Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2013;42:1145–1156
doi:10.1093/ije/dyt095

Differences between blood donors and a population sample: implications for case–control studies

Jean Golding,[1]* Kate Northstone,[1] Laura L Miller,[1] George Davey Smith[2] and Marcus Pembrey[1,3]

[1]School of Social and Community Medicine, University of Bristol, Bristol, UK, [2]MRC Centre for Causal Analyses in Translational Epidemiology, University of Bristol, Bristol, UK, and [3]Institute of Child Health, University College London, London, UK

*Corresponding author. Centre for Child and Adolescent Health, School of Social and Community Medicine, University of Bristol, Barley House, Oakfield Grove, Bristol BS8 2BN, UK. E-mail: jean.golding@bristol.ac.uk

Accepted 29 April 2013

Background Selecting appropriate controls for studies of genetic variation in case series is important. The two major candidates involve the use of blood donors or a random sample of the population.

Methods We compare and contrast the two different populations of controls for studies of genetic variation using data from parents enrolled in the Avon Longitudinal Study of Parents and Children (ALSPAC). In addition we compute different biases using a series of hypothetical assumptions.

Results The study subjects who had been blood donors differed markedly from the general population in social, health-related, anthropometric, and personality-related variables. Using theoretical examples, we show that blood donors are a poor control group for non-genetic studies of diseases related to environmentally, behaviourally, or socially patterned exposures. However, we show that if blood donors are used as controls in genetic studies, these factors are unlikely to make a major difference in detecting true associations with relatively rare disorders (cumulative incidence through life of <10%). Nevertheless, for more common disorders, the reduction in accuracy resulting from the inclusion in any control population of individuals who have or will develop the disease in question can create a greater bias than can socially patterned factors.

Conclusions Information about the medical history of a control and the parents of the control (as a proxy for whether the control will develop the disease) is more important with regard to the choice of controls than whether the controls are a random population sample or blood donors.

Keywords ALSPAC, blood donors, genetic studies, case-control studies, methodology

Introduction

Choice of the optimum control group is important for genetic, biochemical, or other assay-based association studies involving case collections of a particular disease. For example, at a time when case series may test 500 000 genetic variants and many thousands of values of data relating to gene-expression and/or metabolomic, epigenetic, or proteomic factors simultaneously, with the inherent cost of this, there is the temptation to use a multi-purpose set of controls so that, once typed, they can be compared with a variety

of different case series. A technically easy option is to use blood donors. Many examples of the use of blood donors as controls in the study of genetic variants[1–3] and other blood markers[4–6] exist in the literature. However, in using blood donors as controls, it is tacitly assumed that blood donors are representative of the population from which are drawn the case series for which they are to serve as controls. We investigated the differences in a population of prospective parents enrolled between 1990 and 1992 in the Avon Longitudinal Study of Parents and Children (ALSPAC), a study designed to determine how interactions between genotypes and environmental factors influence health and development, to identify differences between those who had donated blood and those who had not.

In Britain, the published guidelines of 1989 for blood transfusion services in the United Kingdom[7] state that blood donors should appear healthy and range in age from 17–65 years; their medical history should be assessed to identify persons who should be permanently excluded (e.g. persons with circulatory disorders, drug allergy, diseases of the central nervous system, diseases of possible viral or autoimmune origin, chronic renal disease, and significant chest disease); among those to be temporarily excluded were persons who had recently had an infectious disease (for whom the time of exclusion varied from 2 years for infectious mononucleosis to the end of a course of antibiotic therapy for a sore throat), a minor or major accident (exclusion for 3–6 months), individuals taking most medications, and pregnant individuals (excluded until 1 year after delivery), among many others. The guidelines also specified that on presentation for blood donation, an individual's haemoglobin (Hb) level was to be measured and that women with an Hb <12.5 g/dL and men with an Hb <13.5 g/dL were to be excluded from blood donation. Individuals weighing <41 kg were also to be excluded.

Given such permanent and temporary exclusions, as well as confounding by other factors related to becoming a blood donor, it would be expected that blood donors would be less prone than average to infections, autoimmune diseases, asthma, and allergies to drugs, and would be less likely to be smokers. Additionally, women who had had a number of pregnancies would be less likely to be blood donors. In our study we assessed social, medical, and familial differences between persons who had and those who had not given blood.

Methods

ALSPAC began by enrolling pregnant women resident in Avon, UK, who were willing to take part in the study (an estimated 80–85% of the eligible population).[8] Criteria for eligibility were an expected date of delivery between 1 April 1991 and 31 December 1992. During pregnancy, prospective mothers in ALSPAC received a variety of different questionnaires that they completed in their own homes and posted back to the study headquarters. The information, although anonymised, has been linked to other information from the same mothers and their partners. The partners were contacted via the mothers, who were asked to invite them to take part in ALSPAC and to give them a questionnaire to complete if they were willing to do so. In all, 14 541 women completed at least one questionnaire, and more than 10 000 of their partners did so.[9] For the present analysis, we restricted the sample to those of white ethnicity aged 18 years and older.

Members of this study sample were asked, during pregnancy, whether they had ever donated blood, to which 29.4% of 12 350 mothers and 34.1% of the 8426 partners responded positively. Categorisation of parental education was based on the highest level of academic achievement (with five categories ranging from no education to a university degree or equivalent); social class of their current or most recent occupation (with five categories[10] ranging from I, for higher professional, to IV/V, for semi-skilled and unskilled manual); place of residence at the time of the respondent's birth [Avon (the study area) or not]; and parity (the number of previous pregnancies resulting in a live or still-birth).

Medical histories of the individuals in the study sample, and of their natural parents, were obtained through specific questions.[9] In addition, the individuals were asked for their perception of their own health (using a four-point scale with the categories of always well; usually well; sometimes unwell; often unwell). The women in the study sample were asked to make this assessment for the time before they became pregnant. The personality of each individual in the sample was assessed on the basis of the answers to 36 questions from the Boyce and Parker Interpersonal Sensitivity Scale[11] for the five domains of interpersonal awareness, need for approval, separation anxiety, timidity, and fragility of the inner self. An individual participant's height and weight (before pregnancy for prospective mothers) were those reported by the individual (a woman's reported height was more likely than a man's to be accurate because it would have been measured during her prenatal care).

Given the age limits on blood donation in the 1989 UK guidelines, and taking into account the probability that women who were 17 years old at delivery would have been too young to be donors or would already have been pregnant and therefore excluded from blood donation, we confined our analyses to persons aged 18 years and older. Only 4–5% of the individuals enrolled in the study were non-white. This group was less likely to have donated blood (16.8% of non-white women and 25.4% non-white men had ever donated blood, compared with 29.4% and 34.1% of white

women and men respectively); the numbers were too small for valid analysis by the time the groups were split into specific ethnic minority groups. We therefore restricted our analysis to the white European population. In terms of our findings, this makes any differences between people who do and do not donate blood a conservative estimate, although many genetic epidemiology studies would also be likely to restrict cases and controls to members of a single ethnic group.

Statistical analyses

For unadjusted analyses, chi-square tests and logistic regression were used with categorical variables, and *t*-tests with continuous variables using SPSS v.12.0.1 (SPSS Inc., Chicago, IL, USA). Binomial probability tests using STATA v. 9.2 were employed to assess the number of values of $P < 0.05$ observed as compared with the number of such values expected if we assumed that the null hypotheses were valid. For these two groups (the 24 parental medical histories and the 52 associations with medical histories) we additionally used a cut-off value of $P < 0.001$. We did not do a correction for multiple testing because our analysis was conducted with the purpose of looking at the number of nominally statistically significant differences compared with the expected number, rather than identifying those factors that were least likely to have occurred by chance.

Results

Demographic effects

Table 1 shows the prevalence of blood donation for both parent groups. There was a strong relationship

Table 1 Prevalence and unadjusted odds ratio (OR) of having given blood by selected demographic factors:

	Women		Men	
	% (*n*) blood Donors	**OR [95% CI]**	**% (*n*) blood donors**	**OR [95% CI]**
Highest educational qualification				
⩽ CSE (lowest)	15.0% (303)	1.00 Reference	24.0% (297)	1.00 Reference
Vocational	20.9% (223)	1.49 [1.23, 1.81]	25.1% (172)	1.06 [0.85, 1.31]
O level	29.4% (1162)	2.35 [2.04, 2.71]	31.7% (835)	1.47 [1.26, 1.72]
A level	40.2% (1042)	3.80 [3.29, 4.40]	42.6% (747)	2.35 [2.00, 2.76]
University degree (highest)	47.7% (705)	5.15 [4.40, 6.04]	50.0% (557)	3.16 [2.65, 3.76]
	***P*<0.0001**		***P*<0.0001**	
Social class (based on mother's occupation)				
IV, V Manual semi-skilled	21.6% (232)	1.00 Reference	26.1% (183)	1.00 Reference
III Manual	19.2% (134)	0.86 [0.68, 1.09]	28.2%(122)	1.11 [0.85, 1.45]
III Non-manual	30.2% (1213)	1.58 [1.34, 1.85]	33.0% (890)	1.40 [1.16, 1.68]
II Professional	40.6% (1196)	2.47 [2.10, 2.91]	43.1% (906)	2.14 [1.71, 2.59]
I Higher professional	51.3% (284)	3.81 [3.05, 4.75]	51.8% (212)	3.05 [2.36, 3.94]
	***P*<0.0001**		***P*<0.0001**	
Parity of mother (no. of previous births)				
0	32.7% (1575)	1.50 [1.23, 1.82]	35.6% (1231)	1.18 [0.93, 1.49]
1	30.9% (1206)	1.38 [1.13, 1.69]	34.7% (872)	1.13 [0.89, 1.44]
2	29.1% (446)	1.27 [1.02, 1.57]	37.0% (365)	1.25 [0.97, 1.62]
3+	24.5% (145)	1.00 Reference	31.9% (114)	1.00 Reference
	***P* < 0.001**		***P* = 0.533**	
Place of birth				
Avon	25.6% (1467)	1.00 Reference	29.2% (1083)	1.00 Reference
Not-Avon	37.1% (1933)	1.71 [1.58, 1.86]	41.1% (1588)	1.69 [1.53, 1.86]
	***P* < 0.0001**		***P* < 0.0001**	

between blood donation and level of academic achievement, with a higher level of academic achievement correlating positively with a greater likelihood of ever having given blood. With regard to social class, parents in the manual semi-skilled (IV) and unskilled (V) groups were considerably less likely to have given blood than those in the non-manual groups. For women, increasing parity was inversely correlated with a likelihood of having previously given blood. However, parity was not associated with womens' partners' likelihood of having given blood. Parents who had been born in Avon (the study area) were much less likely to have given blood than those who had been born elsewhere but were currently resident in Avon.

Health factors

Table 2 compares anthropometric and personality measures of blood donors and non-donors. Individuals who were blood donors were on average taller and heavier than those who never donated. Personality traits also showed differences in mean scores for both sexes between those who had and had not given blood for interpersonal awareness, need for approval, timidity, and fragile inner self, although women showed a smaller effect size for each trait than did men (score differences of 0.28, 0.39, 0.34, and 0.12 vs. 0.91, 0.74, 0.69, and 0.39, respectively).

In general, the more unwell the study participants rated themselves, the less likely they were to have donated blood (Table 3). When analysed by medical condition (Appendix Table 1), women who had a history of eczema or allergies were slightly more likely to have given blood, as were men with a history of allergies, indigestion, and haemorrhoids. Both women and men were more likely to have given blood if they had less than perfect eyesight, as were men if they had poor hearing (data not shown). In contrast, a reduced likelihood of being a blood donor was evident for hay fever (women), epilepsy (both sexes), febrile convulsions (men only), migraine (women), kidney disease (women), drug addiction (both), and severe depression (women only). Men with alcoholism were also less likely to have given blood, but the numbers in this comparison were small. In all, of the 52 analyses, including the data on hearing, 18 (35%) were statistically significant at the conventional level of

Table 2 Variation in mean [SD] of continuous traits: blood donors (BD) *vs.* non-blood donors (NBD)

	Women			Men		
Trait	BD	NBD	*P*	BD	NBD	*P*
Anthropometry						
Height (cm)	165.0 [6.46]	163.7 [6.73]	<0.0001	176.6 [6.65]	175.9 [7.01]	<0.001
Weight (kg)	63.11 [10.3]	61.29 [11.2]	<0.0001	78.91 [11.3]	77.87 [11.67]	<0.001
Personality						
Interpersonal awareness	18.54 [4.42]	18.26 [4.80]	0.004	16.78 [4.70]	15.87 [4.91]	<0.0001
Need for approval	26.14 [3.15]	25.75 [3.70]	<0.0001	25.09 [3.64]	24.35 [4.48]	<0.0001
Separation anxiety	16.07 [4.48]	16.48 [4.76]	<0.0001	14.74 [4.19]	14.67 [4.39]	0.479
Timidity	20.84 [4.31]	20.50 [4.61]	<0.001	19.29 [4.49]	18.60 [4.77]	<0.0001
Fragile inner self	8.82 [2.86]	8.70 [2.99]	0.036	8.46 [2.75]	8.07 [2.68]	<0.0001

Table 3 Proportion and unadjusted odds ratio (OR) of blood donors among women and their partners by their perception of their health prior to the study pregnancy

	Women		Men	
	% (*n*) blood donors	**OR [95% CI]**	**% (*n*) blood donors**	**OR [95% CI]**
Always well	34.8% (1149)	2.32 [1.32, 4.08]	32.0% (1604)	1.23 [0.67, 2.23]
Usually well	30.7% (1898)	1.92 [1.09, 3.37]	30.8% (1631)	1.16 [0.64, 2.11]
Sometimes unwell	19.2% (126)	1.03 [0.57, 1.87]	29.3% (113)	1.08 [0.57, 2.03]
Often unwell	18.8% (15)	1.00 Reference	27.8% (15)	1.00 Reference
	P_T[1] <0.0001		P_T[1] =0.097	

[1] P_T – Result of test for linear trend

$P < 0.05$, whereas only 2.6 would have been expected ($P < 0.0001$); eight (15%) values of $P < 0.001$ were observed, as compared with 0.052 expected ($P < 0.0001$).

In regard to the medical history of their natural parents (Appendix Table 2) 12 conditions were assessed: 10 (37%) were different at the 0.05 level including five (21%) at the 0.001 level. In contrast only 1.2 and 0.024 such differences, respectively, would have been expected by chance ($P < 0.0001$). Individuals were more likely to have given blood if their parents had a history of coronary heart disease (both men and women), arthritis (women), cancer (both), hypertension (both), and type II diabetes (women). Men were less likely to have given blood if a parent had a history of alcoholism.

Hypothetical computation of likelihood of detecting real associations in case–control studies

Having shown that the likelihood of an individual giving blood varied with the individual's background, medical history, and family history, we sought to assess the effect that this might have on case–control studies of the association of these variables with disease. Table 4 shows the numbers of study participants with a candidate genetic marker that one might expect under various assumptions. In each instance, 1000 cases of disease are compared with 1000 controls for 3 possible frequencies of genotype. The assumption throughout is that the true odds ratio (OR) of the disease being associated with the genotype is 2.0. We compared the performance of three types of controls according to a variety of hypothetical assumptions: (i) the optimal control (representative of the population but omitting the individuals who have had, currently have, or will ultimately have the disorder); (ii) a random group (representative of the general population); and (iii) blood donors.

Optimal controls

Assuming that the population of 1000 controls have not had, do not have, and never will have the disease for which a genotype association is being investigated, the numbers of controls who will have a genotype of frequency (*f*) 10%, 30%, and 50% will be 100, 300, and

Table 4 Odds ratios of genotype association with cases for different sets of controls, where the 'true' odds ratio was 2.0 (number of individuals in control samples shown in parentheses)

	Frequency of Disease in Population	Frequency of Genotype in Non-Affected[a]		
		10%	30%	50%
Cases	100%	181	461	667
Optimal Controls[b]		2.0 (100)	2.0 (300)	2.0 (500)
Random controls[c]				
	50%	1.35 (141)	1.39 (381)	1.42 (584)
	30%	1.56 (124)	1.60 (348)	1.64 (550)
	10%	1.83 (108)	1.85 (316)	1.87 (517)
Differential blood donation A[d,e]				
	50%	1.21 (154)	1.25 (407)	1.27 (611)
	30%	1.39 (137)	1.43 (374)	1.47 (577)
	10%	1.70 (115)	1.74 (329)	1.77 (530)
Differential blood donation B[d,f]				
	50%	1.52 (127)	1.56 (354)	1.60 (556)
	30%	1.72 (114)	1.75 (328)	1.78 (529)
	10%	1.90 (104)	1.92 (308)	1.93 (509)

[a]Assuming 1000 cases are being compared with 1000 controls and that cases are twice as likely to have the genotype than controls with no disease.
[b]Optimal controls are individuals who do not have, and never will have, the disease.
[c]Random controls are members of the general population, no attention being paid to whether they have the disease or not.
[d]Blood donors are a sub sample of the general population with various biases.
[e]Assumes that of individuals with the history of the disease (in past, present or future), 40% will give blood but amongst those with no such history, only 20% will do so.
[f]Assumes that of individuals with a history of the disease (in past, present or future), 20% will give blood but amongst those with no such history, 40% will do so.

500, respectively. From this, and an OR of 2.0, one can calculate the numbers of individuals (x) who will have the genotype according to each frequency. For example, if $f = 30\%$, then

$$2.0 = [x/300] \times [700/(1000 - x)]$$

which gives $6000 - 6x = 7x$, or $x = 6000/13 = 461$

This method of calculation gives the number of cases shown in the first line of Table 4.

Random controls

Calculations for control samples in analyses that include cases (either in the past, present, or future) of a disease are calculated as follows: If d is the number of individuals in the control group with the disease (past, present, or future) for which a genetic association is being examined, then $(1000 - d)$ do not have the disease. If it is assumed that p_d and p_n are the proportions of individuals with and without the disease who are in the control sample, and that q_d and q_n are the proportions of individuals with and without the disease who have the candidate genotype, then the number in the control sample expected to have the genotype will be

$$e = 1000 \times [dp_dq_d + (1000 - d)p_nq_n]/ [dp_d + (1000 - d)p_n]$$

and the OR is given by

$$\mathrm{OR} = \{c/(1000 - c)\} \times \{(1000 - e)/e\}$$

where c is the number of cases with the genotype.

With this method, the numbers of cases of each genotype and disease frequency in the random controls can be calculated. The results are shown in Table 4. Note that for each genotype frequency the number of cases increases with the frequency of the disease in the population; conversely, the apparent OR falls with the frequency of disease. However for each disease frequency, the OR varies only marginally with genotype frequency.

It is unlikely that a genetic-association study based on a random control sample would be able to exclude controls who have, have had, or will have the disease for which the study is being conducted.

Blood donors as controls

For blood donors we examined two scenarios: (i) when more donors who have had or are destined to have the disease for which a genetic association study is being done give blood (40% *vs*. 20%); and (ii) when such individuals are less likely to donate blood (20% *vs*. 40%). Calculation of the number of cases involved will be similar to that used for the random controls described above. The results are compared in Table 4.

Blood donors show larger than expected numbers of the genotype for which a disease-association study is being done if there is an increase in donations by individuals with the disease (e.g. with a gene frequency of 10% the numbers of affected controls increase from 115 to 154, with a corresponding reduction in ORs from 1.70 to 1.21). For the second scenario, in which the affected individuals are less likely to donate blood, the ORs become larger (from 1.52 to 1.90) as the number of donors with the disease decreases. Similar changes are shown with different frequencies of the genotype in non-affected individuals (Table 4).

Controls biased demographically

If it is assumed that a disorder, K, is more prevalent in shorter individuals, and has prevalences of 4%, 3%, 2%, and 1% in the four quartiles of height in the general population, the case frequency of K will be 40%, 30%, 20%, and 10% of each quartile. Accordingly, in a population of 10 000 individuals, the numbers of those with the disease in the four height quartiles will be 4000, 3000, 2000, and 1000, respectively. Comparison of K with random controls should reveal this.

Suppose also that there is an important genetic variant that is related to height and is present in 30%, 24%, 18%, 12% of each height quartile. Then the case collection K of 10 000 individuals should include 2400 individuals (30% of 4000 + 24% of 3000 + 18% of 2000 + 12% of 1000) = (1200 + 720 + 360 + 120) with the genetic variant.

This frequency, when compared with 2100 (30% of 2500 + 24% of 2500 + 18% of 2500 + 12% of 2500) = (750 + 600 + 450 + 300) random controls, gives a true risk ratio of 1.14. However, with the strong bias among blood-donor controls towards taller individuals, there is little chance of showing the true effect. Conversely, if the true effect is null but the genetic trait under consideration is linked positively to height, the effect with blood donors will give erroneously significant results, whereas that with the random controls will give the true null effect.

Discussion

We believe that the present study is the first to attempt to define the differences between blood donors and non-donors using a large data set. Epidemiologists have long been aware of the need for appropriate controls for any study of cases, but this is often the most difficult aspect of study design. In studies comparing genetic or other blood markers, there is the added complexity of taking blood. It is therefore tempting to use, as controls, samples from blood donors that have been routinely collected, albeit for a different purpose. This strategy has been widely used for studies of genetic[1–3] as well as cytogenetic[12] and other blood[4–6] markers.

In this study we compare individuals who have donated blood with those who have not within a population resident in Avon, UK, and who were

willing to take part in a longitudinal cohort study. Although reasonably representative of the pregnant Avon population[9] in the early 1990s, the ALSPAC population does not include infertile individuals (whether voluntary or involuntary) or the older population. The study should therefore not be considered to provide a comprehensive view of biases among blood-donors.

Within the ALSPAC study, we have shown that adults who give blood differ in social class, cognitive ability (using educational achievement as a proxy), personality, anthropometry, their medical history and the medical histories of their parents. Our data are similar to those in the literature, showing differential uptake according to occupation[13] and higher levels of education.[14] As shown in other studies, proportionately fewer women give blood than men, and ethnic minorities tend to have a lower rate of blood donation than majority populations. Because criteria for blood donation vary across the world, different biases may occur. Unexpectedly, we found that individuals who were born and still resident in the study area were less likely to be blood donors than those who had moved into the area. If true nationally, this may have implications for genetic studies where matching is done by area.

There is some evidence that blood donors tend to have low self-esteem and give blood in order to improve their self-esteem.[15] We have been able to examine the personalities of the individuals who were blood donors, and have shown that they had different personality traits, particularly in regard to the 'Need for approval' scale, which reflects a wish to make others happy and to ensure that others will like and not reject them.

There is also a strong unexpected association of blood donation with a family history of several common chronic diseases (such as coronary heart disease) in which genetic factors are known to play a part. This bias in blood donors suggests that there are, or could be, more genotype differences between cases and controls than expected by chance. Few studies have compared genetic markers in different types of controls, and those studies that have done this were small and gave negative results.[16] The analysis of the Wellcome Trust Case Control Consortium compared 1500 controls from the 1958 British National Cohort Study with 1500 blood-donor controls using a genome wide array and showed very few differences between markers, although there were more significant outliers than would be expected by chance.[17] Thus, in general, the distribution of genetic variants among blood donors shows fewer differences with non-donors than is found for environmental factors.

Hypothetical calculations

The take-home messages from Table 4 are that when a real association exists:

(i) The frequency of a genotype in the non-affected population makes little difference in terms of the sizes of the ORs that will be demonstrated, whatever the control type.
(ii) If a family history of the condition, the condition itself, or the candidate genotype are associated with an increased likelihood of the individual being a blood donor, the ORs that can be demonstrated will be reduced in comparisons with random controls, and vice versa.
(iii) The ORs depend on the proportion of affected individuals included among the controls; consequently, the rarer a condition, the less impact control bias will have.
(iv) For common disorders, ORs for associated genotypes are likely to be greatly underestimated.

Thus, we have shown in this hypothetical example that when the true OR of a genotype-related disease for cases as compared with controls is 2.0, the comparison with random or blood-donor controls approaches 2.0 only if the proportion of the population who have ever had or will get the disease is no more than 10%. Conversely, where there is no real association of a disease with a genetic variant (i.e. OR = 1.00), the biases inherent in controls may produce a spurious relationship, although this will be of relatively small magnitude (data not shown).

Obviously the ideal design for a case–control study (i.e. the selection of optimal controls) is not generally feasible unless historic controls are used, with full details of their medical histories until the time of death, so that individuals with the disorder being examined in the study can be excluded. Even with this, however, one would have to take into account that individuals dying early may have developed the disease had they survived, and the same exclusion would have to have been likely in the case series (i.e. individuals would have died before becoming cases of the disease). One way around this involves the definition of a 'case' by using the diagnosis of a disease within a specific age window, in which instance the control sample can validly consist of individuals in whom the diagnosis was not made within that window.

From the general computations shown here, it is clear that both random and blood-donor controls can be suitable for detecting real associations of genotypes with relatively rare disorders (e.g. schizophrenia or Hodgkin's lymphoma) for which an onset before middle age ensures that a population of 50-year-old persons, for example, would not be likely to develop the disorder in the future. These controls may not be satisfactory for particularly common disorders, especially those of later onset (e.g. hypertension, type II diabetes) unless there is information on the individuals' family history of the disorder together with their own medical history. Exclusion of such individuals would allow more accurate estimation of true genetic effects, even though some of the individuals without a family history of the disorder may develop the

disorder in the future. For rarer disorders there is no serious problem with controls selected in such a way that these exclusions cannot be made.

Choice of controls should ideally be based on the rule that they should be chosen from the same population as the cases, the only distinguishing feature being 'caseness' (in whatever way that is defined and ascertained). The case series that have been developed, however, often do not have any coherent definition of the baseline population from which the cases have been selected, nor how representative of the disorder the cases are in general. Thus there may be major biases both in the individual cases as well as in their controls.

From these considerations we believe that the key message emanating from our study is that control samples chosen as representative of the population may not be appropriate as comparison populations for common diseases unless they include information that will allow the exclusion of individuals who have developed or will develop the disease. There are also likely to be biases among cases that differ substantially from the biases within a control sample. Consequently, control populations without such basic data will generally not be appropriate for the study of social, behavioural, and environmental exposures or for the study of gene-by-environment interactions, or, indeed, for studies of DNA methylation status where the environment is known to have strong effects.[18] A powerful design would select cases and controls enrolled in the same (prospective) study. This would be feasible in large studies, such as the UK Biobank study.[19] However, other issues, such as the late average age of onset of cases in such cohorts, and the often lower genetic contribution to disease risk at older ages, the long time lag to onset, and the high costs involved need to be considered. For smaller detailed studies (e.g. the 1958 cohort),[20] the duration of follow up, the accuracy of data, and the study of relevant traits within the population will be likely to provide as much (or more) information of value for common diseases, and especially for quantitative disease traits. It is tempting to use control samples stripped of personal information, since this overcomes the challenging and costly issue of protecting confidentiality, but this may be a false economy in the study of some aspects of common disease aetiology.

Funding

Collection of the data used in this paper was funded by a variety of contributors including UK Government departments, charities and research councils amongst others. The UK Medical Research Council, the Wellcome Trust and the University of Bristol currently provide core support for ALSPAC.

Acknowledgements

We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses.

Conflict of interest: None declared.

References

[1] Williams NM, Preece A, Morris DW *et al*. Identification in 2 independent samples of a novel schizophrenia risk haplotype of the dystrobrevin Binding Protein Gene (DTNBP1). *Arch Gen Psychiatry* 2004;**61:**336–44.

[2] Bellamy R, Ruwende C, Corrah T. Tuberculosis and chronic hepatitis B virus infection in Africans and variation in the Vitamin D Receptor Gene. *J Infect Dis* 1999; **179:**721–24.

[3] Calhoun ES, McGovern RM, Janney CA. Host genetic polymorphism analysis in cervical cancer. *Clin Chem* 2002;**48:**1218–24.

[4] Holtmann G, Talley NJ, Mitchell H. Antibody response to specific H. pylori antigens in functional dyspepsia, duodenal ulcer disease, and health. *Am J Gastroenterol* 1998; **93:**1222–26.

[5] McGill S, Wesslen L, Hje E. Serological and epidemiological analysis of the prevalence of Bartonella spp. antibodies in Swedish elite orienteers 1992-93. *Scand J Infect Dis* 2001;**33:**423–28.

[6] Winston AP, Jamieson CP, Madira W *et al*. Prevalence of thiamin deficiency in anorexia nervosa. *Int J Eat Disord* 2000;**28:**451–54.

[7] Department of Health. *Guidelines for the Blood Transfusion Services in the United Kingdom. Volume I*. London: Her Majesty's Stationery Office, 1989.

[8] Fraser A, Macdonald-Wallis C, Tilling K *et al*. Cohort profile: the Avon Longitudinal Study of parents and Children: ALSPAC mothers cohort. *Int J Epidemiol* 2013; **42:**97–110.

[9] Avon Longitudinal Study of Parents and Children; http://www.bristol.ac.uk/alspac (27 June 2013, date last accessed).

[10] Office of Population Censuses and Surveys. Standard Occupational Classification. London: Her Majesty's Stationery Office, 1991.

[11] Boyce P, Parker G. Development of a scale of interpersonal sensitivity. *Aust N Z J Psychiatry* 1989;**23:**341–51.

[12] Zhang ZF, Morgenstern H, Spitz MR. Environmental tobacco smoking, mutagen sensitivity, and head and neck squamous cell carcinoma. *Cancer Epidemiol Biomarkers Prev* 2000;**9:**1043–49.

[13] Houston DJ. ''Walking the Walk'' of public service motivation: Public employees and charitable gifts of time, blood, and money. *J Publ Adm Res Theor* 2006;**16:** 67–86.

[14] Burnett JJ. Psychographic and demographic characteristics of blood donors. *J Consum Res* 1981;**8:**62–66.

[15] Fernández-Montoya A, López-Berrio A, Luna del Castillo JD. How some attitudes, beliefs and motivations of Spanish blood donors evolve over time. *Vox Sang* 1998; **74:**140–47.

[16] Rafii S, O'Regan P, Xinarianos G. A potential role for the XRCC2 R188H polymorphic site in DNA-damage repair and breast cancer. *Hum Mol Genet* 2002;**11:**1433–38.
[17] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* 2007;**447:** 661–78.
[18] Relton CL, Davey Smith G. Is epidemiology ready for epigenetics? *Int J Epidemiol* 2012;**41:**5–9.
[19] Watts G. Genes on ice. *BMJ* 2007;**334:**662–3.
[20] Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int J Epidemiol* 2006;**35:**34–41.

Appendix Table 1 Proportion and unadjusted odds ratios (OR) of women and their partners who had donated blood by their medical history

	Women			Men		
Medical condition	**% (*n*) blood donors**	**OR [95% CI]**	***P***	**% (*n*) blood donors**	**OR [95% CI]**	***P***
Any allergies						
Yes	32.1% (1536)	1.11 [1.02, 1.20]		37.9% (993)	1.21 [1.10, 1.34]	<0.0001
No	30.0% (1862)	1.00 Reference	0.015	33.5% (1578)	1.00 Reference	
Hay fever						
Yes	29.4% (1007)	1.00 Reference		36.7% (822)	1.10 [1.00, 1.23]	0.060
No	31.5% (2433)	1.10 [1.01, 1.21]	0.027	34.4% (1725)	1.00 Reference	
Indigestion						
Yes	31.0% (2390)	1.03 [0.95, 1.12]		36.2% (1692)	1.13 [1.02, 1.25]	0.016
No	30.4% (1050)	1.00 Reference	0.499	33.4% (881)	1.00 Reference	
Bulimia						
Yes	34.8% (92)	1.12 [0.93, 1.56]		28.1% (9)	1.00 Referencee	0.419
No	30.7% (3348)	1.00 Reference	0.154	35.0% (2361)	1.37 [0.64, 2.97]	
Asthma						
Yes	29.0% (367)	1.00 Reference		35.3% (334)	1.01 [0.87, 1.16]	0.925
No	31.1% (3073)	1.10 [0.97, 1.25]	0.140	35.1% (2256)	1.00 Reference	
Eczema						
Yes	32.6% (838)	1.12 [1.01, 1.23]		36.4% (371)	1.07 [0.93, 1.22]	
No	30.3% (2602)	1.00 Reference	0.024	34.9% (2197)	1.00 Reference	0.360
Epilepsy						
Yes	15.9% (20)	1.00 Reference		23.0% (17)	1.00 Reference	
No	31.0% (3420)	2.38 [1.48, 3.85]	<0.001	35.3% (2583)	1.83 [1.06, 3.15]	0.027
Febrile convulsions						
Yes	27.2% (59)	1.00 Reference		25.0% (37)	1.00 Reference	
No	30.9% (3337)	1.20 [0.89, 1.62]	0.241	35.4% (2545)	1.64 [1.13, 2.39]	0.009
Migraine						
Yes	28.8% (1398)	1.00 Reference		33.7% (689)	1.00 Reference	
No	32.4% (2042)	1.19 [1.09, 1.29]	<0.0001	35.7% (1877)	1.09 [0.98, 1.22]	0.109
Back pain						
Yes	31.5% (1636)	1.06 [0.98, 1.15]		36.2% (1263)	1.09 [0.98, 1.20]	
No	30.2% (1804)	1.00 Reference	0.136	34.2% (1338)	1.00 Reference	0.072
Kidney disease						
Yes	26.9% (136)	1.00 Reference		31.8% (49)	1.00 Reference	
No	31.0% (3304)	1.22 [1.00, 1.50]	0.048	35.3% (2550)	1.17 [0.83, 1.64]	0.375
Varicose veins						
Yes	30.9% (402)	1.00 [0.89, 1.14]		38.6% (81)	1.16 [0.88, 1.54]	
No	30.8% (3038)	1.00 Reference	0.946	35.1% (2521)	1.00 Reference	0.297
Haemorrhoids						
Yes	32.0% (1177)	1.08 [0.99, 1.18]		41.1% (638)	1.39 [1.24, 1.56]	
No	30.3% (2263)	1.00 Reference	0.067	33.4% (1938)	1.00 Reference	<0.0001

(continued)

Appendix Table 1 Continued

Medical condition	Women			Men		
	% (*n*) blood donors	OR [95% CI]	*P*	% (*n*) blood donors	OR [95% CI]	*P*
Rheumatism						
Yes	32.1% (155)	1.06 [0.87, 1.29]		33.8% (103)	1.00 Reference	
No	30.8% (3285)	1.00 Reference	0.543	35.2% (2468)	1.07 [0.84, 1.36]	0.600
Arthritis						
Yes	28.1% (108)	1.00 Reference		33.5% (110)	1.00 Reference	
No	30.9% (3332)	1.15 [0.91, 1.44]	0.241	35.2% (2455)	1.08 [0.85, 1.36]	0.536
Psoriasis						
Yes	32.4% (129)	1.08 [0.87, 1.34]		34.9% (81)	1.00 Reference	
No	30.8% (3311)	1.00 Reference	0.489	35.1% (2451)	1.01 [0.77, 1.33]	0.943
Stomach ulcer						
Yes	27.7% (38)	1.00 Reference		31.1% (75)	1.00 Reference	
No	30.9% (3402)	1.16 [0.80, 1.70]	0.429	35.3% (2497)	1.21 [0.92, 1.60]	0.177
Pelvic inflammatory disease						
Yes	29.7% (81)	1.00 Reference		35.7% (322)	1.03 [0.89, 1.19]	
No	30.9% (3359)	1.06 [0.81, 1.38]	0.672	35.1% (2273)	1.00 Reference	0.685
Drug addiction						
Yes	16.7% (8)	1.00 Reference		24.4% (22)	1.00 Reference	
No	30.9% (3432)	2.24 [1.05, 4.78]	0.033	35.3% (2577)	1.69 [1.04, 2.72]	0.032
Alcoholism						
Yes	24.8% (25)	1.00 Reference		24.9% (50)	1.00 Reference	
No	30.9% (3415)	1.36 [0.86, 2.14]	0.183	35.4% (2523)	1.65 [1.20, 2.28]	0.002
Schizophrenia						
Yes	23.1% (3)	1.00 Reference		18.2% (2)	1.00 Reference	
No	30.9% (3437)	1.49 [0.41, 5.41]	0.544	35.1% (2584)	2.45 [0.53, 11.30]	0.239
Anorexia nervosa						
Yes	27.8% (64)	1.00 Reference		62.5% (5)	3.09 [0.74, 12.92]	
No	30.9% (3376)	1.16 [0.87, 1.55]	0.317	35.1% (2587)	1.00 Reference	0.104
Severe depression						
Yes	24.6% (236)	1.00 Reference		31.8% (140)	1.00 Reference	
No	31.4% (3204)	1.40 [1.21, 1.64]	<0.0001	35.3% (2438)	1.17 [0.95, 1.44]	0.134
Other psychiatric problem						
Yes	32.0% (79)	1.06 [0.81, 1.38]		38.1% (48)	1.14 [0.79, 1.64]	
No	30.8% (3361)	1.00 Reference	0.694	35.1% (2533)	1.00 Reference	0.479
Eyesight problem						
Yes	34.0% (1624)	1.29 [1.19, 1.40]		42.1% (1186)	1.62 [1.47, 1.79]	
No	28.6% (1753)	1.00 Reference	P < 0.0001	30.9% (1386)	1.00 Reference	<0.0001

Appendix Table 2 Proportion and unadjusted odds ratios (OR) among women and their partners who had donated blood according to medical history of one or both of their natural parents

	Women			Men		
Medical history of parent	**% (*n*) blood donors**	**OR [95% CI]**	***P***	**% (*n*) blood donors**	**OR [95% CI]**	***P***
Type I diabetes[1]						
Yes	30.9% (71)	1.00 [0.76, 1.33]	0.992	40.7% (72)	1.28 [0.94, 1.73]	0.116
No	30.8% (3369)	1.00 Reference		35.0% (2543)	1.00 Reference	
Type II diabetes[2]						
Yes	35.8% (187)	1.26 [1.05, 1.52]	0.013	32.1% (110)	1.00 Reference	0.229
No	30.6% (3253)	1.00 Reference		35.2% (2505)	1.15 [0.91, 1.45]	
Coronary heart disease						
Yes	36.1% (519)	1.31 [1.17, 1.47]	<0.0001	40.0% (467)	1.28 [1.13, 1.46]	<0.001
No	30.1% (2921)	1.00 Reference		34.2% (2148)	1.00 Reference	
Stroke						
Yes	31.8% (196)	1.05 [0.88, 1.25]	0.608	37.5% (230)	1.12 [0.95, 1.33]	0.190
No	30.8% (3244)	1.00 Reference		34.9% (2385)	1.00 Reference	
Hypertension						
Yes	33.1% (1234)	1.17 [1.08, 1.27]	<0.001	36.9% (791)	1.12 [1.00, 1.24]	0.041
No	29.7% (2206)	1.00 Reference		34.4% (1824)	1.00 Reference	
Rheumatism						
Yes	32.2% (683)	1.08 [0.98, 1.20]	0.136	37.1% (557)	1.11 [0.99, 1.25]	0.072
No	30.5% (2757)	1.00 Reference		34.6% (2058)	1.00 Reference	
Arthritis						
Yes	33.1% (1120)	1.16 [1.07, 1.27]	0.001	36.3% (733)	1.08 [0.97, 1.20]	0.172
No	29.9% (2320)	1.00 Reference		34.6% (1882)	1.00 Reference	
Cancer						
Yes	35.4% (576)	1.27 [1.14, 1.42]	<0.0001	37.8% (448)	1.15 [1.01, 1.31]	0.031
No	30.1% (2864)	1.00 Reference		34.6% (2167)	1.00 Reference	
Chronic bronchitis						
Yes	31.2% (296)	1.02 [0.88, 1.17]	0.825	38.4% (214)	1.16 [0.97, 1.39]	0.094
No	30.8% (3144)	1.00 Reference		34.8% (2401)	1.00 Reference	
Alcohol problem						
Yes	28.6% (257)	0.89 [0.77, 1.03]	0.127	29.3% (157)	0.75 [0.62, 0.91]	0.004
No	31.0% (3183)	1.00 Reference		35.5% (2458)	1.00 Reference	
Schizophrenia						
Yes	26.3% (20)	0.80 [0.48, 1.34]	0.391	40.0% (18)	1.23 [0.68, 2.25]	0.490
No	30.9% (3420)	1.00 Reference		35.1% (2597)	1.00 Reference	
Depression						
Yes	30.1% (884)	0.95 [0.87, 1.05]	0.317	36.6% (657)	1.09 [0.98, 1.22]	0.131
No	31.1% (2556)	1.00 Reference		34.6% (1958)	1.00 Reference	

[1]Insulin dependent diabetes
[2]Non-insulin dependent diabetes