue bars), UKB-eligible Census (solid yellow bars), and Weighted UKB (open green circles) (as in Figure 4), after (linearly) controlling for year of birth and sex. Bar widths indicate 95% confidence intervals (heteroskedasticity-robust standard errors). All blue and yellow bars are highly significantly different from one another ($P < 10^{-8}$).

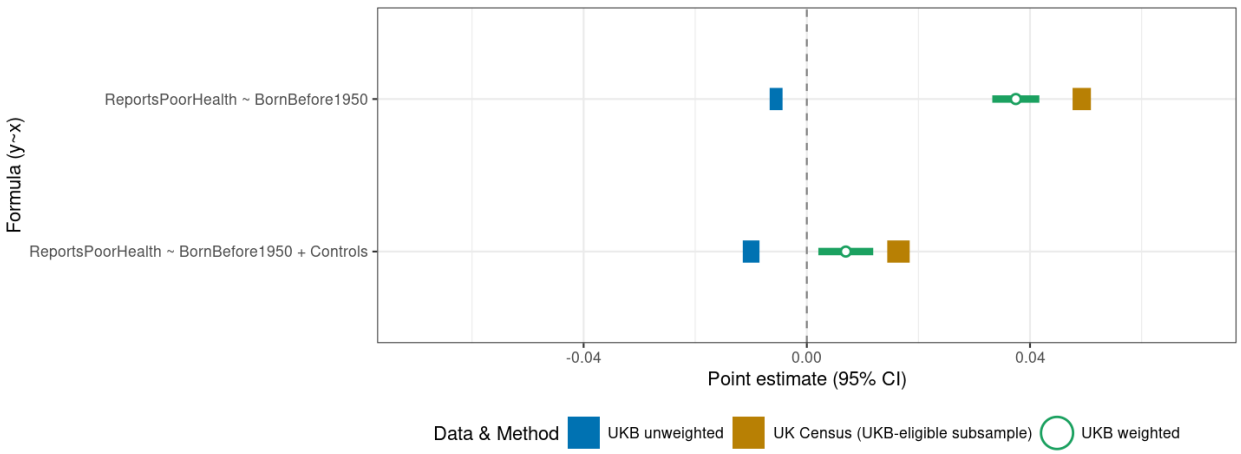


Figure S4: Estimated coefficients for the effect of being born before 1950 on reporting poor health in UKB (solid blue bars), UKB-eligible Census (solid yellow bars), and Weighted UKB (open green circles). The first model shown does not include any control variables and is the same as the one reported in Figure 4. The second model shows the same coefficients after adding various control variables to the regression: sex, years of education, number of cars owned, a single household indicator, tenure of household (4 dummy variables), employment status (5 dummy variables), ethnicity (4 dummy variables), and region (142 dummy variables).

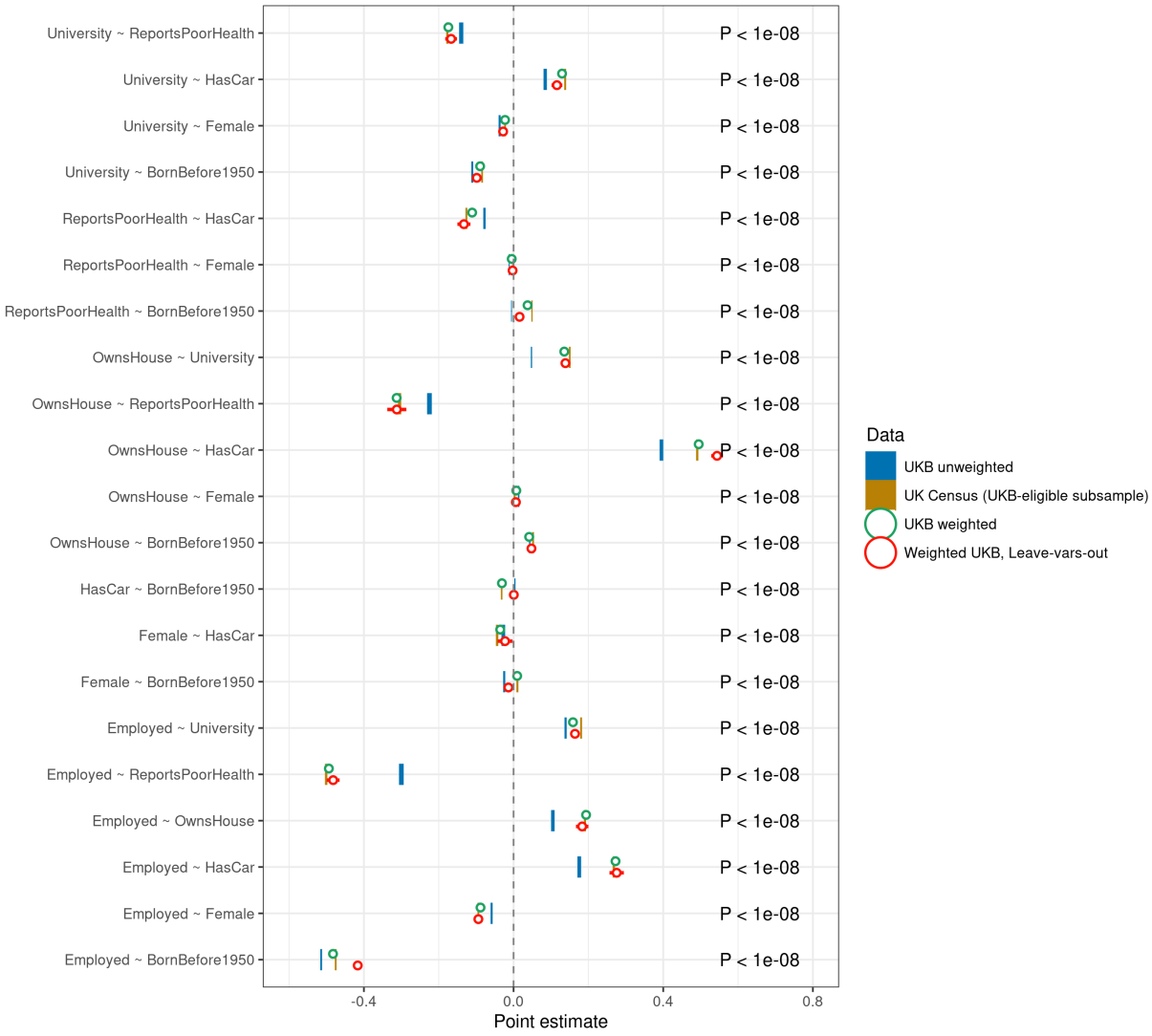
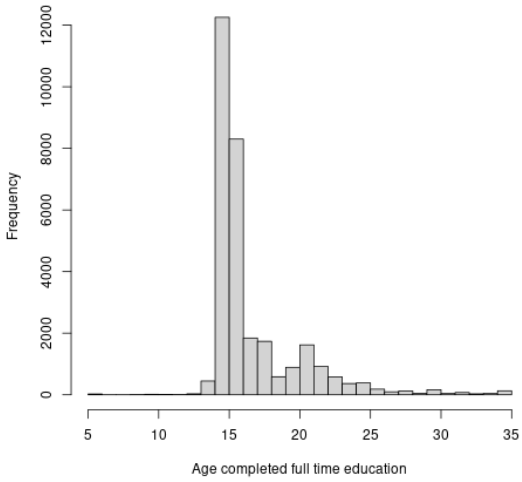
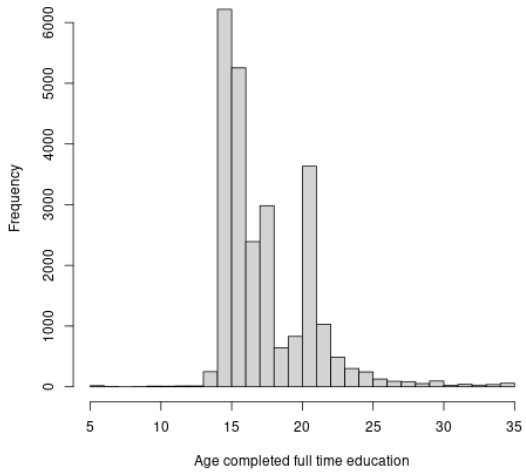


Figure S5: Estimated coefficients for bivariate linear models in UKB and UK Census when leaving relevant model variables out of IP-weight construction. Associations are as shown in Figure 4. Additionally, the red open circles show the results from weighted models in the UKB that are based on "leave-variables-out" weights. These weights are constructed in the same manner as the IP weights, but are based on a LASSO model that did *not* include the dependent and independent variable included in the association model shown. All results estimated on the first holdout sample only.



(a) Distribution of "age completed full time education", for those with an NVQ or HNC or HND or equivalent



(b) Distribution of "age completed full time education", for those with a professional qualification not elsewhere classified

Figure S6

References

1 Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. American journal of epidemiology. 2017;186(9):1026-34.

2 Lee JJ, Wedow R, Okbay A, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nature genetics. 2018;50:1112-21.

3 Education GPS;. Accessed: 2022-02-11. https://gpseducation.oecd.org/CountryProfile?primaryCountry=GBR&treshold=10&topic=EO.

4 Okbay A, Wu Y, Wang N, et al. Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. Nature genetics. 2022;54:437-49.

5 Pirastu N, Cordioli M, Nandakumar P, Mignogna G, Abdellaoui A, Hollis B, et al. Genetic analyses identify widespread sex-differential participation bias. Nature Genetics. 2021;53(5):663-71.