Acta Genet Med Gemellol 36:5-20 (1987)

Prospects for Detecting Genotype X Environment Interactions in Twins with Breast Cancer

N.G. Martin[1], **L.J. Eaves, A.C. Heath**

Department of Human Genetics, Medical College of Virginia, Richmond, USA

Abstract. We consider a study of MZ and DZ twin pairs ascertained because one or both twins have a disease. Genotypes at a major locus are known and putative environmental risk factors have been measured for all individuals. The power of the study to estimate the effect on liability of the measured and residual genetic and environmental effects (G_m, G_r, E_m, E_r) and all two-way interactions between them (except $G_r \times E_r$) is estimated by simulation. If liabilities can be indexed on a continuous scale (eg, blood pressure as an index of liability to hypertension), then a study of 600 MZ and 600 DZ pairs would have sufficient power to detect quite subtle interaction effects, even if ascertainment is greatly biased toward MZ twins. If liabilities cannot be measured and only affection status is known, then the power of the study would be much lower, although not impracticably so. There appears to be no advantage in augmenting the twins with a sample of control individuals who have been drawn at random from the population regardless of disease status, at least for the case we have considered in which the disease threshold on the liability scale is assumed to be known without error. The argument is developed in terms of the utility of the design for research into breast cancer.

Key words: Twins, Genotype × environment interaction, Breast cancer, Disease liability, Oncogenes

INTRODUCTION

Most genetic models for the etiology of disease assume either that environmental effects are random or, if they are not random, that they contribute additively with the genotype

[1] Presently at: Queensland Institute of Medical Research, Brisbane, Australia.

https://doi.org/10.1017/S0001566000004542 Published online by Cambridge University Press

to liability for disease. Thus, traditional models for segregation and linkage treat the environment as a random variable whose contribution to disease liability can be summarized adequately by a reduction in the "penetrance" of a given genotype.

There is, however, an extensive animal and plant literature which shows that the additive model for genetic and environmental effects is far from adequate because sensitivity to the environment is itself under genetic control. That is, there are genes which affect the phenotype by making the organism more or less sensitive to particular environmental effects. The genetic architecture of such "genotype × environment interaction" (G × E) has been analyzed extensively in species from fungi to mice [12] with the result that certain important principles have been established:

1) Genes which affect sensitivity to the environment are often quite distinct (ie, are different loci) from those which affect average response over a range of environments [9];

2) Genes affecting sensitivity to the environment have distinct additive and dominance effects from those which affect overall response (ie, they have probably quite different relationships to fitness);

3) Different sets of genes control sensitivity to different specific environmental factors.

What do these findings mean for human disease? In the first place, they warn us that many of the models geneticists use for family resemblance in man may not capture the essence of gene expression in particular cases. Eaves [2], for example, has shown that ignoring the effects of G × E in family data, when it is actually present, may lead to seriously biased estimates of gene frequencies from segregation analyses. Ignoring the effects of G × E in counseling may lead to misleading estimates of risk to relatives [6].

More importantly, however, many of the ideas currently formulated by epidemiologists about the etiology of common disease relate much more closely to notions of G × E interaction than they do to the models for segregation and family resemblance traditionally employed in genetic epidemiology. The notion of inherited vulnerability to psychiatric disorder, for example, amounts to a recognition that certain individuals are genetically more sensitive to environmental stress. The idea that some forms of hypertension may be the result of inherited sensitivity to sodium can only be expressed mathematically in models for G × E interaction.

Nowhere is the need to consider models for G × E interaction more pressing than in the area of cancer genetics. Despite the extensive evidence for familial aggregation, at least for some forms of breast cancer, and the many clues implicating proto-oncogenes in the etiology of breast cancer, the MZ twin concordance is low, around 15%, and the DZ concordance only a few percent points lower. The main known risk factors, early menarche, late first full-term pregnancy (FFTP) and late menopause [4], are probably largely environmental, although they almost certainly also have polygenic components of variation [11]. One explanation consistent with all these facts is that genes and environment are both important in cancer etiology, but not necessarily in a simple additive way. Large genetic effects influencing sensitivity to environmental risk factors to which the individual is exposed relatively rarely would produce low twin concordance [2,6].

Recently, Krontiris et al [7] have found a disproportionately high number of certain "rare" alleles (defined by restriction enzyme techniques) at the Harvey ras oncogene in a sample of cancer patients compared with normal controls. Their patient sample was diag-

https://doi.org/10.1017/S0001566000004542 Published online by Cambridge University Press

nostically heterogeneous but the "rare" alleles appeared to be more frequent in many of the different forms of cancer. In particular, we note that in 13 breast cancer patients, 4 out of 26 alleles were of the "rare" type, a significantly higher frequency than in the control sample (9 out of 230). What sets this finding apart from run-of-the-mill associations between diseases and genetic markers is that many independent lines of evidence prior to this finding had implicated ras and other oncogenes in the etiology of cancer.

Given that certain oncogene variants confer greater risk of cancer, the critical question is how. Two extreme hypotheses can be proposed. The first is that the alleles confer increased risk quite independently of any exogenous factors, simply having a higher probabiltiy of being involved in transformation of a normal to a precancerous cell. We shall call this the constitutive hypothesis. The second hypothesis is that the alleles confer increased risk by causing the individual to be more sensitive to exogenous factors (eg, carcinogens). This hypothesis posits a major role for genotype × environment interaction (G × E) and we shall call it the environmental sensitivity hypothesis. Surprisingly, this second hypothesis is commonly neglected by human geneticists, despite extensive evidence, noted above, for the importance of G × E from the study of animal and plant species.

One way to resolve constitutive and environmental sensitivity hypotheses would be to study genetic and environmental risk-factors in a series of incident breast cancer cases and controls. This approach would be very powerful if all relevant genetic and environmental risk factors had been successfully identified and measured. However, this approach cannot detect interactions involving unmeasured environmental or genetic risk factors, and hence has low power when some of these risk factors have still to be identified.

A more powerful way to distinguish between the constitutive and environmental sensitivity hypotheses is to type polymorphisms and measure environmental factors putatively associated with increased risk, in twin pairs both concordant and discordant for cancer. The theoretical basis for testing the principal and subsidiary hypotheses will be developed formally below and for the present we shall merely note the results to be gained by various comparisons.

By including both monozygotic (MZ) and dizygotic (DZ) twins we can estimate the effects of background genotype on cancer risk, ie, the effect of unidentified genes. It is the principal purpose of the conventional twin study, in which concordance or correlation for a trait is compared in MZ and DZ twins, to estimate the component of variance due to genetic differences. By typing twins at particular loci, this variance can be further partitioned into that due to measured genotype and the residual due to unmeasured or background genotype. Background genotype, for example, could include variability at other, untyped oncogenes, or indeed at a host of other loci with greater or lesser influences on oncogenesis. The advantage of the proposed design is that one can not only estimate the effect of background genotype, but also its interactions with measured genotypes (epistasis) and measured environmental risk factors. If any of these interaction effects, or the main effect of genetic background is large, then one inference would be that one is not typing (all) the right genes.

It can be seen that the study of cancer risk in MZ and DZ twins in whom some, but not all, pertinent genotypes and environmental risk factors are measured, is potentially a most useful design for unravelling etiology. But how powerful is it?

We have been stimulated to attempt an answer to this question by the existence of

https://doi.org/10.1017/S0001566000004542 Published online by Cambridge University Press

the Twin Cancer Registry, compiled by Dr. T.M. Mack and his colleagues at the University of Southern California. They have already identified more than 1400 pairs of twins where at least one has breast cancer. Detailed diagnostic and putative risk factor information is available for about 1200 of these pairs, half MZ and half DZ. We were tempted to ask what the prospect would be of detecting the various types of interaction discussed above (if they exist!) if one could type all these twins for a polymorphic candidate gene like Ha-ras.

A subsidiary question is whether there would be any gain in the power of such a study by including randomly selected controls. One possibility is a matched sample of female controls for whom disease status, major genotype and environmental risk factor data would be collected. Far easier would be to obtain blood samples of the twins' husbands for genotyping, but no comparable environmental risk factor data could be collected for them.

Because there are many questions about the design and power of the proposed study, it was decided to simulate it in some detail in order to assess the consequences of various sampling strategies. Key questions to be answered include:

1) What is the power to detect main effects of the measured and residual genotype (G_m and G_r) and the measured and residual environment (E_m and E_r) and all six of their two way interactions?
2) Is there an advantage in including controls, even if measured environmental indices (E_m) are not available for them (as would be the case if we used husbands as controls)?
3) What is the power of the study when 1) we assume that we can measure (or obtain an index of) liability, and 2) only affection status for individuals is known?
4) How are estimates and power affected by different assumptions about ascertainment? In particular, how are they affected if ascertainment differs in MZ and DZ twins?

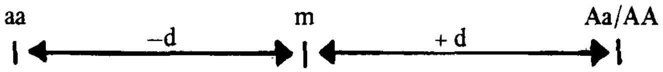
METHODS

Simulation

We have only considered the case of a rare dominant gene predisposing to a disease and we have done all our modelling as if for the rare Ha-ras alleles and breast cancer, ie, with the frequency of the rare dominant $p = 0.05$ and a lifetime risk of 7% [7,16].

In generating the data, components of variance in liability due to G_m, G_r, E_m and E_r are chosen to add to unity. These variance components have the values B_1, B_2, B_3, B_4 respectively, so $B_1 + B_2 + B_3 + B_4 = 1.0$.

For the case of complete dominance, the contribution of a pair of alleles A,a to variation in disease liability is shown in the following picture:



The average liability of the dominant and recessive phenotypes is denoted m. The mean liability of the dominant phenotype is then m + d and of the recessive, m – d. In our simulations m and d are calculated such that the population mean of the major gene effect is zero and the population variance in liability due to the major gene is B_1. Thus, if the frequency of the dominant allele is p and $p + 1 = 1$, then $d = B_1/(1 - (p - q + 2pq)^2)$ and $m = d(q - p - 2pq)$. Major genotypes of MZ twins and of controls are obtained by sampling individuals from a population in Hardy-Weinberg equilibrium. Pairs of DZ twins are generated by first sampling two parents from this population and then sampling a gamete from each of them, once for each twin.

Background (or residual) genotypes are obtained for MZ twins and controls simply by sampling at random from an N(0,1) distribution. For DZ twins they are obtained by sampling from the bivariate normal distribution $N(0,0,1,1, \rho = 0.5)$. Contributions of the measured environment and unmeasured environment are also sampled from N(0,1) distributions, independently for each twin and control individual. Thus, we have only considered the case of additive polygenic variation, a reasonable restriction in view of the low power of the twin study to detect non-additive genetic variance [10]. Likewise our treatment is restricted to environmental influences which are specific to the individual and not shared by the cotwin (ie, E1/ES and not E2/EC).

We can generate all possible two-way interactions between the four main effects by use of the terms b_5 to b_{10} which are the regression coefficients of the phenotype on the appropriate product terms of the main effects. Thus:

b_5 generates $G_m \times E_m$ interaction
b_6 generates $G_m \times E_r$ interaction
b_7 generates $G_r \times E_m$ interaction
b_8 generates $G_r \times E_r$ interaction
b_9 generates $G_m \times G_r$ interaction
b_{10} generates $E_m \times E_r$ interaction

One of these interactions, $G_r \times E_r$, the interaction of residual genetical and specific environmental effects (generated by the coefficient b_8), is completely confounded with E_r (B_4) and therefore cannot be estimated separately from it. Furthermore, its presence would violate the assumption of multivariate normality implicit in the estimation procedure and might therefore bias the estimates of other parameters. How serious this problem might be and how to deal with it need further investigation [5]. For the present, we shall avoid the problem by omitting this type of interaction from further consideration. In any case, it is the interactions involving the measured effects which are the focus of our attention and the stimulus for this work.

Having obtained the major genotype of an individual and measures of the background genotype, measured and residual environmental deviations, the liability (X) for the i'th individual is calculed as:

$$(1)\quad X_i = m + s_i \cdot d + b_2 \cdot g_i + b_3 \cdot em_i + b_4 \cdot er_i + s_i \cdot b_5 \cdot em_i + s_i \cdot b_6 \cdot er_i$$
$$+ b_7 \cdot g_i \cdot em_i + s_i \cdot b_9 \cdot g_i + b_{10} \cdot em_i \cdot er_i$$

where g_i is the residual polygenic deviation, em_i and er_i are the measured and unmeasured environmental deviations and s_i is -1 if the i'th individual is of recessive phenotype and $+1$ if of dominant phenotype. In the above, b_2, b_3, b_4 are the positive square roots of the quantities B_2, B_3, B_4 as previously defined.

Because the present exercise was prompted by the case of breast cancer, we wished to generate data in which approximate 7% of cases were affected. This would correspond to a standard normal deviate of 1.48 if liability were distributed N(0,1). However, the expected population variance of liability will differ from unity as a complex function of the parameters chosen to generate the data. Furthermore, since a major dominant gene is segregating, the distribution will be skewed. We calculated an approximate value of the population variance V(P), ignoring the covariance terms between effects and took the threshold T above which individuals are deemed affected as:

$$(2)\quad T = 1.48/\sqrt{V(P)}$$

For the range of parameter values we considered, this procedure generated samples of unselected controls in which 5.5-7% of individuals were affected.

Liability scores for pairs of MZ and DZ twins and for control individuals are generated according to (1) and the procedures outlined above. All control individuals are accepted into the sample. A twin pair is accepted into the sample if at least one of them has a liability $X \geqslant T$. Twins and controls accepted into the sample have the following information about them stored for data analysis: whether they are affected or unaffected, their liability X, their genotype at the major locus, and their measured environmental index, em.

Data may also be generated with different values of Π, the probability of inclusion in the sample given that an individual is affected. Thus, for values of $\Pi < 1$, a twin pair is only included in the sample if at least one of them has $X \geqslant T$ and, for this individual, a random number from a U(0,1) distribution is $\leqslant \Pi$. MZ and DZ twins may be generated with the same or with different values of Π but the control sample, by definition, is never subject to any selection.

To test the power of the proposed study under the most stringent conditions, data were simulated in which all two-way interaction effects except $G_r \times E_r$ (see above), were generated.

Since our tentative experimental design calls for n pairs of MZ, n pairs of DZ and n control individuals, we have simulated $r = 5$ replicates of unit size n for each design modification. The three designs considered were:

Design I: MZ, DZ, controls
Design II: As in I but no measured environmental indices for controls
Design III: MZ, DZ, no controls

Thus, design I envisages collecting the same detailed environmental risk factor data in controls as we have for the twins, design II recognizes the difficulty of this and design III questions whether there is any benefit in having controls at all.

Model Fitting

Models incorporating different main effect and interaction parameters may be fitted to the raw observations by the method of maximum likelihood. We shall consider two cases: 1) liabilities are known for individuals; 2) liabilities are not known, only whether an individual is affected or unaffected.

1) Liabilities are known

We follow the approach of Lange et al [8] which assumes that the distribution of scores in

https://doi.org/10.1017/S0001566000004542 Published online by Cambridge University Press

a pedigree is multivariate normal, conditional upon the measured genotypes and environments. For a given pedigree of n individuals we define a vector of observed scores, **x** and a corresponding vector of expected values, E**x**. The values of the elements of E**x** will depend upon the known genotype at the major locus and the measured environmental index.

Similarly, we define the expected covariance matrix, **Σ**, of individuals in the pedigree. The elements of **Σ** will depend upon the major genotypes, the environmental indices and the relationship between the individuals, in our case whether twins are MZ or DZ.

For a given E**x** and **Σ** the log likelihood of obtaining the pedigree of individuals of given observation vector **x** is:

$$(3)\quad L = -\frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x} - E\mathbf{x})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - E\mathbf{x}) + \text{constant}$$

The joint log-likelihood of obtaining a sample of pedigrees of varying measured genotypes and environmental indices, some MZ twins, some DZ, and some singleton controls, is simply the sum of the log-likelihoods of the individual pedigrees. Estimation involves the selection of parameter values under a given hypothesis which maximize the joint likelihood of observing the given set of pedigrees. Conventional methods of minimization may be used to minimize −L with respect to the parameters of the model. We use the optimization routine E04JAF from the NAG Library (NAG, Mark 11) which allows the user to specify bounds for parameters but no other constraints.

The elements of E**x** are calculated as:

$$E(X_i) = m + s_i\,d + em_i \cdot (b_3 + s_i\,b_5)$$

The parameters are as defined above, except that they are now replaced by their estimates.

The expected variance-covariance matrix, **Σ**, will contain the appropriate variances for twin 1 and twin 2 as the diagonal elements and the expected covariance of the twins in the off-diagonal element. For controls, **Σ** will contain only one element, namely the variance appropriate for an individual of that major genotype and environmental index. The expected variance for the i'th individual is:

$$V(X_i) = B_2 \cdot (1 + s_i\,b_9 + b_7 \cdot em_i)^2 + B_4 \cdot (1 + s_i\,b_6 + b_{10} \cdot em_i)^2$$

The covariance of twins i and j is:

$$W(X_i, X_j) = [B_2 \cdot (1 + s_i\,b_9 + b_7 \cdot em_i)(1 + s_j\,b_9 + b_7 \cdot em_j)]/z$$

where z is 1 for MZ and 2 for DZ twins. Note that since both measured and residual environmental effects are specific to the individual, there is no term in B_4 in the covariance.

Correction for ascertainment

In the above we have assumed that pairs are ascertained at random from the population of twins and then examined to determine their disease status. Clearly this is not the case in our study since ascertainment requires that at least one twin already have breast cancer. The likelihoods computed on the assumption of random sampling, therefore, have to be modified to allow for our ascertainment procedure.

https://doi.org/10.1017/S0001566000004542 Published online by Cambridge University Press

No correction for ascertainment is completely free of assumptions. Typically, it is assumed in human genetic applications that an affected individual has a probability, Π, of being ascertained and that affected individuals in a family are ascertained independently. In our simulated data sets we know that these assumptions are correct but this will not necessarily be the case in the real world. Two extreme cases are normally distinguished [1]. Ascertainment is complete when $\Pi = 1$. The limiting case as $\Pi \rightarrow 0$ is described as "single ascertainment".

The usual correction for ascertainment [3] requires multiplying the likelihood (3) above by the ratio A/B where A is the probability that a pair with a affected individuals will actually be ascertained, and B is the probability that a pair of a given measured genotype and environmental index will be ascertained.

A is simply $1 - (1 - \Pi)^a$, where $a = 1$ or 2 in the case of twins, and is the probability that as least one individual is ascertained from a pair in which a are affected.

The probability that both members of a randomly selected twin pair will be affected, given their genotype and environmental indices is Φ_{11}. Let the probability that the first twin will be affected and the second twin not be Φ_{10}. Φ_{01} is the probability that the first twin will be unaffected and the second twin affected. These probabilities can be calculated from the bivariate normal density function, N.

If $t_i = [T - E(X_i \mid G_i, em_i)]/\sqrt{E(V(X_i))}$, then

$$\Phi_{11} = \int_{t_1}^{\infty} \int_{t_2}^{\infty} N(0,0,1,1, \rho, x_1, x_2) \partial x_i \, \partial x_2,$$

$$\Phi_{10} = \int_{t_1}^{\infty} \int_{-\infty}^{t_2} N(0,0,1,1, \rho, x_1, x_2) \partial x_1 \, \partial x_2 \quad \text{and}$$

$$\Phi_{01} = \int_{-\infty}^{t_1} \int_{t_2}^{\infty} N(0,0,1,1, \rho, x_1, x_2) \partial x_1 \, \partial x_2 .$$

where G_i is the major genotype of the i'th individual and ρ is the correlation between twins predicted under the model. The same threshold value, T, is employed as was used to generate the data (see equation 2 above).

Thus, the probability that a pair will be ascertained, given their genotypes and environmental indices is:

$$B = \Phi_{11}(1 - (1 - \Pi)^2) + (\Phi_{10} + \Phi_{01})\Pi .$$

B is thus the sum of 1) the probability that at least one individual is ascertained from pairs in which both are affected and 2) the probability that the affected individual in pairs where only one is affected will be ascertained. The Φ values are functions of the genetic and environmental parameters of the model. Π can also be estimated as another parameter (and we have done this successfully in a variety of simulations) but in the real world we are more likely to supply a fixed, independently estimated value of Π during our estimation procedure.

2) Liabilities are not known: only affection status is known

If liabilities are not known then the likelihood of a twin pair, given their major genotypes ans environmental indices is one of the three conditional probabilities given above, viz, Φ_{11} if both twins are affected, Φ_{10} if the first is affected and the second not, and Φ_{01} if the first is unaffected and the second is affected. This likelihood must, as for the continuous case, be multiplied by the appropriate correction for ascertainment. For the discontinuous case, then, likelihoods of twin pairs are a direct function of the three Φ's and Π. For control individuals the likelihood is simply the normal integral from minus infinity to t_i (defined above) if the individual is unaffected, or one minus this probability if the individual is affected.

Estimation of power

To test our design under the most exacting circumstances, data sets have been simulated with all main effects and interactions (except $G_r \times E_r$) contributing simultaneously to liability. We first obtain the log-likelihood, L_0, for the full model which contains all the parameters used to generate the data, viz, m, d, B_2, B_3, B_4, b_5, b_6, b_7, b_9, b_{10}. To the same data we then fit seven subsidiary models, omitting in turn d, B_2, b_5, b_6, b_7, b_9 and b_{10}. Preliminary studies showed that there is always ample power to detect E_m (B_3) and E_r (B_4) and omitting them from the model frequently causes considerable numerical difficulties. Log-likelihoods for each of these models, $L_1 - L_7$, are then used to calculate likelihood ratio chi squares:

$$\chi^2 = 2(L_0 - L_i), \text{ each on 1 df,}$$

to test the effect of the parameter omitted from the i'th model. These are summed across replicates to obtain $\Sigma\chi^2$.

To estimate the sample size, N, required to detect each effect at the $\alpha = 0.05$ significance level with a probability (power), β, of 0.95, 0.80 and 0.50, we first obtain an estimate of the noncentrality parameter,

$$\lambda' = \Sigma\chi^2/nr, \text{ and then}$$

$$N = \lambda_{a,b,c}/\lambda',$$

where λ is the non-central chi square value for $\alpha = a$, $\beta = b$, and c degrees of freedom [10]. For $\alpha = 0.5$, $c = 1$ and $\beta = 0.95$, 0.80 and 0.50, λ is 12.995, 7.849 and 3.841 respectively [14].

RESULTS

We shall first consider the results of simulations in which liabilities are assumed to be known. Then we shall consider the power of studies in which only affection status is known. Finally, the consequences of different assumptions about ascertainment will be explored.

Numbers required when liabilities are known

For each of designs I, II and III, r = 5 replicates of n = 600 were generated. Numbers

https://doi.org/10.1017/S0001566000004542 Published online by Cambridge University Press

required to detect effects with 95% power are shown in Table 1 for data generated with two different sets of parameter values, the interaction effects in Table 1(b) being considerably greater than those in Table 1(a). For approximately 80% power, multiply these numbers by 0.6 and for 50% power by 0.3.

Table 1 - Estimated numbers required to detect effects at $\alpha = 0.05$ with 95% power when liabilities are known, there is complete ascertainment, $B_1 = 0.1$, $B_2 = B_3 = B_4 = 0.3$, and

(a) $b_5 = b_6 = b_7 = b_9 = b_{10} = 0.1$

	d	B_2	b_5	b_6	b_7	b_9	b_{10}
Design I	37	25	183	443	931	1421	627
Design II	42	23	251	552	714	1119	618
Design III	50	39	247	308	1302	1627	331

(b) $b_5 = b_6 = 0.1$, $b_7 = b_9 = b_{10} = 0.3$

	d	B_2	b_5	b_6	b_7	b_9	b_{10}
Design I	34	28	211	349	157	345	80
Design II	39	31	211	804	242	386	109
Design III	43	52	231	333	288	415	92

In design I, for the parameter values used in Table 1(a), mean MZ pairwise concordance over five replicates was 16.8% and mean DZ concordance was 9.3%. The "disease" gene was 7.4 times more common in affected than unaffected individuals.

Thus in Table 1(b), we see that in order to detect the direct effect of the major gene on liability (d) at the 5% significance level in 95% of studies, we should need 34 pairs of MZ twins, 34 pairs of DZ twins and 34 control individuals (Design I), or if no environmental index measurements were available for controls we should need 39 in each group (Design II). Alternatively, we could simply use 43 pairs of MZ and 43 pairs of DZ twins (Design III).

Liabilities are not known; only affection status is known

Now we consider what happens when liabilities are not known and the only information we have is whether an individual is affected or unaffected. All data were generated and estimated with the ascertainment probability $\Pi = 1$. For the discontinuous case we found that power was much lower so a unit sample size of n = 2000 was used, with 5 replicates as before. Because power to detect interaction effects of the sizes used in Table 1 was low, data were generated with larger values of $b_5 - b_{10} = 0.3$. Numbers required for both 95% and 50% power are shown in Table 2. In the 10,000 pairs generated in each category in design I, the MZ concordance for the disease was 19%, the DZ concordance was 9.7% and the relative frequency of the "disease" gene in affected vs unaffected controls was 9.33.

https://doi.org/10.1017/S0001566000004542 Published online by Cambridge University Press

Table 2 - Estimated numbers required to detect effects at $\alpha = 0.05$ with 95% and 50% power when liabilities are not known, there is complete ascertainment and $B_1 = 0.1$, $B_2 = B_3 = B_4 = b_5 = b_6 = b_7 = b_9 = b_{10} = 0.3$.

$\beta = 0.95$	d	B_2	b_5	b_6	b_7	b_9	b_{10}
Design I	1525	238	2260	6762	1587	5274	1067
Design II	1879	538	1773	14315	4582	9174	1124
Design III	1954	283	1558	3823	2218	4944	961
$\beta = 0.50$							
Design I	450	70	668	1998	469	1559	315
Design II	555	159	524	4231	1354	2711	332
Design III	557	83	460	1130	655	1461	248

How accurate are the parameter estimates from the discontinuous case? The mean and sd of the estimates of each parameter under the full model over the five replicates are shown in Table 3 beneath their expected values.

Table 3 - Means (and sd's) of five estimates of parameters under the full (correct) model when only affection status is known and ascertainment is complete. Expected values of parameters are shown.

	m	d	B_2	B_3	B_4	b_5	b_6	b_7	b_9	b_{10}
Exp. value	0.43	0.53	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30
Design I	0.43	0.56	0.29	0.28	0.31	0.25	0.23	0.34	0.29	0.26
sd	0.11	0.13	0.05	0.06	0.03	0.10	0.16	0.12	0.12	0.06
Design II	0.50	0.53	0.24	0.44	0.38	0.26	0.19	0.25	0.26	0.24
sd	0.07	0.09	0.03	0.04	0.05	0.12	0.14	0.07	0.12	0.03
Design III	0.47	0.48	0.33	0.34	0.30	0.36	0.39	0.27	0.33	0.32
sd	0.06	0.12	0.04	0.08	0.05	0.10	0.11	0.11	0.13	0.12

Effect of ascertainment bias

A further potential complication arises if ascertainment differs between MZ and DZ twins. To investigate the effect of such ascertainment bias on estimation we have simulated one of the more difficult situations we could imagine. It has been alleged that Π is greater for MZ than for DZ twins. We simulated data for the continuous case using the same parameters as in Table 1(a) but with $\Pi_{MZ} = 0.20$ and $\Pi_{DZ} = 0.05$. Thus a twin pair is only included in the sample if at least one twin is affected and, for this individual, a random number from a U(0,1) distribution is $\leqslant \Pi$. In designs I and II the control sample, by definition, is never subject to any selection.

https://doi.org/10.1017/S0001566000004542 Published online by Cambridge University Press

Using the methods above, models were then fitted to these data sets assuming $\Pi_{MZ} = \Pi_{DZ} = 0.10$. As before, r = 5 replicates of n = 600 were generated. Results are shown in Table 4. In order to see whether biassed ascertainment causes bias in parameter estimates, means and standard deviations of estimates from five replicates of each design were calculated and these are shown in Table 5 beneath their expected values.

Table 4 - Effects of ascertainment bias when liabilities are known. Estimated numbers required to detect effects at $\alpha = 0.05$ with 95% power when $B_1 = 0.1$, $B_2 = B_3 = B_4 = 0.3$, $b_5 = b_6 =$ $= b_7 = b_9 = b_{10} = 0.1$. Data were generated with $\Pi_{MZ} = 0.20$ and $\Pi_{DZ} = 0.05$. Estimation assumed $\Pi_{MZ} = \Pi_{DZ} = 0.1$.

	d	B_2	b_5	b_6	b_7	b_9	b_{10}
Design I	35	28	190	329	1216	1464	443
Design II	39	24	240	691	803	1489	792
Design III	48	43	195	371	1673	3304	458

Table 5 - Means and sd's of five estimates of parameters under the full (correct) model when data were generated with $\Pi_{MZ} = 0.20$ and $\Pi_{DZ} = 0.05$, and parameters were estimated with $\Pi_{MZ} = \Pi_{DZ} = 0.10$.

	m	d	B_2	B_3	B_4	b_5	b_6	b_7	b_9	b_{10}
Exp. value	0.43	0.53	0.30	0.30	0.30	0.10	0.10	0.10	0.10	0.10
Design I	0.44	0.53	0.28	0.29	0.31	0.10	0.11	0.08	0.09	0.11
sd	0.04	0.05	0.03	0.01	0.01	0.01	0.03	0.04	0.03	0.04
Design II	0.40	0.54	0.34	0.29	0.31	0.10	0.08	0.11	0.07	0.08
sd	0.01	0.02	0.04	0.03	0.02	0.01	0.02	0.06	0.04	0.03
Design III	0.46	0.53	0.28	0.28	0.30	0.11	0.11	0.08	0.07	0.12
sd	0.03	0.03	0.03	0.04	0.01	0.02	0.02	0.05	0.02	0.04

DISCUSSION

We set out to assess, by detailed simulation, the utility of a study of 600 MZ and 600 DZ twin pairs in which one or both have a disease (breast cancer) and for whom genotypes at a putative major locus and measurements for putative environmental risk factors are available. In particular, we wished to know the power to detect interactions between measured and residual genetic and environmental effects (G_m, G_r, E_m, E_r) which might influence disease liability. We also wished to know whether there was any value in augmenting the twins with a sample of controls for whom environmental risk data might, or might not, be available. Further, how might the estimation procedure be influenced by differences in the ascertainment of MZ and DZ twins?

These questions have been considered under the best and worst case assumptions; in the best case we can assign a liability to every individual, in the worst case, no such information is available and we only know whether an individual is affected or unaffected. The power to detect effects when we have an index of liability which is correlated with the true liability to varying degrees must lie in between.

The pairwise concordance rates are not unlike those obtained for breast (and other) cancer, in that the concordances are both small and not very much higher in MZ than DZ pairs. The important point is that, in data ascertained through probands, small twin concordance and small MZ-DZ differences in concordance can still be compatible with large genetic main and interaction effects. To conclude, as some have done, that low MZ and DZ concordances belie an important role for genes, is unwarranted.

We have shown (Table 1) that our study would provide sufficient power to detect quite subtle interactions of each kind when liabilities are known as, for example, if the disease in question were hypertension and blood pressure were a measure of liability. The power to detect a main effect of the measured major gene on liability is remarkable and in other simulations we have shown that our proposed study would have 80% power to detect the effect of a typed gene accounting for as little as 1% of the variance in liability. Furthermore, main and interaction effects can still be detected with hardly any less power and no notable bias if ascertainment is grossly biased in favour of MZ twins (Tables 4 and 5).

When we turn to the worst case in which no information on liability is assumed beyond what can be inferred from affection status, then we see (Table 2) that our power to detect the interaction effects and even the direct effect of the major gene, plummets. Even so, all parameters can still be estimated with little evidence of bias (Table 3). Three things should be remembered, however.

First, even interaction effects generated with $b_5 - b_{10} = 0.3$ are not large. The expected polygenic variance of liability in the dominant phenotype is approximately 3.5 times that of the recessive and the environmental variance of the dominant about 1.8 times greater than the recessive. Over all, the variance in liability of the dominant is approximately twice that of the recessive. These effects fall far short of the dramatic switching on and off of gene action by major environmental or epistatic triggers, of which there are now many well documented examples.

Second, all the required numbers we quote are for 95% power of detecting an effect at the 5% level of significance. This is a powerful experiment indeed. In the second part of Table 2 we also give the numbers required for $\beta = 0.5$ and we can see that our proposed study would have approximately a 50% chance of detecting all main interaction effects except $G_m \times E_r$ and $G_m \times G_r$. Given the potential importance of any positive finding, one could make a convincing case that an experiment with 50% power is worth doing. Remember too that we have set ourselves the demanding task of detecting the effect of a major gene accounting for only 10% of the variance in liability. In the zeitgeist that currently prevails, many experimenters would begin with the expectation that the direct effect of a major gene would account for at least 20% of the variance in liability and we would certainly have sufficient power to detect effects of this magnitude.

Finally, we have only considered the most stringent case in which we have tried to detect each interaction effect against the background of the four main effects and the other four interaction effects. Complicated as the real world is, one wonders whether it is

https://doi.org/10.1017/S0001566000004542 Published online by Cambridge University Press

likely to be this cruel to the experimenter. Nevertheless, the experimenter begins his analysis from a position of ignorance and must first fit the full model in order to see which effects can be eliminated from further consideration.

As to the value of including controls, with or without measured environmental indices, one surprising feature of Tables 1 and 2 is that, in many respects Design 2 appears to be the worst, particularly when liabilities are not known (Table 2), and for the detection of $G_m \times E_r$ throughout. Comparing Design I (with controls) and Design III (no controls), the gain in having controls seems very marginal indeed, whether liabilities are known or not. It seems that including controls without environmental indices is worse than no controls at all, as if the demands made on the twin data to explain liability in controls exceed the contribution to total information made by the measured genotypes of the controls.

Our results regarding controls may, however, reflect the way in which we have defined them. Our "controls" are simply a random sample of the population measured for genotype at the marker locus, disease status and, possibly, the environmental risk factor. However, in many epidemiological studies, controls are selected as being free from the disease. How this change in definition would affect our results and conclusions is not clear but will be the subject of future simulations.

It should be remembered that all the required numbers on which the tables are based are estimates and therefore subject to stochastic error. Since the number of replicates of each study is only five, it is possible that stochastic error is playing a larger role in our results than we would like. However, it is important to realise that these simulations are extremely computer intensive. Each row of the tables represents as much as eight hours of CPU time on an IBM 3081D machine. Clearly it would have been desirable to use a larger number of replicates for each condition and to extend the study to a wider range of conditions and parameter values. It is not easy to predict the extent to which our conclusions, obtained for a set of parameter values which seem reasonable for the case of breast cancer, will generalise to parameter values appropriate for other diseases. The purpose of our paper has not been to provide definitive guides for the conduct of research into breast cancer or any other particular disease, but to show how such projected studies can be simulated in some detail in order to assess their potential for elucidating complex issues of G × E interaction.

We have simulated a case in which we have not taken account of ascertainment which is grossly biassed. The "true" probability of ascertainment of an affected MZ twin (0.2) was four times that of an affected DZ twin (0.05), but we assumed for the purposes of estimation that this probability was equal (0.1) in the two groups. Ascertainment biases in the real world are unlikely to be more extreme than this. It can be seen that main and interaction effects are still detected with hardly any less power (Table 4) than when data were generated and models fitted assuming complete ascertainment ($\Pi = 1$) throughout, as was the case in Table 1(a). Neither is there any notable bias in the estimates (Table 5). It was a general finding of our simulations that estimates of genetical and environmental parameters are remarkably insensitive to false assumptions about the value of Π when the true value of Π lies in the range 0.001 – 0.50. Since this is the range in which true values are almost centain to lie, it seems unlikely that ascertainment bias, whether due to age or zygosity, is likely to prove a major obstacle to our aims. It is unclear, however, what effect violations of the assumption of independent ascertainment

https://doi.org/10.1017/S0001566000004542 Published online by Cambridge University Press

would have on parameter estimation.

Taking all the above into account it seems that a study of the dimensions proposed might have a reasonable chance of finding important main and interaction effects of genes influencing liability to breast cancer if they were there. This chance would be improved if we had any information on liability. A most important topic for further development then is the relationship of liability to age-of-onset and any measures of severity such as laterality, aggressiveness and rate of metastasis. Such relationships are likely to be complex but given good population data on age-specific incidence, prevalence, survival time and ascertainment probability, classified by laterality and other diagnostic data, it seems reasonable to suppose that more information could be gleaned about liability than mere presence or absence of the disease [15].

Use of the design to detect G × E interactions in diseases (eg, hypertension) for which continuous indices of liability (blood pressure) are readily available is an attractive proposition and only awaits the identification of suitable polymorphic candidate genes.

Acknowledgements. This work was first presented at the Genetic Analysis Workshop IV held at Snowbird, Utah on October 8, 1985 and was supported by NIH grants GM32732, GM30250, AA06781, AGO4954. We are grateful to Drs. G. Carey, D. Deapen, K. Kendler, T. Mack and D. Thomas for valuable discussions.

REFERENCES

1. Cavalli-Sforza LL, Bodmer WF (1970): The Genetics of Human Populations. San Francisco: WH Freeman.
2. Eaves LJ (1984): The resolution of genotype × environment interaction in segregation analysis of nuclear families. Genet Epidemiol 1:215-228.
3. Elston RC, Sobel E (1979): Sampling considerations in the gathering and analysis of pedigree data. Am J Hum Genet 31:62-69.
4. Henderson BE, Pike MC, Ross RK (1984): Epidemiology and risk factors. In G Bonadonna (ed.): Breast Cancer: Diagnosis and Management. New York: Wiley.
5. Jinks JL, Fulker DW (1970): Comparison of the biometrical genetical, MAVA and classical approaches to the analysis of human behavior. Psychol Bull 73:311-349.
6. Kendler KS, Eaves LJ (1986): Models for the joint effect of genotype and environment on liability to psychiatric illness. Am J Psychiatry 143:279-289.
7. Krontiris TG, DiMartino NA, Colb M, Parkinson DA (1985): Unique allelic restriction fragments of the human Ha-ras locus in leukocyte and tumor DNAs of cancer patients. Nature 313:369-373.
8. Lange K, Westlake J, Spence MA (1976): Extensions to pedigree analysis III. Variance components by the scoring method. Ann Hum Genet 39:485-491.
9. Lewontin RC (1974): The Genetic Basis of Evolutionary Change. New York: Columbia UP.
10. Martin NG, Eaves LJ, Kearsey MJ, Davies P (1978): The power of the classical twin study. Heredity 40:97-116.
11. Martin NG, Eaves LJ, Eysenck HJ (1977): Genetical, environmental and personality factors influencing the age of first sexual intercourse in twins. J Biosoc Sci 9:91-97.
12. Mather K, Jinks JL (1982): Biometrical Genetics (2nd ed). London: Chapman and Hall.
13. Numerical Algorithms Group (1978) E04JAF. Oxford: NAG, Mark 8.
14. Pearson ES, Hartley HO (eds): Biometrika Tables for Statisticians (Vol. 2). London: Cambridge UP.

15. Thomas DC, Langholz B, Mack T, Deapen D, Floderus-Myrhed B (1986): Bivariate lifetable models for analysis of gene-environment interaction in twins (submitted).

16. Williams WR, Anderson DE (1984): Genetic epidemiology of breast cancer: segregation of 200 Danish pedigrees. Genet Epidemiol 1:7-20.

Correspondence: Dr. N.G. Martin, Queensland Institute of Medical Research, Bramston Terrace, Brisbane 4006, Australia.

https://doi.org/10.1017/S0001566000004542 Published online by Cambridge University Press