# The Genetics of Participation: Method and Analysis

Stefania Benonisdottir[1,*] and Augustine Kong[1,*]

[1]Big Data Institute, Li Ka Shing Centre for Health Information Discovery, University of Oxford, Oxford, UK.

*Corresponding authors: stefania.benonisdottir@bdi.ox.ac.uk and augustine.kong@bdi.ox.ac.uk

**Abstract:** Participation in a genetic study likely has a genetic component. Identifying such component is difficult as we cannot compare genetic information of participants with non-participants directly, the latter being unavailable. Here, we show that alleles that are more common in participants than non-participants would be further enriched in genetic segments shared by two related participants. Genome-wide analysis was performed by comparing allele frequencies in shared and not-shared genetic segments of first-degree relative pairs of the UK Biobank. A polygenic score constructed from that analysis, in non-overlapping samples, is associated with educational attainment ($P = 2.1 \times 10^{-52}$), body mass index ($P = 1.5 \times 10^{-19}$), and participation in a dietary study ($P = 6.9 \times 10^{-21}$). Further analysis shows that inclination to participate is a behavioural trait in its own right, and not simply a consequence of other established phenotypes. Understanding the basis of this trait is important for data analyses and the design of future surveys, genetic or otherwise.

## Introduction

For all sample surveys, ascertainment bias, *i.e.* the sample is not representative of the population, is a problem that could lead to seriously misleading conclusions[1,2]. By its very nature, ascertainment bias usually cannot be evaluated based on the sample alone[3]. For example, with a target variable such as the presidential candidate one is voting for or whether one has been vaccinated, other variables (covariates) that have known distributions for both sample and population are needed for potential adjustments[1-3]. Such adjustments are inherently imperfect as the covariates are unlikely to fully capture the correlation between participation and the target variable[1,3]. For genetic investigations, among participants of the primary study who have contributed DNA, further engagement in optional components of the study has been demonstrated to have associations with both genotypes and phenotypes[4-7]. That, however, does not address the potential genotypic difference between the primary study participants and the target population. Given this background, it is striking to see that it is actually possible to investigate how the sampled genotypes are biased based on themselves alone. A recent study identified single nucleotide polymorphisms (SNPs) that had significant allele frequency differences between males and females in the samples, and proposed that those variants have differential participation effects for the sexes[8]. This relies on the assumption that genotypes of autosomal variants should have the same frequencies in the population for both sexes. This approach, however, cannot identify variants that affect primary study participation of both sexes in a similar manner, which are presumably more abundant and have broader impact. Here we show a way to study these variants.

There are two properties that make genetic data distinct. Firstly, all individuals are genetically related by various degrees, and the relatedness are captured by the genotypes. Secondly, each individual has two copies of genetic segments on autosomal chromosomes, one inherited from

each parent. Some of these segments are identical by descent (IBD), *i.e.* inherited from a recent common ancestor, with genetic segments in a relative. Conceptually, instead of comparing individuals, we compare genetic segments. The key idea introduced here is that if an allele has a higher frequency in participants than non-participants, it would also have higher frequency in segments that are in two participants compared to segments that are only in one participant. This alternative view of the data leads to the three principles of genetic induced participation bias described below. These principles allow us to perform a genome-wide association scan (GWAS) of study participation variants using only the genetic data of the study without phenotypes. Most importantly, this analysis is only sensitive to direct genetic effects[9-11], and is immune to indirect genetic effects and confounding effects such as those induced by population stratification[12]. Examination of the GWAS results reveals that, while a person's participation is related to other characteristics such as educational attainment, the effect of the genetic component of participation is not simply manifested through them. This partially explains why standard covariate adjustment cannot fully eliminate the effect of ascertainment bias in general.

*The First Principle of Genetic Induced Ascertainment Bias:*

*On average, between two ascertained individuals, genetic segments shared identical by descent (IBD), relative to segments that are not, are enriched with alleles that have positive direct effects on ascertainment probability.*

This principle (Fig. 1A) relies on individuals in the population being genetically related, closely or distantly. Gene alleles that are enriched in the participants compared to the non-participants are expected to be further enriched in segments shared by ascertained relatives. With a large sample, at a specific SNP locus, many pairs of individuals would share one long haplotype, inherited identical by descent from a not-very-distant common ancestor. For each of such pairs, there is one distinct shared haplotype, and two distinct not-shared haplotypes. Tabulating over

all such pairs the SNP alleles in the shared haplotypes and the not-shared haplotypes, the SNP allele that promotes participation/ascertainment should tend to have a higher frequency in the shared than the not-shared haplotypes. This becomes a case-control analysis where the shared and not-shared alleles are the cases and controls respectively, and matched to the extent that they are in the same individuals. Still, that does not remove potential confounding entirely as haplotypes in the population that were driven to higher frequency through natural selection would also be shared by more individuals, both in the population and in the sample. Essentially, ascertainment bias is a form of selection and to cleanly distinguish it from other forms of selection requires more stringent matching of shared and not-shared haplotypes. We achieve that by using ascertained parent-offspring and sibling pairs.

*Ascertained parent-offspring pairs (Fig. 1B).* For an ascertained parent-offspring pair where the other parent is not ascertained, there are three distinct alleles by descent, ($S$) the allele in the genomic segment transmitted from parent to offspring, ($NS_P$) the allele in the parental genomic segment not transmitted to the offspring, and ($NS_{OP}$) the allele inherited by the offspring from the other parent. Thinking of the offspring as the proband, the $NS_P$ allele is a perfect match for the $S$ allele as they are both in the ascertained parent. Mendelian inheritance dictates that each would have the same chance to be transmitted to the offspring to become the shared allele. The principle is similar to that underlying the transmission disequilibrium test[13]. The $NS_{OP}$ allele is not as perfectly matched, and while it could be used for subsequent validation, it is not used in our initial analysis. With alleles coded as 0/1 and $n_{PO}$ ascertained parent-offspring pairs,

$$DPO \stackrel{\text{def}}{=} \frac{\sum_i (S_i - NS_{Pi})}{n_{PO}} = \frac{\sum_i DPO_i}{n_{PO}} = \frac{\sum_i S_i}{n_{PO}} - \frac{\sum_i NS_{Pi}}{n_{PO}} \stackrel{\text{def}}{=} F_{PT} - F_{PNT}\,,$$

where $i = 1, \ldots, n_{PO}$ indexes the pairs, can be used to test for association between SNP and ascertainment. In particular, conditional on the genotypes of the ascertained parents, $DPO$ has expectation zero under the null hypothesis of no ascertainment bias, and is not subject to any population stratification induced bias. The case where both parents are ascertained together with an offspring can be treated as two parent-offspring pairs as the transmission from the two parents are independent under the null hypothesis.

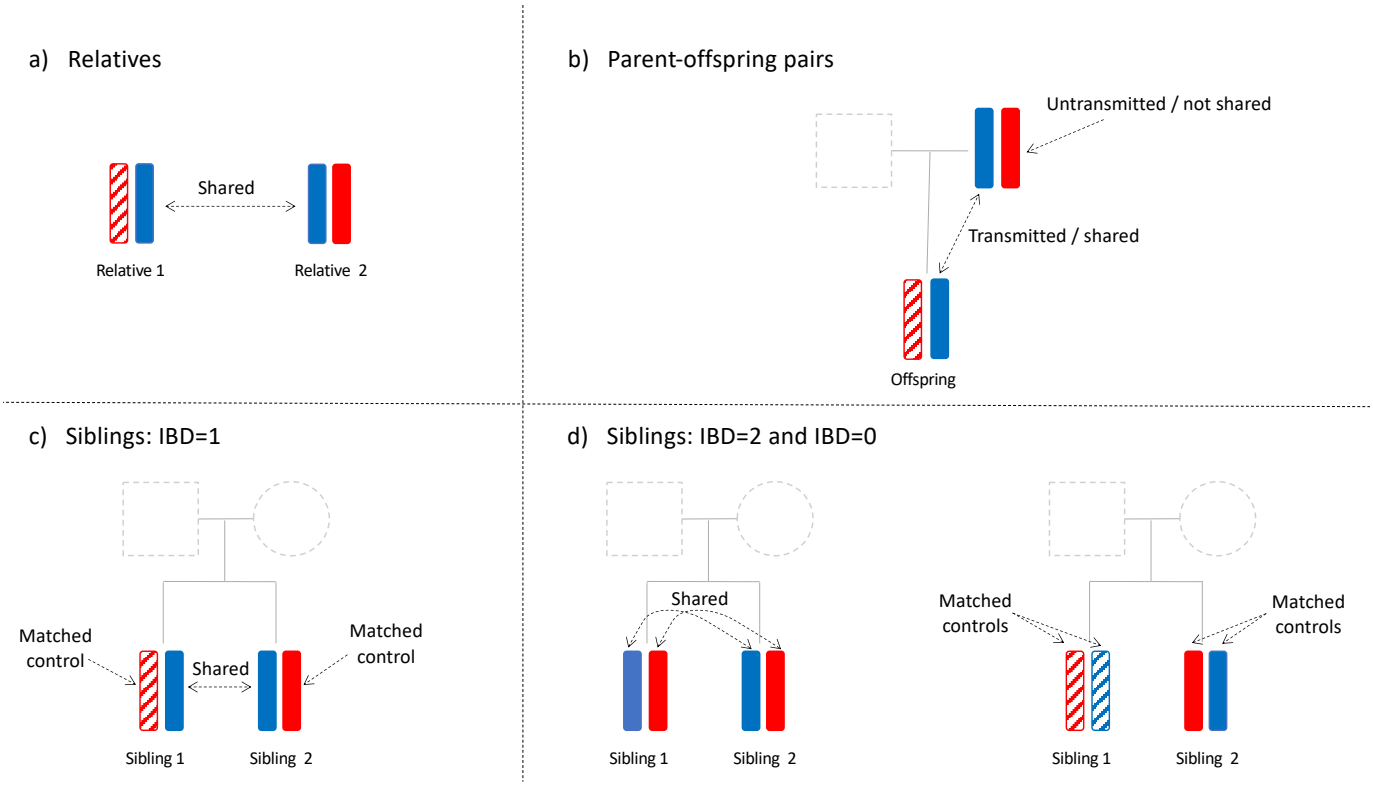


**Figure 1. The First Principle for Genetically Induced Participation Bias; comparing shared and not-shared alleles**. a) General relative pairs. b) Parent-offspring pairs. c) Sibling pairs with IBD = 1 at SNP locus. d) Sibling pairs with IBD = 2 and IBD = 0 at SNP locus.

*Ascertained sib-pairs, parents not ascertained.* Let $n_{SIB}$ be the number of ascertained sib-pairs with unascertained parents. At a specific locus, assuming random Mendelian transmission, a sib-pair has probability ½ of inheriting the same allele IBD from the father, and the same independently from the mother. It follows that a sib-pair would share 2, 1, or 0 alleles IBD with

probabilities ¼, ½ and ¼ respectively. With data from dense SNPs, the IBD state of a locus can usually be determined with high accuracy[14]. For a specific SNP, based on the sib-pairs with IBD states that are assumed known, two frequency-difference statistics are derived as follows. For each sib-pair that has IBD state 1 (Fig. 1C), there are one distinct share allele ($S$) and two distinct not-shared alleles ($NS_1$ and $NS_2$). Define

$$DSIB1 = \frac{\sum_i (S_i - (NS_{1i} + NS_{2i})/2)}{n_{SIB1}} = \frac{\sum_i S_i}{n_{SIB1}} - \frac{(NS_{1i} + NS_{2i})}{2 \times n_{SIB1}} = F_{IBD1S} - F_{IBD1NS}$$

where $i = 1, \ldots, n_{SIB1}$. Note that for any one of these sib-pairs, if the shared allele is paternally inherited, then the two not-shared alleles are maternally inherited, and vice versa. Despite that, conditional on the genotypes of the two parents, without ascertainment bias, $DSIB1$ has expectation 0. This holds even in the most extreme case where fathers carry only allele 1 and mothers carry only allele 0. That can be demonstrated by considering the four parent-offspring transmissions --- one paternal transmission for each sib and one maternal transmission for each sib --- jointly (Fig. S1). Notably, when the IBD state is 1, the shared allele is equally likely to be paternal or maternal, and thus any systematic differences between fathers and mothers cancel in expectation. For the approximately one-quarter of the sib-pairs ($n_{IBD2}$) at a locus that share two alleles IBD, the average allele frequency is

$$F_{IBD2} = \frac{\sum_i (G_{1i} + G_{2i})}{4 \times n_{IBD2}} = \frac{\sum_i G_{1i}}{2 \times n_{IBD2}}$$

where $i = 1, \ldots, n_{SIB2}$, indexes the sib-pairs with IBD state 2, and $G_{1i}$ and the identical $G_{2i}$ are the genotypes of sib 1 and sib 2 respectively in pair $i$. Similarly, for the approximately one-quarter of the sib-pairs ($n_{IBD0}$) at a locus that share zero alleles IBD, let

$$F_{IBD0} = \frac{\sum_j (G_{1j} + G_{2j})}{4 \times n_{IBD0}}.$$

where $j = 1, \ldots, n_{SIB0}$, indexes the sib-pairs with IBD state 0, and $G_{1j}$ and $G_{2j}$ are the genotypes of sib 1 and sib 2 respectively in pair $j$. The difference (Fig. 1D)

$$DSIB20 = F_{IBD2} - F_{IBD0}$$

is another test statistic for ascertainment bias. Here $F_{IBD2}$ and $F_{IBD0}$ are allele frequencies from different sib-pairs. However, for a sib-pair at a particular locus, the chance to be in IBD state 0 or 2 is the same, and thus, without ascertainment bias, $DSIB20$, has expectation zero.

In summary, $DPO$, $DSIB1$, and $DSIB20$ are only sensitive to the direct effects[9,15] on ascertainment, and are unaffected by population stratification induced bias[12,15] or indirect genetic effects from relatives [9-11]. However, direct genetic effects can manifest in various ways, their relative contributions depending on the sampling scheme. For example, the effect can be 'voluntary', a person is more likely to participate when invited, or 'involuntary', *e.g.* a person is less likely to be invited because of its attributes. The three statistics can be combined, *e.g.* as a weighted linear combination, into one test statistic. However, having separate values of them are useful because they can have different expectations depending on the nature of the ascertainment bias, and they are impacted differently by genotyping and data processing errors.

**UK Biobank and Data Processing Artefacts**

The UK Biobank (UKBB) is a large-scale database with genetic and phenotypic information of individuals from across the United Kingdom (UK)[16]. Invitations to participate were sent to 9,238,453 individuals who were aged between 40 and 69 years and lived within 25-mile radius of any of the 22 UKBB assessment centres[17]. In the end, 5.45% of those participated in the study (~500,000 individuals) and went through baseline assessments that took place from 2006 to 2010.[17] In addition to phenotypic details collected at the baseline visit, information continued to be added, including follow-up studies for large subsets of the cohort[17-19]. It is known that the

UKBB sample is not fully representative of the UK population[16,17,20]. The participants were more likely to be female[17], less likely to smoke[17] and were older[17] than non-participants. Also, compared to the national average, participants were more educated[20], taller[17] and had a lower BMI[17].

We applied our methods to 4,427 parent-offspring pairs and 16,668 sibling pairs of white British (WB) descent (Fig. S2, Table S1 and Supp. Text). Association analysis was performed for 597,039 directly genotyped phased SNPs[16] that passed quality control filtering (Supp. Text) and had a minor allele frequency (MAF) > 1%. For each sibling pair, IBD sharing status (0, 1 or 2) of every SNP was 'called' using the program KING[14] (Supp. Text). Initial results had the major allele of a SNP significantly more likely to be positively associated with participation. Examination of possible artefacts exposed a few data processing issues, leading to adjustments outlined below.

For many loci, the IBD fractions of the sibling pairs, calculated from the calls, deviated significantly from the expected IBD fractions (Supp. Text, Fig. S3) indicating IBD calling errors. Errors were most likely to occur around recombination events and ends of chromosomes (Fig. S3) and to reduce their impact, for each sibling pair, we identified the IBD segments --- contiguous SNPs with the same IBD status called --- and trimmed away 250 SNPs from each end of a segment before the association analysis. This removed 250 SNPs at each end of the chromosomes entirely from the association analysis and reduced slightly the sample size of the other SNPs (Fig. S4, median reduction of 1,516 pairs).

For a parent-offspring pair and for SNPs that have IBD status 1 for a sibling pair, the shared allele and not-shared alleles are clear unless both members of the pair are heterozygotes, in which case the phasing with neighbouring SNPs was used to resolve the uncertainty (Fig. S5 and Supp. Text). Our analysis estimated that the phasing provided in the UK Biobank data

release[16] had an error rate of approximately 0.5% and an adjustment was made to the test statistics *DPO* and *DSIB1* to account for the induced bias (Supp. Text).

After the above adjustments, the tendency for the major allele to be positively associated with participation was substantially reduced but not completely eliminated. Even though part of this 'major allele effect' might be real, we made further adjustments based on MAF to remove this tendency from our association results (Supp. Text). After all these adjustments, a number of highly significant associations remained in the major histocompatibility complex (MHC), a region with extended linkage disequilibrium (LD) and high SNP density. Further investigations led us to believe these signals were most likely artefacts (Supp. Text and Fig. S6). After removing 152,582 SNPs from the MHC and other extended LD regions (Table S2) as well as additional quality control filtering (Supp. Text and Fig. S7), 444,457 SNPs remained in our participation genome scan.

**Results**

For each SNP, we computed three t-statistics by dividing each of $DPO$, $DSIB1$, and $DSIB20$, by its standard error (SE). The three t-statistics were then combined into one, based on their SEs, and then converted to a nominal chi-square statistic. The latter were adjusted by genomic control using the genomic inflation factor. A QQ-plot of the p-values ($P$) based on the adjusted chi-square statistics (Fig. S8) shows that one SNP, rs113001936 on chromosome 16, is significant with Bonferroni correction, and the excess of SNPs with $P < 0.01$. A literature search did not provide any obvious reasons for rs113001936 to have a participation effect. To validate our results, we derived the weights of a participation polygenic scores (*pPGS*) based on our participation GWAS (Supp. Text). Values of the *pPGS*, standardized to have variance one, were computed for 272,409 WB individuals with no close relatives in UKBB (>3$^{rd}$ degree for all pairs). They are referred to as the 'unrelateds' and notably do not overlap with the first-

degree relatives used. Associations between the *pPGS* and various quantitative traits were examined using the unrelateds. Table 1 shows some of the strongest associations and includes a few non-significant ones. The strongest association is with EA where the effect (correlation) is 0.0307 with $P = 2.1 \times 10^{-52}$. The effect, 0.0299, is nearly as strong for age-at-first-birth (AFB) of women ($P = 1.2 \times 10^{-20}$). The next strongest association is with BMI ($P = 1.5 \times 10^{-19}$) where the effect is notably negative. These results, consistent with the known differences in EA and BMI between sample and population[17,20], validated that our GWAS performed without any phenotype data can nonetheless capture genetic associations with participation.

After the baseline assessments, subsets of the UKBB participants were invited to further participate in a dietary study in 2011-2012 where invited participants were asked to answer a dietary questionnaire[19] and a physical activity study in 2013-2015 where invited participants were asked to wear an accelerometer for a week[18]. Not everybody was invited (the criteria for invitation included having a valid email address) and only a subset of those invited actually participated (Table S3). We refer to participation in these follow-up studies as 'secondary participation'[4-7]. For the dietary study, the estimated effect of *pPGS* on being invited, in $\log_e$ odds-ratio ($\log(OR)$), is of 0.0348 ($P = 1.7 \times 10^{-16}$, Table 1). For those invited, the estimated effect on actual participation in $\log(OR)$ is 0.0290 ($P = 3.6 \times 10^{-8}$). For the physical activity study, the corresponding effect estimates are 0.0275 ($P = 1.6 \times 10^{-12}$) and 0.0328 ($P = 3.4 \times 10^{-9}$) respectively. Comparing the *pPGS* values of the three groups --- uninvited, invited but did not participate, participated ---, a 2-*df* test gives *P* of $6.9 \times 10^{-21}$ and $8.5 \times 10^{-19}$ respectively for the dietary and physical activity studies.

| Phenotype | | | Adjusted for EA | | |
|---|---|---|---|---|---|
| **Quantitative traits** | **Effect** | **P-value** | **Effect** | **P-value** | **N** |
| Educational attainment (EA) | 0.0307 | $2.1 \times 10^{-52}$ | - | - | 260,950 |
| Age at first birth | 0.0299 | $1.2 \times 10^{-20}$ | 0.0196 | $1.2 \times 10^{-10}$ | 98,653 |
| BMI | -0.0177 | $1.5 \times 10^{-19}$ | -0.0137 | $4.4 \times 10^{-12}$ | 271,535 |
| HDL cholesterol | 0.0131 | $7.2 \times 10^{-10}$ | 0.0098 | $5.7 \times 10^{-6}$ | 237,785 |
| Height | 0.011 | $7.7 \times 10^{-8}$ | 0.0067 | $1.3 \times 10^{-3}$ | 271,820 |
| HbA1c | -0.0073 | $3.2 \times 10^{-4}$ | -0.0055 | $8.2 \times 10^{-3}$ | 259,594 |
| N. children | -0.0055 | $4.4 \times 10^{-3}$ | -0.0038 | 0.052 | 271,317 |
| N. siblings | -0.0048 | 0.020 | -0.0001 | 0.94 | 268,191 |
| Vitamin D | -0.0031 | 0.13 | -0.0015 | 0.45 | 249,079 |
| Glucose | -0.0027 | 0.19 | -0.0013 | 0.54 | 237,629 |
| Grip strength | -0.0021 | 0.29 | -0.0034 | 0.091 | 270,525 |
| Lipoprotein A | -0.0014 | 0.50 | -0.0010 | 0.64 | 251,698 |
| SHBG | 0.0013 | 0.55 | 0.0012 | 0.58 | 235,960 |
| Testosterone | 0.0004 | 0.85 | 0.0006 | 0.76 | 257,570 |
| **Binary traits** | **log(OR)** | **P-value** | **log(OR)** | **P-value** | **Ncases/Ncontrols** |
| Dietary study invitation | 0.0348 | $1.7 \times 10^{-16}$ | 0.0222 | $5.0 \times 10^{-7}$ | 166,993/105,416 |
| Dietary study participation | 0.0290 | $3.6 \times 10^{-8}$ | 0.0220 | $5.0 \times 10^{-5}$ | 54,124/112,869 |
| Physical activity study invitation | 0.0275 | $1.6 \times 10^{-12}$ | 0.0184 | $5.2 \times 10^{-6}$ | 132,633/139,776 |
| Physical activity study participation | 0.0328 | $3.4 \times 10^{-9}$ | 0.0281 | $8.1 \times 10^{-7}$ | 59,455/73,178 |

**Table 1: Participation PGS and association with phenotypes.** This table depicts how the participation polygenic score (PGS) is associated with a range of phenotypes. For the quantitative traits, we performed linear regression with the corresponding phenotype as a response and the PGS and genotyping array (BiLEVE or Axiom) as explanatory variables in the subset of individuals of White British descent who have no close relatives in UK Biobank (>3$^{rd}$ degree for all pairs). Prior to regression analysis, the quantitative traits were adjusted for year of birth (YOB), age at measure and 40 PCs separately for each sex and then the residuals were standardised also separately for each sex (Supp. Text). Grip strength was additionally adjusted for height. BMI: body max index. HDL cholesterol: High-density lipoprotein cholesterol. HbA1c: Glycated Haemoglobin. N. children: Number of children. N. siblings: Number of full siblings. SHBG: Sex hormone binding globulin. Information about age at first birth was only available for women. For the binary traits, we performed a logistic regression including YOB, age at measure up to the order of three, 40 PCs, sex and genotyping array as additional covariates. Columns 2 and 3, labelled 'Effect/log(OR)' and 'P-value', depict the slope/log(odds-ratio) for the participation PGS and the corresponding P-value. The P-values have been adjusted with the LD-score regression intercept of the corresponding phenotype (Supp. Text). Columns 4 and 5 show the slope/log(odds-ratio) and the P-value for the participation PGS when educational attainment (EA) has been added as an additional covariate. SNP-wise weights for the participation PGS were computed using identity-by-descent information from 16,668 siblings and 4,427 parent-offspring pairs in UK Biobank.

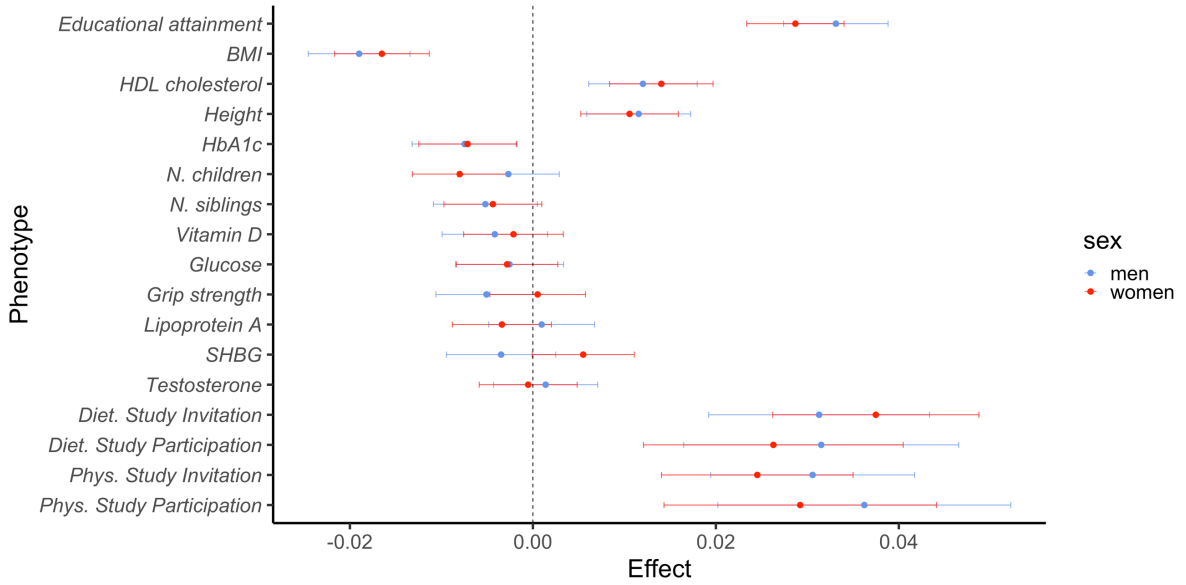


**Figure 2: Participation PGS and sex-specific analysis.** This figure depicts the results from regressing the participation polygenic score (PGS) on a range of phenotypes for men and women of White British descent who have no close relatives in UK Biobank (>3$^{rd}$ degree for all pairs). Each dot depicts the effect in SD or log(OR) of the participation PGS (x-axis) on the corresponding phenotype (y-axis). The vertical lines depict 95% confidence intervals taking the corresponding LD-score regression intercept into account (Supp. Text). The blue dots correspond to regression in the subset of men while the red dots correspond to regression in the subset of women. Genotyping array (BiLEVE or Axiom) was an additional covariate in the regressions. For men, the sample size ranged from 111,219 individuals to 127,186 individuals and for women the sample size ranged from 124,741 to 145,223. SNP-wise weights for the PGS were computed using IBD information from siblings and parent-offspring pairs (not sex-specific weights). Prior linear regression for each sex, the quantitative phenotypes were adjusted for year of birth (YOB), age at measure and 40 PCs separately for each sex and then the residuals were standardised also separately for each sex (Supp. Text). For the binary phenotypes, we performed a logistic regression separately for each sex, including YOB, age at measure up to the order of three, 40 PCs, and genotyping array as additional covariates. Grip strength was additionally adjusted for height. BMI: body max index. HDL cholesterol: High-density lipoprotein cholesterol. HbA1c: Glycated haemoglobin. N. children: Number of children. N. siblings: Number of full siblings. SHBG: Sex hormone binding globulin.

Associations of *pPGS* were further examined with adjustment for EA (Table 1). For traits/variables with unadjusted $P < 1 \times 10^{-3}$, the adjusted effects shrink but all remain significant. Relatively, the effect on participating in the physical study when invited shrinks the least, by 14.3%, from 0.0328 to 0.0281, while the effect on dietary study invitation shrinks by 36.2%, from 0.0348 to 0.0222. This difference in shrinkage is partly because the estimated effect of EA on dietary study invitation is 0.4661 (Table S4), much larger than its effect on physical activity study participation, 0.1826. From the EA effects alone, the ascertainment bias

would appear to be much stronger with dietary study invitation. By contrast, on the genetic level, the ascertainment bias for physical study participation is stronger than that for invitation after adjustment for EA. Thus, while the genetic component to participation is associated with EA, its effects on other traits is not manifested mainly through EA. Importantly, phenotypes known to correlate with participation, do not fully capture the nature and magnitude of ascertainment bias. The estimated effects of EA are substantially larger than that of the *pPGS*. However, the *pPGS* only captures a small fraction of the full genetic component of participation, the latter could have effects that are at least comparable to those of EA, particularly for the participation traits. When males and females are analysed separately for the traits/variables in Table 1 (Fig. 2), no significant difference is found for the estimated effects of *pPGS*. Furthermore, while it has been reported that UKBB participation rates differ by sex and age[17], the *pPGS* is associated with neither ($P > 0.05$). This implies that the effect of the *pPGS*, based on autosomal variants, is additive to the effects of sex and age on participation, with no detectable statistical interactions.

UKBB did not recruit families and participants were all adults providing their own consent[16]. Under these conditions, for alleles that have an effect on participation, relative allele frequencies in different groups of individuals and genetic segments depend on many factors, most important of which are the overall participation rate and the participation rate of those with a close relative among the participants (Supp. Text and Fig. S9). For UKBB, ignoring confounding factors, by simulations, we estimated that the allele frequency differences between the shared and not-shared in sibling pairs are about 90% of the frequency differences between the alleles in participants than non-participants (Supp. Text). Alleles in the unrelateds are by definition not IBD shared through a recent common ancestor with any other participant. By comparison, participants with close relatives among the participants have both shared and not-shared alleles. Thus:

*The Second Principle --- On average, participants with close relatives among the other participants are enriched with alleles that promote participation relative to other participants.*

To examine this, we randomly partitioned the sibling pairs into two halves, derived a *pPGS* based on a GWAS performed with the first half ($pPGS_1$), and compare the $pPGS_1$ values of sibpairs in the second half with the unrelateds. The splitting reduces power. Nonetheless, the second half of the sibpairs have average $pPGS_1$ that is 0.041 standard deviation (SD) higher than that of the unrelateds ($P = 2.1 \times 10^{-5}$). Switching the roles of the two halves of the sibpairs, the first half of the sibpairs have average $pPGS_2$ that is 0.030 SD higher than that of the unrelateds ($P = 2.0 \times 10^{-3}$).

*The Third Principle --- If genetics contribute to participation, there would be more close relative pairs among the participants than what is expected if participation is random.*

This is true because a participant has a higher than average expected participation genetic component value, and thus so would its close relatives. Even though UKBB did not recruit families on purpose, the dataset contains more close relatives ($\leq 3^{rd}$ degree) than what would be expected from random sampling, *e.g.* sibling pairs are twice as many as expected[16]. They speculated that the cause of overrepresentation of close relatives is mutual consultation and possibly shared environment leading to correlated participation[16]. Here we show that shared genetics most likely also play a substantial role.

**Discussion**

Ascertainment bias, particularly participation bias, is arguably the most challenging problem in applied statistics and it is becoming increasingly relevant in the age of 'Big Data'[1,3]. In addition to obvious pitfalls with data analyses that ignored ascertainment bias, for genetic studies, more subtle consequences include collider bias[20] that decreases correlations between

contributing alleles and the introduction of artificial epistatic effects (Supp. Text). Here we show that ascertainment bias leaves many footprints in the genetic data of the participants. Exploiting that, we are able to perform a GWAS on participation using only genetic data of the participants. Notably, each individual often harbours both 'case' and 'control' alleles, at different sites and the same sites. With nearly perfectly matched controls, our GWAS only captures direct effects and is unaffected by population stratification confounding. While the genetic component to participation can only be studied through a genetic study, the former can play a role in all sampling-based studies, genetic or otherwise.

One of the complications of studying the genetics of participation is the issue of replication. As study design, participation rate and the contribution of genetics can vary hugely from study to study, each study has to be evaluated on its own and the participation GWAS results cannot be expected to replicate across all studies. For this reason, we validated our participation GWAS through a special form of within-study replication. This was done by investigating the relationship between the *pPGS* and various phenotypes in a group of UKBB participants that do not overlap with the UKBB first-degree relatives that the participation GWAS is based on. Notably, the *pPGS* is constructed from a GWAS performed without phenotypes. Thus, its associations with phenotypes such as EA and BMI, that have previously been reported to exhibit ascertainment bias[17,20], is confirmation that our GWAS is indeed capturing the genetic effects of participation.

The results of this study demonstrate that there can be a common component that underlies many different participation events. The *pPGS* we constructed is based on information about primary participation, but it plays a role in both the passive (being invited) and active (deciding to participate when invited) phases of the secondary participation events. Thus, the relevance of a genetic component underlying participation should not be judged based on an individual

selection event only. Its effect could accumulate through its impact on many participation events of a person's lifespan, or it can be magnified through nested participation, *e.g.* the participants in the dietary and physical activity studies have higher average genetic propensity to participate than the other UKBB participants, who already have higher average propensities than the population. In contrast to a common tendency to think of participation as a consequence of other characteristics and established traits, we propose that the propensity to participate in a whole range of events should be studied as a behavioural trait in its own right. A better understanding of this behaviour could lead to better recruitment schemes, for studies genetic and otherwise, that increases participation rate and have lower ascertainment bias.

**Acknowledgements**

A.K. is supported by the Li Ka Shing Foundation. S.B. is supported by the Li Ka Shing Foundation and the Goodger and Schorstein scholarship. This research has been conducted using the UK Biobank Resource under application number 11867. The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

**Author contributions**

Idea: A.K.
Methodology: A.K. and S.B.
Design of study: A.K. and S.B.
Investigation: A.K. and S.B.
Data application: A.K and S.B.
Writing – original draft: A.K. and S.B.
Writing – review and editing: A.K. and S.B.

**Competing interests**

Authors declare that they have no competing interests.

**Supplementary Materials**

Figs. S1 to S9
Tables S1 to S4
Supplementary Text

## References


1 Bradley, V. C. *et al.* Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 1-6, doi:10.1038/s41586-021-04198-4 (2021).

2 Barnes, P. Reality Check: Should We Give Up on Election Polling? *BBC News* (2016). < http://www.bbc.com/news/election-us-2016-37949527>.

3 Meng, X.-L. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics* **12**, 685-726 (2018).

4 Tyrrell, J. *et al.* Genetic predictors of participation in optional components of UK Biobank. *Nature communications* **12**, 1-13, doi:10.1038/s41467-021-21073-y (2021).

5 Taylor, A. E. *et al.* Exploring the association of genetic factors with participation in the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* **47**, 1207-1216, doi:10.1093/ije/dyy060 (2018).

6 Martin, J. *et al.* Association of Genetic Risk for Schizophrenia With Nonparticipation Over Time in a Population-Based Cohort Study. *Am J Epidemiol* **183**, 1149-1158, doi:10.1093/aje/kww009 (2016).

7 Adams, M. J. *et al.* Factors associated with sharing e-mail information and mental health survey participation in large population cohorts. *International journal of epidemiology* **49**, 410-421, doi:10.1093/ije/dyz134 (2020).

8 Pirastu, N. *et al.* Genetic analyses identify widespread sex-differential participation bias. *Nat Genet* **53**, 663-671, doi:10.1038/s41588-021-00846-7 (2021).

9 Kong, A. *et al.* The nature of nurture: Effects of parental genotypes. *Science* **359**, 424-428, doi:10.1126/science.aan6877 (2018).

10 Young, A. I. *et al.* Mendelian imputation of parental genotypes for genome-wide estimation of direct and indirect genetic effects. *BioRxiv*, doi:10.1101/2020.07.02.185199 (2020).

11 Kong, A., Benonisdottir, S. & Young, A. I. Family analysis with Mendelian imputations. *BioRxiv*, doi:10.1101/2020.07.02.185181 (2020).

12 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909, doi:10.1038/ng1847 (2006).

13 Ewens, W. J. & Spielman, R. S. The transmission/disequilibrium test: history, subdivision, and admixture. *Am J Hum Genet* **57**, 455-464 (1995).

14 Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873, doi:10.1093/bioinformatics/btq559 (2010).

15 Young, A. I., Benonisdottir, S., Przeworski, M. & Kong, A. Deconstructing the sources of genotype-phenotype associations in humans. *Science* **365**, 1396-1400, doi:10.1126/science.aax3710 (2019).

16 Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209, doi:10.1038/s41586-018-0579-z (2018).

17 Fry, A. *et al.* Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *American journal of epidemiology* **186**, 1026-1034, doi:10.1093/aje/kwx246 (2017).

18 Doherty, A. *et al.* Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. *Plos One* **12**, e0169649, doi:10.1371/journal.pone.0169649 (2017).

19 Bradbury, K. E., Young, H. J., Guo, W. & Key, T. J. Dietary assessment in UK Biobank: an evaluation of the performance of the touchscreen dietary questionnaire. *J Nutr Sci* **7**, e6, doi:10.1017/jns.2017.66 (2018).

20 Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M. & Davey Smith, G. Collider scope: when selection bias can substantially influence observed associations. *International journal of epidemiology* **47**, 226-235, doi:10.1093/ije/dyx206 (2018).