Sample size determination for studies of gene-environment interaction

JA Luan, MY Wong, NE Day and NJ Wareham

Background The search for interaction effects is common in epidemiological studies, but the power of such studies is a major concern. This is a practical issue as many future studies will wish to investigate potential gene-gene and gene-environment interactions and therefore need to be planned on the basis of appropriate sample size calculations.

Methods and Results The underlying model considered in this paper is a simple linear regression relating a continuous outcome to a continuously distributed exposure variable. The slope of the regression line is taken to be dependent on genotype, and the ratio of the slopes for each genotype is considered as the interaction parameter. Sample size is affected by the allele frequency and whether the genetic model is dominant or recessive. It is also critically dependent upon the size of the association between exposure and outcome, and the strength of the interaction term. The link between these determinants is graphically displayed to allow sample size and power to be estimated. An example of the analysis of the association between physical activity and glucose intolerance demonstrates how information from previous studies can be used to determine the sample size required to examine gene-environment interactions.

Conclusions The formulae allowing the computation of the sample size required to study the interaction between a continuous environmental exposure and a genetic factor on a continuous outcome variable should have a practical utility in assisting the design of studies of appropriate power.

Keywords Genotype, environmental exposure, gene-environment interaction, sample size, quantitative trait

Accepted 7 February 2001

The study of interaction or effect modification is frequently undertaken in epidemiology, but the power of such studies to demonstrate these interactions, and therefore their sample size is a matter of concern. Previous papers have presented power and required sample size calculations for case-control studies of gene-environment interaction where the environmental factor is categorical and the genetic factor is binary. Using published formulae, Hwang et al. presented sample size calculations for a binary environmental exposure and a binary genetic factor. These were extended by Foppa and Spiegelman to consider an environmental exposure that was categorized into multiple levels. Both of these methods were subsequently compared with an approach designed for the general multivariate regression model for the odds ratio. However, in all of these studies the outcome is a binary event variable, such as the occurrence of a disease of interest.

The alternative situation where the outcome variable is continuously distributed has received less attention but is likely to become important as researchers investigate the genetic basis of quantitative traits such as blood pressure and obesity. A method for calculating power in this situation was recently described but was limited to a number of specific situations in which some main and interaction effects were fixed to zero. In the approach presented here, we consider the situation of an effect of a categorical genetic factor on the association between a continuous environmental exposure and a continuously distributed outcome. We illustrate the utility of this approach with an example of the investigation of the interaction between genes and physical activity in the determination of glucose tolerance.
Methods

Suppose that we consider a certain autosomal locus in which there are two different alleles, a and A, where a is the rare allele. There would be three possible genotypes, aa, aA, and AA. For the purposes of this analysis we have considered dominant and recessive models which allow the three genotypes to be reduced to two genetic groups, i.e. dominant (carriers of the rare allele versus homozygotes for the common allele) or recessive (homozygotes for the rare allele versus all others). The relationship between a continuous outcome variable y and the genetic factor with a continuous environmental exposure E can be expressed as two simple linear regressions

\[ y = \alpha_1 + \beta_1 E + \epsilon, \text{ if an individual is in the first group; } \]
\[ y = \alpha_2 + \beta_2 E + \epsilon, \text{ if an individual is in the second group. } \]

The regression parameters \( \alpha_i \) and \( \beta_i \) are weights reflecting the contribution of the genetic factor and the environmental exposure to the continuously distributed outcome y. If there is no gene-environment interaction, then the regression parameters \( \beta_1 \) and \( \beta_2 \) are equal. \( \epsilon \) is a stochastic error term and is assumed to be normally distributed with mean zero and variance \( \sigma^2 \). We assume the distributions of the residual of y in each group are the same, and the variances of exposure E in each group are \( \tau^2 \). In order to give the \( \beta \) parameters a clear interpretation, we have standardized both the outcome and the environmental exposure by making \( \sigma^2_y = \tau^2 = 1 \). \( \sigma^2_y \) is the residual variance of y after adjusting for E. In most situations \( E \) would account for 20% or less of the total variation in y and therefore \( \sigma^2_y \) would be within 10% of the population standard deviation. Thus the \( \beta \) coefficients are interpretable as the approximate proportion of a standard deviation change in y for a standard deviation change in E.

We consider a general situation for a polymorphism where \( p \) is the frequency of the rare allele. Assuming that the polymorphism is in Hardy-Weinberg equilibrium, then the genotype frequencies of aa, aA and AA are \( p^2 \), 2p(1 – p) and (1 – p)^2, respectively. Accordingly, the proportions of individuals in the two genetic groups are \( p^2 \) and 1 – \( p^2 \) for a recessive model, and \( p(2 - p) \) and \( (1 - p)^2 \) for a dominant model, respectively. To study the effect of the environmental exposure on the association of the outcome variable with this genetic factor, we test the null hypothesis that the regression slopes in the genetic sub-groups are equal. If \( n \) individuals are studied, then the test statistic (Appendix) is distributed as a \( F \)-distribution with degrees of freedom 1 and \( n - 4 \) under the null hypothesis, and a non-central \( F \)-distribution with degrees of freedom 1 and \( n - 4 \) under the alternative hypothesis. The non-centrality parameter is

\[ n \left( \frac{2}{\sum_{i=1}^{2} \beta_i^2 p_i} - \frac{1}{\sum_{i=1}^{2} \beta_i p_i} \right)^2. \]

where \( p_i \) and \( p_2 \) are the proportions of individuals in each group. The above expression of the non-centrality parameter can be simplified re-written as \( n p_2 (\beta_1 - \beta_2)^2 \), which means that we can always set the higher risk group as the first group without affecting the sample size and power calculations.

In this paper we adopt the definition of the non-centrality parameter as given by Rencher,9 S-Plus10 and SAS.11 However, in some papers,12,13 it is defined as \( \sqrt{\varphi (k + 1)} \), where \( \varphi \) is the non-centrality parameter defined above, and \( k \) is the numerator degrees of freedom of the test statistic.

Under the situation that the two slopes are equal, we can study the association of the outcome variable with the genetic factor where \( E \) is included as a confounding factor, i.e. to test whether the two intercepts are equal. If the slopes are not equal, then testing the equality of the intercepts is misleading. The test statistic (Appendix) follows a \( F \)-distribution with degrees of freedom 1 and \( n - 3 \) under the null hypothesis, and a non-central \( F \)-distribution with degrees of freedom 1 and \( n - 3 \) under the alternative hypothesis. The non-centrality parameter is

\[ np_2 (\alpha_1 - \alpha_2)^2. \]

Using the distribution and the non-centrality parameter, we are then able to calculate power to detect an interaction effect or alternatively the sample size necessary to detect a given interaction with fixed power and significance. We have not adopted any specific parametric model for describing the interaction. Instead in the results and figures we present power calculations over a range of values for \( \beta_1 \) and \( \beta_2 \).

The range of possible values for \( \beta_1 \) and \( \beta_2 \) are derived from the study of the relationship between physical activity and glucose intolerance. This association is typical of quantitative traits that may be influenced by genetic factors, as evidence from ecological and migration studies suggests the possibility of strong gene-environment interactions.14 In a study by Wareham et al.,15 the relationship between physical activity and a continuous measure of glucose intolerance was quantified using an objective measure of energy expenditure and a multivariate approach to correction for measurement error. The corrected regression coefficient relating habitual energy expenditure to the 2-h plasma glucose was -0.72 mmol/l per standard deviation of the physical activity level, the ratio of the total energy expenditure to basal metabolic rate. The 95% CI for this coefficient were -0.35 to -1.15 mmol/l per standard deviation. As the population standard deviation for the 2-h plasma glucose was 2.2 mmol/l, we may then express this coefficient standardized for the dependent variable too, resulting in a central estimate of -0.33 with 95% CI of -0.16 to -0.52. In the analysis of plausible values for \( \beta_2 \), we have, therefore taken 0.1 to 0.5 as the range of overall effect that would be of interest in the study of gene-environment interactions. We have simplified the reporting of associations by only considering positive associations, as the results would be symmetrical for associations that were in the opposite direction. This range of \( \beta_2 \) values is plausible and would include the central estimates from other studies that have examined the association between continuous outcomes and continuous exposures. For example in the Intersalt study16 the pooled regression coefficient relating 24-h sodium excretion to systolic blood pressure was 0.0354 mm Hg/mmol sodium per day. As the standard deviation of the systolic blood pressure in the UK centres was approximately 15 mm Hg and the standard
deviation of the sodium excretion was 50 mmol per day, this can be converted to a standardized $\beta_2$ value of 0.12, which is within the range we have selected to examine. Although it is possible that stronger effects would be of interest, there are at present few examples of such strong associations and we have limited our attention to those that are less than 0.5.

Results

Figure 1 shows how sample size and the power to detect an interaction for a given allele frequency of the rare allele (5%) vary according to the ratio of the standardized regression coefficients relating the environmental exposure to the outcome in the genetic sub-groups. Using the range of values for $\beta_2$ from the example of glucose intolerance and physical activity, the figure shows that when the effect in those with the common allele is large ($\beta_2 = 0.5$) and there is a moderately strong interaction such that the individuals with the rare allele have a slope that is twice as great, then a study of under 1000 people would be sufficient to detect this interaction with power of greater than 90%. However, if the effect size in those with the common allele were much smaller ($\beta_2 = 0.1$), then even a study of 8000 individuals would be underpowered to detect a doubling of this effect size in those with the rare allele. Figure 1 also shows how power is markedly increased if the interaction is very strong. When the effect size in those with the common allele is small ($\beta_2 = 0.1$), but the effect in those with the rare allele is five times stronger (rather than twice as strong in the previous example), then a study of 600 people would have a power of 80% to detect the interaction.

The example in Figure 1 was constrained by the frequency of the rare allele which was fixed at 5%. Figure 2 shows how the power to detect an interaction effect of 2 with a moderate effect ($\beta_2 = 0.25$) is affected by alterations in the allele frequency. Two different genetic models are considered (recessive and dominant), but the same graph can be used to estimate power and sample size for both. Using the example of the association between physical activity and glucose intolerance where the central estimate was approximately equal to that considered in this figure, then a study of 2000 individuals would have more than 80% power to detect a doubling of this effect size in a sub-group of individuals with the rare allele which occurred with a frequency of more than 5% and was dominant. However, only very large studies (6000 individuals) would be powered sufficiently to detect an interaction of the same magnitude if the rare allele had its biological effect in a recessive manner, and even then the rare allele frequency would need to be high (15%).

In Figures 1 and 2 the magnitude of the gene-environment interaction was considered to be relatively strong as the ratio of the two regression slopes was assumed to be at least 2. Although such strong gene-environment interactions may exist, in any given situation the strength of the interaction will not be known at the point at which power and sample size are being considered. Therefore, we have calculated sample size for possible interactions ranging from close to 1 up to 3. Figure 3 shows a series of sample size plots in which the ratio of $\beta_1 : \beta_2$ is varied for a range of plausible values of $\beta_2$ with power fixed at 80% and significance at 5%. Separate graphs are presented for alleles of differing frequencies. When the rare allele is common (30%) and dominant, then if the effect size is large ($\beta_2 = 0.5$) in those with the common allele, a study of 5000 would be powered to detect an effect in those with the rare allele which was only about 1.2 times greater. Such a small interaction may however, be biologically important and of potential public health significance as the polymorphism is common and has a large effect. Conversely, only an enormous study of almost 50 000 people would be sufficient to detect an interaction effect of 3 for an uncommon dominant allele (0.5%) if the effect in those with the common allele was small (i.e. $\beta_2 = 0.1$). In this situation, one would need to question whether such interactions were worth detecting.

Discussion

In this paper we have presented the formulae and graphs necessary to calculate the statistical power and sample size that is required to study the interaction between a genetic factor and a continuous environmental exposure on a continuously distributed outcome variable. 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 key parameters that determine the sample size are the frequency of the genetic factor and the manner in which it has its biological effect i.e. whether it is dominant or recessive. In addition, power and sample size are critically determined by the absolute magnitude of the slope of the regression linking the environmental exposure and the outcome in people with the common allele, and the ratio of this slope to that in the subgroup of people with the rare allele. As in the case of the example of the association between physical activity and glucose intolerance, estimates of the overall effect size may already be available from previously published studies and pilot work can relatively quickly establish the allele frequency for candidate polymorphisms. The parameter that is uncertain is

Figure 3 Sample size required to achieve 80% power at 5% significance level to detect gene-environment interactions of varying magnitude by different frequencies of the rare allele
the strength of the interaction. This is, of course, the outcome of the study, and in settling on a given value to calculate power one would need to be guided by consideration of what size of interaction would be of biological importance in a given situation.

The fact that power critically depends upon the magnitude of the association between the environmental exposure and the outcome is an argument for utilizing exposure measurement instruments that have small degrees of error, because less precise instruments will result in attenuated regression coefficients, making it harder to detect gene-environment interactions. Given that the cost of epidemiological studies is determined not only by the total sample size but also by the cost of measuring the main exposures, the balance between investing in large studies with imprecise but inexpensive exposure measurement compared to smaller studies with expensive but more precisely measured exposures becomes critical in planning future studies to detect possible gene-environment interactions.

Acknowledgement
Dr Wareham is an MRC Clinician Scientist Fellow. The work of Dr Wong was supported by the British Council and the Royal Society. The idea for this paper was conceived by all authors, who each contributed to the analysis and writing of the paper. NJW will act as guarantor for the paper. There are no conflicts of interest.

KEY MESSAGES
• Existing power and sample size calculations exist for examining interaction in case-control studies.
• This paper presents power and sample size calculations for gene-environment interaction studies in which both the environmental exposure and the outcome are continuous.
• Power is dependent upon:
  the frequency of the genetic polymorphism and whether its biological effect is dominant or recessive,
  the magnitude of the interaction effect, expressed as the ratio of the slopes of the genotype-specific regression coefficients between exposure and outcome,
  the absolute slope of these regression coefficients.

References

Appendix
The test statistic, for testing $H_0: \beta_1 = \beta_2$, is equal to

$$F_\beta = \frac{Y'X(X'X)^{-1}X'Y - Y'Xb_0(X'Xb_0)^{-1}X'b_0Y}{(Y'Y - Y'X(X'X)^{-1}X'Y)/(n - 4)}$$

where $n$ is the sample size, $Y = (y_1, y_2, ..., y_n)'$, $X$ is the design matrix$^{17}$ accommodating the linear regression models in this paper,

$$X = \begin{bmatrix} 1 & 0 & x_1 & 0 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_2 & 0 & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_{k-1} & 0 & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_n \end{bmatrix}$$
The article authored by Luan et al. is most timely. A primary aim of a large National Cohort study currently being designed under the auspices of the United Kingdom Medical Research Council, the Wellcome Trust and the Department of Health is to investigate the role of gene-environment interactions in the aetiology of complex diseases. The study currently aims to recruit and track approximately 500,000 middle-aged subjects across the UK. It represents a major investment in the future of British biomedical science, and it is critical that we produce merely quantitative changes in the association between E and G. It is not only present in the underlying biology but, given adequate power, it should also be detectable as a statistical interaction—variation in the estimated association between E and D for different genotypes at G—regardless of the scale of analysis. Strachan refers to this all-or-nothing phenomenon as ‘effect concentration’. However, if different genotypes at G produce merely quantitative changes in the association between E and a continuous or categorical trait, then the scale of analysis becomes critical in determining whether or not a statistical interaction is detected. For example, interactions can be created or obscured by changing from the natural scale to a logarithmic scale or vice versa. Realistically, it is probable that most true biological interactions cause quantitative rather than all-or-nothing effects at the phenotypic level and Luan et al. concentrate on the former scenario in their article.

Ideally, we need to know more about the underlying biology. If a model properly reflecting the biology (including the true scale of biological action) demonstrates that the association between E and the phenotype does vary with genotype at G, then this is valuable aetiological information. Unfortunately, a

X_β is the design matrix when β_1 = β_2,

\[
X_β = \begin{bmatrix}
1 & 0 & x_1 \\
0 & 0 & \vdots \\
1 & 0 & x_k \\
0 & 0 & \vdots \\
1 & 0 & x_n \\
\end{bmatrix}
\]

where x_i is the environmental variable value of individual i, and k is the number of individuals in the first genetic group. Under the null hypothesis, the test statistic F_β follows F-distribution with degrees of freedom 1 and n - 4. Under the alternative hypothesis, F_β has a non-central F-distribution with degrees of freedom 1 and n - 4 with non-centrality parameter

φ_β = np_1p_2(β_1 - β_2)^2.

p_1 and p_2 are the proportions of individuals in the first and second genetic groups, and p_1 + p_2 = 1. For a recessive model and a dominant model, is p_1 equal to p_2 and p(2 - p), respectively.

The power with 5% significance level for fixed values of n, p, β_1 and β_2 can be obtained easily using any statistical software, e.g. in SAS, the command for the power calculation is

1 - PROBF(FINV(0.95,1,n-4),1,n-4,φ_β)

Under the situation that β_1 = β_2, the test statistics, for testing H_0: α_1 = α_2, is equal to

F_α = \frac{Y'X_α (X_α'X_α)^{-1}X_α Y - Y'X_β (X_β'X_β)^{-1}X_β Y}{(Y'Y - Y'X_β (X_β'X_β)^{-1}X_β Y)/(n - 3)}

where X_α is the design matrix when α_1 = α_2 and β_1 = β_2, that is, a n × 2 matrix with all elements in the first column equal to one and x_1, x_2, ..., x_n in the second column. Under the null hypothesis, the test statistic F_α follows F-distribution with degrees of freedom 1 and n - 3. Under the alternative hypothesis, F_α has a non-central F-distribution with degrees of freedom 1 and n - 3 with non-centrality parameter

φ_α = np_1p_2(α_1 - α_2)^2.

Commentary: Gene-environment interactions: fundamental yet elusive

Paul Burton

The article authored by Luan et al. is most timely. A primary aim of a large National Cohort study currently being designed under the auspices of the United Kingdom Medical Research Council, the Wellcome Trust and the Department of Health is to investigate the role of gene-environment interactions in the aetiology of complex diseases. The study currently aims to recruit and track approximately 500,000 middle-aged subjects across the UK. It represents a major investment in the future of British biomedical science, and it is critical that we properly understand the ‘science’ underpinning the detection and interpretation of gene-environment interactions. However, there are a number of significant problems. One of these is the dependency of statistical interaction on ‘the scale one chooses to measure effects’. If environmental exposure E causes disease D only in the presence of abnormal protein P encoded by allele A* of gene G, then this represents an interaction between E and G. It is not only present in the underlying biology but, given adequate power, it should also be detectable as a statistical interaction—variation in the estimated association between E and D for different genotypes at G—regardless of the scale of analysis. Strachan refers to this all-or-nothing phenomenon as ‘effect concentration’. However, if different genotypes at G produce merely quantitative changes in the association between E and a continuous or categorical trait, then the scale of analysis becomes critical in determining whether or not a statistical interaction is detected. For example, interactions can be created or obscured by changing from the natural scale to a logarithmic scale or vice versa. Realistically, it is probable that most true biological interactions cause quantitative rather than all-or-nothing effects at the phenotypic level and Luan et al. concentrate on the former scenario in their article.

Ideally, we need to know more about the underlying biology. If a model properly reflecting the biology (including the true scale of biological action) demonstrates that the association between E and the phenotype does vary with genotype at G, then this is valuable aetiological information. Unfortunately, a
lack of biological knowledge currently impairs our ability to interpret statistical interactions which in turn hinders attempts to learn more about the biology: ‘Catch-22’.6

Two scientific positions would seem tenable. On the one hand, it could be argued that epidemiology has already reached its limits.7 Rather than incurring the opportunity cost associated with an expensive cohort study aimed at teasing out subtle gene-environment interactions of uncertain biological relevance, an argument could be made for cheaper studies to detect large effects (taking advantage of ‘Mendelian Randomisation’8), with the aim of furthering our understanding of the biology at a simple level.3,4 Alternatively, it could be argued that a cohort study based on 500 000 middle-aged individuals will become ever more valuable over several decades: both as a direct source of information and as a sampling frame for nested sub-studies. The real question is not whether we have enough biological knowledge to definitively interpret gene-environment interactions now, but whether our state of knowledge several decades hence will enable carefully crafted models, faithfully reflecting interactions in the underlying biology, to be used to make useful causal inferences based upon data generated directly or indirectly from the proposed cohort. If so, the prospective and longitudinal nature of the assessment of both exposures and phenotypes should prove invaluable and the benefits accruing from the cohort design could substantially exceed its immediate and on-going costs. From this viewpoint, the key issues—currently being worked on by the Protocol Development Committee—are to ensure that the design and conduct of the proposed study are properly thought out, that its aims are realistic given its funding and that the design is flexible enough to cope with the unpredictable course of bioscience over the next few decades.

Given this background, the paper by Luan et al.1 provides a valuable contribution. Firstly, the power estimates are of direct value in their own right, particularly as most previous papers9,10 have focused on binary rather than quantitative traits. Scientists designing studies with the aim of detecting gene-environment interactions for quantitative traits will be able to make direct, or indirect, use of the documented power profiles. In this regard, it is helpful that the profiles are presented for regression coefficients standardized for dispersion of the trait and of the environmental exposure. Secondly, there is an intrinsic value in performing one’s own power calculations from first principles, particularly for large or expensive studies. This paper clearly outlines the methods used, and should enable an equivalent approach to be used in other settings. Thirdly, it is helpful that the authors illustrate how one might use information from previous studies to properly inform such a power calculation. Finally, the paper will be helpful to those designing the UK National Cohort study. Specifically, it will provide provisional power estimates pertinent to the study of gene-environment interaction in the component of the study that will produce results most rapidly: the study of quantitative traits measured at recruitment.

This having been said, power calculations for studies of gene-environment interaction should still carry a cautionary ‘Health Warning’. It is the biology not the statistics that will ultimately tell us the appropriate analytical scale for any given interaction, and until we know whether it is $X$, $\log(X)$ or $1/X$ that we should be investigating, the link between biological and statistical interaction must remain tenuous. Furthermore, as our knowledge of the complex diseases grows, apparently reasonable power calculations reported for earlier studies look increasingly overoptimistic as new layers of complexity emerge and negative or non-reproducible studies predominate. This is as true for gene-environment interactions as it is for genetic main effects. Nevertheless, one of the distinct advantages of presenting a comprehensive power profile in a form equivalent to that of Luan et al.1 is that it makes all assumptions explicit. An assessor can then decide whether these assumptions are credible or unrealistic given current knowledge.

References