© International Epidemiological Association 2003 Printed in Great Britain

International Journal of Epidemiology 2003;**32**:51–57
DOI: 10.1093/ije/dyg002

The detection of gene–environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement?

MY Wong,[1] NE Day,[2] JA Luan,[2] KP Chan[1] and NJ Wareham[2]

Accepted 18 February 2002

Background The search for biologically relevant gene–environment interactions has been facilitated by technological advances in genotyping. The design of studies to detect interactions on continuous traits such as blood pressure and insulin sensitivity is attracting increasing attention. We have previously described power calculations for such studies, and this paper describes the extension of those calculations to take account of measurement error.

Methods The model considered in this paper is a simple linear regression relating a continuous outcome to a continuously distributed exposure variable in which the ratio of slopes for each genotype is considered as the interaction parameter. The classical measurement error model is used to describe the uncertainty in measurement in the outcome and the exposure. The sample size to detect differing magnitudes of interaction with varying frequencies of the minor allele are calculated for a given main effect observed with error both in the exposure and the outcome. The sample size to detect a given interaction for a given minor allele frequency is calculated for differing degrees of measurement error in the assessment of the exposure and the outcome.

Results The required sample size is dependent upon the magnitude of the interaction, the allele frequency and the strength of the association in those with the common allele. As an example, we take the situation in which the effect size in those with the common allele was a quarter of a standard deviation change in the outcome for a standard deviation change in the exposure. If a minor allele with a frequency of 20% leads to a doubling of that effect size, then the sample size is highly dependent upon the precision with which the exposure and outcome are measured. ρ_{Tx} and ρ_{Ty} are the correlation between the measured exposure and outcome, respectively and the true value. If poor measures of the exposure and outcome are used, (e.g. $\rho_{Tx} = 0.3$, $\rho_{Ty} = 0.4$), then a study size of 150 989 people would be required to detect the interaction with 95% power at a significance level of 10^{-4}. Such an interaction could be detected in study samples of under 10 000 people if more precise measurements of exposure and outcome were made (e.g. $\rho_{Tx} = 0.7$, $\rho_{Ty} = 0.7$), and possibly in samples of under 5000 if the precision of estimation were enhanced by taking repeated measurements.

Conclusions The formulae for calculating the sample size required to study the interaction between a continuous exposure and a genetic factor on a continuous outcome variable in the face of measurement error will be of considerable utility in designing studies with appropriate power. These calculations suggest that smaller studies with repeated and more precise measurement of the exposure and

[1] Department of Mathematics, The Hong Kong University of Science & Technology, Hong Kong.

[2] University of Cambridge Institute of Public Health, Cambridge, UK.

Correspondence: Dr Nicholas J Wareham, University of Cambridge Institute of Public Health, Robinson Way, Cambridge CB2 2SR, UK. E-mail: njw1004@medschl.cam.ac.uk

Downloaded from https://academic.oup.com/ije/article/32/1/51/642808 by guest on 20 August 2022

outcome will be as powerful as studies even 20 times bigger, which necessarily employ less precise measures because of their size. Even though the cost of genotyping is falling, the magnitude of the effect of measurement error on the power to detect interaction on continuous traits suggests that investment in studies with better measurement may be a more appropriate strategy than attempting to deal with error by increasing sample sizes.

Keywords Environmental exposure, gene-environment interaction, genotype, quantitative trait, sample size

The calculation of the number of participants required for traditional forms of epidemiological studies is made relatively straightforward by the publication of formulae and tables allowing estimation of samples size for any given power and significance.[1] An increasing area of interest in epidemiology is the design of studies for the detection of gene-environment interactions. Established methods are already available for the computation of sample size for studies where the outcome is a category (e.g. hypertension or diabetes) and the environmental exposure is considered as a binary or ordered categorical state or as a continuum.[2–5] We have recently produced sample size formulae for situations where both the exposure and the outcome are continuously distributed.[6] The key determinants of power in this context are the allele frequency, the size of the main effect and the magnitude of the interaction effect. However, in planning studies to examine gene-environment interaction on continuous traits, researchers are also presented with choices about how the outcome and exposures are assessed. As with many such studies, the trade-off is one between precision and feasibility. If the exposure of interest is dietary, then the gold standard method may be a 7-day weighed diary, but if large numbers of participants are required a less precise instrument such as a food frequency questionnaire may be employed. The measurement error introduced by using a less precise exposure measurement gives rise to an attenuation of the true effect.[7] A similar phenomenon also occurs when the outcome measurement used in a particular study is a proxy for the true outcome of interest. In the case of examination of usual or habitual blood pressure for example, using a single measure as an estimate of the usual level leads to an attenuation of the true association.[8,9] Although there is an established literature on the impact of such error,[10] and also techniques for adjusting observed associations to take its impact into account,[11,12] the effect of the measurement error on the power to detect gene-environment interactions has not previously been considered. In the statistical literature, the general issues about the effects of measurement error on the power to detect interactions between two continuously distributed measures have been considered[13] but sample size formulae have not been presented. In this paper we describe power calculations that include information about the measurement error in the continuously distributed exposures and outcome. In addition we describe the impact of misclassification in the genotyping.

Methods

For the purpose of these calculations, we designate two different alleles at a certain locus as *A* and *a*, where *a* is the minor allele, giving three possible genotypes, *aa*, *aA* and *AA*. We have restricted our attention in this paper to the dominant genetic models only, allowing the three genotypes to be reduced to two genetic groups, i.e. carriers of the minor allele versus homozygotes for the common allele but extension of our approach to the recessive model is simple.

The relationship between the outcome and the genetic factor with a non-genetic exposure can be expressed as two simple linear regressions shown below.

$$y = \alpha_1 + \beta_1 E + \varepsilon \text{ for an individual in the first group;}$$

$$y = \alpha_2 + \beta_2 E + \varepsilon \text{ for an individual in the second group}$$

where y is a continuous outcome variable; E represents a continuously distributed environmental exposure; ε represents a stochastic error term and is assumed to be normally distributed with mean zero and variance σ_y^2. We assume that the variances of exposure E in each group are equal. The regression coefficient β_i reflects the magnitude of the contribution of the environmental exposure to outcome, y, for the ith group. If the outcome is not significantly associated with the non-genetic exposure or if that relationship cannot be expressed in terms of a linear function, then subsequent examination of the data for possible interaction would not be appropriate. In addition, if the model chosen to describe the linear relationship is in fact log-linear, then the interaction term is specific to the manner in which the data are transformed and cannot be generalized either to other transformations or the situation where data are untransformed.

To study the effect of the environmental exposure, E, on the association of the dependent variable with the genetic factor, we test the hypothesis $H_0 : \beta_1 = \beta_2 = \beta$. If β_1 and β_2 are equal, we have two parallel lines and thus there is no interaction. Because of measurement error, instead of the true exposure E_t, we observe its corresponding surrogate E_o. We assume that the measurement error is non-differential with regard to the outcome variable y, i.e. E_o contributes no information about y beyond what is available in E_t. E_o is related to E_t by an additive error model as $E_o = E_t + \varepsilon_e$ with $\mathrm{E}(\varepsilon_e) = 0$ and $\mathrm{Var}(\varepsilon_e) = \sigma_e^2$. The true exposure E_t is assumed to have mean μ and variance τ^2. The correlation coefficient of the true exposure E_t and its corresponding surrogate E_o is defined as ρ_{Tx}. A similar phenomenon exists of the outcome variable that is assessed by the surrogate Y_o which is correlated to the true outcome Y_t by the coefficient ρ_{Ty}.

We consider a general situation for a polymorphism which is in Hardy-Weinberg equilibrium[14] with a minor allele frequency

Downloaded from https://academic.oup.com/ije/article/32/1/51/642808 by guest on 20 August 2022

of p, giving genotype frequencies for aa, aA and AA of p^2, $2p(1 - p)$ and $(1 - p)^2$, respectively. Accordingly, the true proportions of individuals in the two genetic groups, p_{T1} and p_{T2}, are $p(2 - p)$ and $(1 - p)^2$ for a dominant model. We assume that misclassification of each allele is independent of the other. If the probabilities of misclassification of A and a are P_A and P_a, respectively, then the observed genotype frequency of the rare gene is equal to $p' = (1 - p)P_A + p(1 - P_a)$. The observed frequencies of aa, aA and AA are thus p'^2, $2\,p'(1 - p')$ and $(1 - p')^2$, respectively. When the exposure is subject to classical measurement error model, the conditional mean and variance of y on the observed exposure in the ith group are $\alpha_i + \beta_i(1 - \rho^2{}_{Tx})\ \mu + \beta_i\rho^2{}_{Tx}E_o$ and $\sigma^2{}_y + \beta^2{}_i\ \tau^2(1 - \rho^2{}_{Tx})$, respectively.[15]

In a situation where the true exposure cannot be observed and the genotype cannot be assessed correctly, the likelihood ratio test statistic W_β (Appendix), under the alternative hypothesis, is approximately distributed as a non-central χ^2 with one degree of freedom. The non-centrality parameter ϕ_n is given in the Appendix. Using the distribution and the non-centrality parameter, we are then able to calculate power for detecting an interaction effect or alternatively the sample size necessary to detect a given interaction with fixed power and significance level using any statistical software, e.g. SAS. Power at the significance level α for a fixed sample size n is equal to the probability of a chi-square random variable with one degree of freedom and a non-centrality parameter ϕ_n, greater than $\chi^2_{1-\alpha}(1)$, where $\chi^2_{1-\alpha}(1)$ is the $100(1-\alpha)$th percentile of the chi-square distribution with one degree of freedom. In SAS, it is calculated as

$$\text{PROBCHI}(\text{CINV}(1-\alpha,\ 1),\ 1,\ \phi_n).$$

The sample size needed to achieve a power of at least (1 – β) is thus the smallest positive integer n satisfying the inequality of

$$\Pr(\chi^2 > \chi^2_{1-\alpha}\,(1)) \geqslant 1 - \beta,$$

where χ^2 has a chi-square distribution with one degree of freedom and non-centrality parameter ϕ_n.

In the figures we present power calculations over a range of values for β_1 and β_2. In order to give the β coefficients a clear interpretation, we standardize the environmental exposure by making $\tau^2 = 1$. In most situations E would account for 20% or less of the total variation in y and therefore the residual variance of y after adjusting for E, σ^2_y, would be within 10% of the population standard deviation. We take $\sigma^2_y = 1/\rho^2_{Ty} - 1$. Thus the β coefficients are interpretable as the approximate proportion of a standard deviation change in y for a standard deviation change in E.

Results

As with all power calculations, the required sample sizes are dependent upon the level of significance and power assumed. For the purpose of illustration we have selected a power of 95% and a significance of 10^{-4} but calculation of sample size for different values of power and significance is straightforward, given the formula. The main determinants of the sample size required to detect interaction between a gene and a continuous trait for a continuous outcome are the strength of the true association in those with the common allele, the magnitude of the interaction, the measurement error of the exposure and outcome, the frequency of the minor allele and the degree of genetic misclassification. Rather than attempt to show the impact on sample size of varying all these parameters at once, we have elected to show the effects of varying combinations of them.

Table 1 shows the effect on sample size of varying allele frequencies (p) and allele misclassification rates (P_A and P_a) to detect an interaction effect of 2 with a moderate effect ($\beta_2 = 0.25$). The effect of the allele misclassification on power is greatest when the minor allele frequency is low. Thus for common alleles reduction of measurement error in the classification of genotypes by repeated measurement would not greatly increase power and would, of course, result in considerable additional expense. However, for potentially important but less common alleles, it may be important to reduce genetic misclassification. Repeated measurement would only diminish that component of error that was random. If the misclassification were nonrandom, this would not be the case. It may be that the degree of misclassification varies according to which polymorphism is being examined, in which case computation of the extent of measurement error in a pilot study of a particular polymorphism would be as important as estimating its frequency.

For the purposes of the remaining calculations we have assumed that the genetic misclassification is fixed at 2.5%, a figure in the range demonstrated in empirical studies.[16–19] In Table 2 we have fixed the measurement error in the exposure

Table 1 Sample size for detecting a gene-environment interaction ($\beta_1 = 0.5$, $\beta_2 = 0.25$) with 95% power at a significance level of 10^{-4} for different minor allele frequencies (p) and varying degrees of genetic misclassification (P_A and P_a)

Allele misclassification	Minor allele frequency			
	0.05	0.10	0.15	0.20
0.0	16 909	9766	7568	6644
0.01	20 690	10 949	8232	7122
0.025	26 890	12 887	9321	7904
0.05	38 857	16 626	11 418	9411
0.10	70 983	26 651	17 039	13 447

The parameters fixed in this calculation are the exposure measurement error $\rho_{T_x} = 0.6$, the outcome measurement error $\rho_{T_y} = 0.7$, the effect size in the common allele group $\beta_2 = 0.25$ and the interaction

$$\frac{\beta_1}{\beta_2} = 2.$$

Table 2 Sample size required to detect with 95% power and a significance level of 10^{-4} different degrees of interaction between genotype and a continuous exposure on a continuously distributed outcome for different minor allele frequencies

	Minor allele frequency			
β_1/β_2	0.05	0.10	0.15	0.20
1.5	106 886	50 926	36 631	30 906
2.0	26 890	12 887	9321	7904
3.0	6843	3333	2447	2103
4.0	3116	1551	1160	1014

The parameters fixed in this calculation are the exposure measurement error $\rho_{Tx} = 0.6$, the outcome measurement error $\rho_{Ty} = 0.7$, the effect size in the common allele group $\beta_2 = 0.25$, and the gene misclassification rate $P_A = P_a = 0.025$.

Downloaded from https://academic.oup.com/ije/article/32/1/51/642808 by guest on 20 August 2022

and outcome at $\rho_{Tx} = 0.6$ and $\rho_{Ty} = 0.7$. The magnitude of the true effect in the group homozygous for the common allele is fixed at 0.25, which can be interpreted as a quarter of a standard deviation difference in the outcome for a standard deviation difference in the exposure. The table demonstrates how relatively small interaction effects on uncommon alleles will be difficult to detect unless study samples exceed 100 000 individuals. Conversely, interactions for common alleles that are very strong may be detected in study samples with as few as 1014 individuals. The important fixed variables in Table 2 are the error in the assessment of the exposure and outcome. The values of 0.6 and 0.7 for the correlation between the true and observed exposure and outcome, respectively, would be typical of studies where relatively precise methods are employed. In reality, such methods are rarely employed in large studies where less accurate methods are often employed in the interests of feasibility.

Table 3 shows how study sample size is heavily dependent upon the measurement error in the exposure and outcome. For studies with poor assessment of exposure and outcome ($\rho_{Tx} = 0.3$ and $\rho_{Ty} = 0.4$), sample sizes in excess of 100 000 individuals would be required to detect an interaction that was detectable in under 20 000 people with studies employing even reasonably accurate measurement ($\rho_{Tx} = \rho_{Ty} = 0.6$). Improving the measurement can be achieved by taking repeated measurements provided the error in repeated measures is uncorrelated.[7] For a measurement with a validity coefficient of 0.6, taking two independent repeated measures increases the overall validity coefficient (ρ_{Ta}) to just under 0.8. Referring to Table 3 this would reduce the necessary sample size from 13 086 to 2410.

Discussion

In this paper, we present the formulae necessary to calculate the statistical power and the sample size for the study of interaction between a continuous environment exposure and a genotype on a continuous outcome variable when there is measurement error in the assessment of both exposure and outcome and misclassification error in assessing the genotype. The need for such sample size calculations is likely to increase as we attempt to design studies aimed at understanding the genetic basis of common diseases. The impact of the misclassification in the assessment of the genotype is relatively minor except when the frequency of the minor allele is low. Given that the misclassification may differ between specific polymorphisms, some assessment of typing error may need to be built into pilot phases of association studies, which will be necessary in any event to calculate allele frequencies. When the allele frequency is low but the error is high, it may be worth undertaking repeat genotyping to reduce that error, provided that error is random. An assumption of the analyses presented here is that the error is non-differential and different results would be found if genotyping or exposure measurement were subject to non-random error.

A greater impact on power comes not from genotyping errors, but from the precision with which the exposure and outcome are estimated. The practical consideration when designing studies aimed at detecting gene-environment interactions will be the trade-off between sample size and measurement precision. Our calculations suggest that this trade-off should be weighted towards better measurement. This general point may be illustrated with an example, the study of the relationship between physical activity and insulin sensitivity. This association has been demonstrated in previous epidemiological studies[20,21] and is biologically plausible as intervention studies demonstrate improvements of insulin sensitivity with increasing activity.[22] Ecologic studies would suggest that certain sub-groups of the population e.g. people from specific at-risk ethnic groups or those with a family history of diabetes, are more susceptible to the adverse effects of sedentary living than others.[23] There is also reasonable evidence that insulin sensitivity has a genetic component,[24] and thus the search for gene-physical activity interactions in this context would be logical. In designing a study to examine this association and possible gene-physical activity interactions, one would be left with difficult choices for the assessment of both the exposure and the outcome. The accepted optimal methods for assessing insulin sensitivity are either the frequently sample minimal model intravenous glucose tolerance test or the euglycaemic clamp technique.[25]

Table 3 Sample size required to detect with 95% power and a significance level of 10^{-4} a given interaction for different degrees of precision in the continuously distributed exposure and outcome

		ρ_{Tx}						
β_2	ρ_{Ty}	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.10	0.4	926 208	520 848	333 225	231 306	169 852	129 966	102 620
	0.5	530 688	298 368	190 837	132 426	97 205	74 346	58 673
	0.6	315 838	177 515	113 491	78 713	57 743	44 132	34 801
	0.7	186 290	104 644	66 854	46 326	33 948	25 915	20 407
	0.8	102 208	57 348	36 585	25 306	18 505	14 091	11 064
0.25	0.4	150 989	84 787	54 146	37 501	27 464	20 950	16 484
	0.5	87 705	49 191	31 364	21 680	15 841	12 051	9453
	0.6	53 329	29 854	18 988	13 086	9527	7217	5633
	0.7	32 602	18 195	11 526	7904	5720	4302	3330
	0.8	19 149	10 627	6683	4541	3249	2410	1836

The parameters fixed in this calculation are the minor allele frequency $p = 0.2$, the gene misclassification $P_A = P_a = 0.025$, the interaction $\frac{\beta_1}{\beta_2} = 2$.

Downloaded from https://academic.oup.com/ije/article/32/1/51/642808 by guest on 20 August 2022

Both of these tests are difficult to do in populations greater than a few hundred individuals and therefore epidemiological studies have relied on proxy measures. One such measure is the fasting insulin concentration which has a correlation with the gold standard method of whole-body glucose uptake of about 0.66 in normoglycaemic individuals.[26] A study with two repeats of fasting insulin would provide an overall assessment that had a correlation with the true outcome of 0.84, assuming that the error in each repeat was uncorrelated with each other. However, many large epidemiological studies cannot study fasted individuals and may employ indicative measures of insulin sensitivity that are more distantly related. One such example could be the waist-hip ratio, an indicator of the degree of central obesity which has previously been shown to be associated with insulin resistance.[27,28] However, the correlation between the waist-hip ratio and fasting insulin is only of the order of 0.3.[29] If such a poor measure of outcome were employed, large numbers would be required to overcome the measurement error. As Table 3 indicates, there would be a 10-fold difference in sample size between a study employing a good measure of outcome i.e. repeated fasting insulin, compared to one relying on a poor measure such as waist-hip ratio.

The paradox is that the larger study employing the poorer measurement would, for practical reasons, also compromise on the exposure measurement. If that exposure of interest were physical activity, then large studies would probably only consider a questionnaire. Even a comprehensive questionnaire covering occupational and recreational activity is unlikely to have a correlation with the true exposure of interest of habitual energy expenditure in a general population of above 0.3. For example, the correlation of the ARIC/Baeke questionnaire with objective movement sensor derived estimates of energy expenditure was only 0.24 in men and 0.19 in women.[30] Our study of the EPIC-Norfolk physical activity questionnaire (EPAQ2) demonstrated an overall correlation of 0.44 with repeated measures of energy expenditure over one year. The correlation after adjustment of age and sex was 0.28.[31] Other questionnaires such as the Cardia questionnaire used in the Insulin Resistance Atherosclerosis Study (IRAS) to demonstrate an association between physical activity and insulin sensitivity have not been associated with energy expenditure measured by objective methods.[32,33] The use of a relatively poor measure of physical activity ($\rho_{Tx} = 0.3$) together with a proxy for insulin sensitivity such as waist-hip ratio ($\rho_{Ty} = 0.4$) would mean that over 150 000 individuals would need to be genotyped to detect a doubling of an effect in individuals with a minor allele that was present in 20% of the population. Although genotyping on this scale will undoubtedly become increasing feasible and less costly, there may be considerable cost savings in investing in better measurement.

As an alternative to physical activity questionnaires more direct measurements have been proposed. The optimal method for measuring energy expenditure in free-living individuals, the doubly-labelled water technique is very expensive and isotopes are not always readily available.[34] Less expensive objective methods such as heart rate monitoring with individual calibration have been correlated with the gold standard methods over the short term (r = 0.93) and are applicable in medium-sized epidemiological studies.[35–37] Studies with repeated assessment of energy expenditure by heart rate monitoring suggest that a single measure has a correlation with the latent variable of habitual energy expenditure of 0.73.[38] Thus a study that had even two repeat measurements would increase the overall ρ_{Tx} to 0.88. Referring to Table 3 one can see that if such a method were employed in a study with repeated measures of fasting insulin as the outcome, then a sample size of 2000 would be sufficient to detect the interaction that required a study of more that 150 000 individuals with poorer measurement.

Although incorporating increased precision of measurement into a study requires additional resources, these would certainly be dwarfed by the savings on study infra-structure and genotyping costs when compared to bigger studies with less accurate methods. The magnitude of the impact of measurement precision on power to detect gene-environment interaction on continuous traits would suggest that smaller studies with better measurement may be preferable to very large studies with less precise measurement.

Acknowledgements

The work of MYW on this project was funded by the Royal Society. NJW is an MRC Clinician Scientist Fellow.

Downloaded from https://academic.oup.com/ije/article/32/1/51/642808 by guest on 20 August 2022

KEY MESSAGES

- The sample size needed to detect the interaction of a genetic factor with a continuously distributed exposure on a continuous outcome is dependent upon the magnitude of the interaction, the allele frequency and the strength of the association between exposure and outcome.
- Sample size is highly dependent upon the measurement error in the assessment of the exposure and outcome variables.
- Studies employing imprecise exposure and outcome assessment may need to be 20 times larger than studies that utilize repeated and more precise measurement.
- Investment in better measurement may be a more cost-effective strategy for the detection of this form of gene-environment interaction than simply increasing sample size.

References

[1] Machin D, Campbell M, Foyers P, Pinol A. *Sample Size Tables for Clinical Studies*. Oxford: Blackwell Science, 1997.

[2] Hwang SJ, Beaty TH, Liang KY, Coresh J, Khoury MJ. Minimum sample-size estimation to detect gene environment interaction in case-control designs. *Am J Epidemiol* 1994;**140:**1029–37.

[3] Foppa I, Spiegelman D. Power and sample size calculations for case-control studies of gene-environment interactions with a polytomous exposure variable. *Am J Epidemiol* 1997;**146:**596–604.

[4] Garcia-Closas M, Lubin JH. Power and sample size calculations in case-control studies of gene-environment interactions: comments on different approaches. *Am J Epidemiol* 1999;**149:**689–92.

[5] Garcia-Closas M, Rothman N, Lubin J. Misclassification in case-control studies of gene-environment interaction: assessment of bias and sample size. *Cancer Epidemiol Biomark Prev* 1999;**8:**1043–50.

[6] Luan JA, Wong MY, Day NE, Wareham NJ. Sample size determination for studies of gene-environment interaction. *Int J Epidemiol* 2001;**30:** 1035–40.

[7] Armstrong BK, White E, Saracci R. *Principles of Exposure Measurement in Epidemiology*. Oxford: Oxford University Press, 1994.

[8] International Co-operative Research Group. Intersalt: an international study of electrolyte excretion and blood pressure. Results for 24 hour urinary sodium and potassium excretion. *BMJ* 1988;**297:**319–28.

[9] MacMahon S, Peto R, Cutler J *et al.* Blood pressure, stroke, and coronary heart disease. *Lancet* 1990;**335:**765–74.

[10] Devine OJ, Smith JM. Estimating sample size for epidemiologic studies: the impact of ignoring exposure measurement uncertainty. *Stat Med* 1998;**17:**1375–89.

[11] Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *Am J Epidemiol* 1990;**132:**734–45.

[12] Wong MY, Day NE, Wareham NJ. Measurement error in epidemiology: The design of validation studies II: bivariate situation. *Stat Med* 1999;**18:**2831–45.

[13] Aiken LS, West SG. *Multiple Regression: Testing and Interpreting Interactions. First Edn*. Newbury Park, CA: SAGE Publications, 1991.

[14] Khoury MJ, Beaty TH, Cohen BH. *Fundamentals of Genetic Epidemiology*. Oxford: Oxford University Press, 1993.

[15] Mood AM, Graybill FA, Boes DC. *Introduction to the Theory of Statistics. Third Edn.* London: McGraw-Hill Book Company, 1974.

[16] Pastinen T, Raitio M, Lindroos K, Tainola P, Petonen L, Syvanen AC. A system for specific, high-throughput genotyping by allele-specific primer extension on microarrays. *Genome Res* 2000;**10:**1031–42.

[17] Prince JA, Feuk L, Howells WM *et al.* Robust and accurate single nucleotide polymorphism genotyping by dynamic allele-specific hybridization (DASH): design criteria and assay validation. *Genome Res* 2001;**11:**152–62.

[18] Wang DG, Fan JB, Siao CJ *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998;**280:**1077–82.

[19] Hacia JG, Fan JB, Ryder O *et al.* Determination of ancestral alleles for human single nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat Genet* 1999;**22:**164–67.

[20] Mayer-Davis EJ, D'Agostino R Jr, Karter AJ *et al.* Intensity and amount of physical activity in relation to insulin sensitivity: The Insulin Resistance Atherosclerosis Study. *JAMA* 1998;**279:**669–74.

[21] Regensteiner JG, Mayer EJ, Shetterly SM *et al.* Relationship between habitual physical activity and insulin levels among nondiabetic men and women. *Diabetes Care* 1991;**14:**1066–74.

[22] Holloszy JO, Schultz J, Kusnierkiewicz J, Hagberg JM, Ehsani AA. Effects of exercise on glucose tolerance and insulin resistance. Brief review and some preliminary results. *Acta Med Scand Suppl* 1986;**711:** 55–65.

[23] Hamman RF. Genetic and environmental determinants of non-insulin-dependent diabetes mellitus (NIDDM). *Diabetes Metab Rev* 1992;**8:**287–38.

[24] Mayer EJ, Newman B, Austin MA *et al.* Genetic and environmental influences on insulin levels and the insulin resistance syndrome: an analysis of women twins. *Am J Epidemiol* 1996;**143:**323–32.

[25] Anderson RL, Hamman RF, Savage PJ *et al.* Exploration of simple insulin sensitivity measures derived from frequently sampled intravenous glucose tolerance (FSIGIT) Tests: the Insulin Resistance Atherosclerosis Study. *Am J Epidemiol* 1995;**142:**724–32.

[26] Laakso M. How good a marker is insulin level for insulin resistance? *Am J Epidemiol* 1993;**137:**959–65.

[27] Groop LC. Insulin resistance: the fundamental trigger of type 2 diabetes. *Diabetes Obes Metab* 1999;**1(S1):**S1–S7.

[28] Abate N, Garg A, Peshock RM *et al.* Relationship of generalized and regional adiposity to insulin sensitivity in men with NIDDM. *Diabetes* 1996;**45:**1684–93.

[29] Byrne CD, Wareham NJ, Day NE, McLeish R, Williams DRR, Hales CN. Decreased non-esterified fatty acids suppression and features of the insulin resistance syndrome occur in a sub-group of individuals with normal glucose tolerance. *Diabetologia* 1995;**38:**1358–66.

[30] Richardson MT, Ainsworth BE, Wu H-C, Jacobs DR, Leon AS. Ability of the Atherosclerosis Risk in Communities (ARIC)/Baeke questionnaire to assess leisure-time physical activity. *Int J Epidemiol* 1995;**24:** 685–93.

[31] Wareham NJ, Jakes RW, Rennie KL, Mitchell J, Hennings S, Day NE. Validity and repeatability of the EPIC-Norfolk Physical Activity Questionnaire. *Int J Epidemiol* 2002;**31:**168–74.

[32] Jacobs DR, Ainsworth BE, Hartman TJ, Leon AS. A simultaneous evaluation of 10 commonly used physical activity questionnaires. *Med Sci Sports Exerc* 1993;**25:**81–91.

[33] Pereira MA, Fitzgerald SJ, Gregg EW *et al*. A collection of physical activity questionnaires for health-related research. *Med Sci Sports Exerc* 1997;**29(Suppl.6):**S1–S205.

[34] Prentice AM (ed.). *The Doubly-Labelled Water Method for Measuring Energy Expenditure: Technical Recommendations for Use in Humans*. A consensus report by the International Dietary Energy Consultancy Working Group. Vienna: International Atomic Energy Agency, 1990.

[35] Ceesay SM, Prentice AM, Day KC, Murgatroyd PR, Goldberg GR, Scott W. The use of heart rate monitoring in the estimation of energy expenditure: a validation study using indirect whole-body calorimetry. *Br J Nutr* 1989;**61:**175–86.

[36] Spurr GB, Prentice AM, Murgatroyd PR, Goldberg GR, Reina JC, Christman NT. Energy expenditure from minute-by-minute heart-rate recording: comparison with indirect calorimetry. *Am J Clin Nutr* 1988;**48:**552–59.

[37] Livingstone MBE, Prentice AM, Coward WA *et al.* Simultaneous measurement of free-living energy expenditure by the doubly labeled water method and heart-rate monitoring. *Am J Clin Nutr* 1990;**52:** 59–65.

[38] Wareham NJ, Wong M-Y, Day NE. Glucose intolerance and physical inactivity: the relative importance of low habitual energy expenditure and cardiorespiratory fitness. *Am J Epidemiol* 2000;**152:**132–39.

[39] Myers RH. *Classical and Modern Regression with Application*. Boston: PWS-Kent; 1990.

[40] Campbell MJ. *Statistics at Square Two: Understanding Modern Statistical Applications in Medicine*. London: BMJ Publishing, 2000.

Downloaded from https://academic.oup.com/ije/article/32/1/51/642808 by guest on 20 August 2022

Appendix

The likelihood ratio test, for testing $H_0 : \beta_1 = \beta_2 = \beta$, is equal to $(\hat{\sigma}^2/\hat{\sigma}^2_\beta)^{n/2}$, where n is the sample size, $\hat{\sigma}^2$ equals to $Y'(I - X(X'X)^{-1}X')Y/n$ and $\hat{\sigma}^2_\beta$ equals to $Y'(I - X_\beta(X'_\beta X_\beta)^{-1}X'_\beta)Y/n$, where $Y = (y_1,y_2,\ldots,y_n)'$, X is the design matrix accommodating the linear regression model in this paper, i.e.,

$$X = \begin{bmatrix} 1 & 0 & x_1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_k & 0 \\ 0 & 1 & 0 & x_{k+1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & x_n \end{bmatrix},$$

and X_β is the design matrix when $\beta_1 = \beta_2$, i.e.,

$$X_\beta = \begin{bmatrix} 1 & 0 & x_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_k \\ 0 & 1 & x_{k+1} \\ \vdots & \vdots & \vdots \\ 0 & 1 & x_n \end{bmatrix},$$

where x_i is the environmental variable of the ith individual and k is the number of individuals in the first genetic group.[39,40] When the sample size n is large and H_0 is true, $W_\beta = n\log(\hat{\sigma}^2_\beta/\hat{\sigma}^2)$ is approximately distributed as a chi-squared distribution with one degree of freedom. The statistic W_β has a limiting non-central chi-squared distribution with one degree of freedom.[15] The non-centrality parameter ϕ_n is

$$\frac{n\tau^2\rho^2_{T_x}\left\{\sum_{i=1}^{2}\sum_{j=1}^{2}\beta_i\beta_j\left(\sum_{k=1}^{2}\frac{\delta_{ki}\delta_{kj}}{p_{o_k}} - p_{Ti}p_{Tj}\right)\right\}}{\sigma^2_y + \tau^2\left(1-\rho^2_{T_x}\right)\sum_{i=1}^{2}\beta_i^2 p_{Ti}}$$

where δ_{st} is the joint probability that an individual has been assessed to the sth group but it, in fact, belongs to the tth group, p_{Ti} and p_{oi} is the true and observed proportions of individuals in the ith genetic group with $\Sigma p_{Ti} = \Sigma p_{oi} = 1$ and $p_{oi} = \delta_{i1} + \delta_{i2}$.

It can be shown that for a dominant model, if we assign the carriers of the rare allele into the first group, $p_{T_1} = p\ (2 - p)$, $p_{T2} = (1 - p)^2$, $\delta_{11} = p^2\ (1 - P_a^2) + 2p(1 - p)(1 - P_a + P_aP_A)$, $\delta_{12} = (1 - p)^2\ P_A(2 - P_A)$, $\delta_{21} = p^2P_a^2 + 2p(1 - p)P_a(1 - P_A)$ and $\delta_{22} = (1 - p)^2(1 - P_A)^2$. For a recessive model, if we assign the homozygotes for the rare allele into the first group, $p_{T1} = p^2$, $p_{T2} = 1 - p^2$, $\delta_{11} = p^2(1 - P_a)^2$, $\delta_{12} = 2p(1 - p)(1 - P_a)P_A + (1 - p)^2 P_A^2$, $\delta_{21} = p^2P_a(2 - P_a)$ and $\delta_{22} = 2p(1 - p)(1 - P_A + P_aP_A) + (1 - p)^2(1 - P_A^2)$.

The non-centrality parameter, if the true exposure can be observed and the genotype can be assessed correctly, is obtained by setting P_A and P_a to be zero and ρ_{Tx} to be one. It is, thus, equal to

$$\frac{n\tau^2\left\{\sum_{i=1}^{2}\beta_i^2 p_{Ti} - \left(\sum_{i=1}^{2}\beta_i\, p_{Ti}\right)^2\right\}}{\sigma^2_y}.$$

The same power can be achieved in different situations if the non-centrality parameter is identical. Hence, the ratio of the sample size required to attain the desired power for a likelihood ratio test based on the surrogate exposure to that based on the true exposure is equal to

$$\frac{\left\{\sigma^2_y + \tau^2\left(1-\rho^2_{T_x}\right)\sum_{i=1}^{2}\beta_i^2 p_{Ti}\right\}\left\{\sum_{i=1}^{2}\beta_i^2 p_{Ti} - \left(\sum_{i=1}^{2}\beta_i\, p_{Ti}\right)^2\right\}}{\sigma^2_y\rho^2_{T_x}\left\{\sum_{i=1}^{2}\sum_{j=1}^{2}\beta_i\beta_j\left(\sum_{k=1}^{2}\frac{\delta_{ki}\delta_{kj}}{p_{o_k}} - p_{Ti}p_{Tj}\right)\right\}}.$$

When the probabilities of misclassification are equal to zero, the ratio becomes

$$\frac{1}{\rho^2_{T_x}} + \frac{\tau^2}{\sigma^2_y}\left(\frac{1}{\rho^2_{T_x}} - 1\right)\sum_{i=1}^{2}\beta_i^2 p_{Ti}.$$

Under the situation that $\beta_1 = \beta_2$, the likelihood ratio test statistic, for testing $H_0 : \alpha_1 = \alpha_2 = \alpha$, is equal to $W_a = n\log(\hat{\sigma}^2_\alpha/\hat{\sigma}^2_\beta)$, where $\hat{\sigma}^2_\alpha$ is equal to $Y'(I - X_\alpha(X'_\alpha X_\alpha)^{-1}X'_\alpha)Y/n$ for X_α being the design matrix when $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$, i.e., a $n \times 2$ matrix with all elements in the first column equal to one and $x_1, x_2, \ldots, x_n$ in the second column.

When $\beta_1 = \beta_2 = \beta$, the test statistic W_a for testing the equality of intercepts follows a chi-squared distribution with one degree of freedom under the null hypothesis and a non-central chi-squared distribution with non-centrality parameter

$$\frac{n\left\{\sum_{i=1}^{2}\sum_{j=1}^{2}\alpha_i\alpha_j\left(\sum_{k=1}^{2}\frac{\delta_{ki}\delta_{kj}}{p_{o_k}} - p_{Ti}p_{Tj}\right)\right\}}{\sigma^2_y + \beta^2\tau^2\left(1-\rho^2_{T_x}\right)},$$

under the alternative hypothesis. Without errors on measuring exposure and on assessing the genotype, the non-centrality parameter becomes

$$\frac{n\left\{\sum_{i=1}^{2}\alpha_i^2 p_{Ti} - \left(\sum_{i=1}^{2}\alpha_i\, p_{Ti}\right)^2\right\}}{\sigma^2_y}.$$

Thus, to achieve the same power at a fixed significance level, the ratio of sample sizes based on the surrogate exposure and its true one is equal to

$$\frac{\left(\sigma^2_y + \beta^2\tau^2\left(1-\rho^2_{T_x}\right)\right)\left\{\sum_{i=1}^{2}\alpha_i^2 p_{Ti} - \left(\sum_{i=1}^{2}\alpha_i\, p_{Ti}\right)^2\right\}}{\sigma^2_y\left\{\sum_{i=1}^{2}\sum_{j=1}^{2}\alpha_i\alpha_j\left(\sum_{k=1}^{2}\frac{\delta_{ki}\delta_{kj}}{p_{o_i}} - p_{Ti}p_{Tj}\right)\right\}}$$

Even when there is no misclassification in assessing the genotype, the ratio becomes

$$1 + \frac{\tau^2}{\sigma^2_y}\beta^2\left(1-\rho^2_{T_x}\right).$$

The non-centrality parameter of the test statistic is thus smaller when the true exposure is unobservable. The loss of power can be substantial. It is noted that the loss of power is related to the strength of the exposure as measured by β^2.

Downloaded from https://academic.oup.com/ije/article/32/1/51/642808 by guest on 20 August 2022