J. psychiat. Res., Vol. 21, No. 4, pp. 639–647, 1987.
Printed in Great Britain.

0022-3956/87 $3.00+.00

INCLUDING THE ENVIRONMENT IN MODELS FOR GENETIC SEGREGATION

L. J. EAVES*

Department of Human Genetics, Medical College of Virginia, Richmond, Va 23298, U.S.A.

(*Received* 1 *December* 1986; *revised* 18 *February* 1987)

Summary—Current models for linkage and segregation analysis do not allow for many of the different ways in which the environment may affect risk for a psychiatric disorder. A variety of mechanisms for the contribution of the environment and the correlation and interaction of environment effects with genetic differences are described. The approach to including these effects in models for segregation and linkage is outlined.

INTRODUCTION

ONE OF THE most appealing facets of genetic epidemiology and human population genetics in the last 15 years has been the development of models for non-genetic inheritance. Genetic epidemiologists have devised a variety of models for the effects of the environment on continuous variables (MORTON, 1974; EAVES, 1976a,b; RAO *et al.*, 1976; RICE *et al.*, 1978; HEATH *et al.*, 1985). The seminal work of CAVALLI-SFORZA and FELDMAN (1973, 1981) has extended the mathematical models of classical population genetics to the non-genetic transmission of discontinuous variables. The latter approach has been independent of the others for most practical purposes and has still to be incorporated into the mainstream of genetic epidemiology. However, psychiatric genetics deals primarily with discontinuous variables and clinicians are justifiably uneasy about "transforming away" discontinuities (for example, by switching to a multifactorial threshold model) merely to satisfy mathematical convenience. The cost of this to the genetic study of common disease has been the adoption of "naive monogenism" which sees the key to psychiatric disorder in the routine application of standard methods of segregation and linkage analysis with little regard for the fact that the principal purpose of the brain is the storage and exploitation of information presented by the social and physical environment.

The present state of theory relevant to the analysis of psychiatric disease is quite comparable to that of quantitative genetics in the early 1970's. There was an excellent genetic model for family resemblance, due primarily to FISHER (1918), but virtually no mathematical theory of cultural inheritance. This deficiency was remedied during that decade by the work already cited. Unfortunately, the progress in understanding the theory of non-genetic inheritance applied to discontinuous traits has not yet had as great a practical impact on the subject as path-analytic approaches have had on the analysis of continuous variation. Although CAVALLI-SFORZA and FELDMAN'S (1973) early theory of cultural inheritance

*To whom reprint requests should be addressed.

included parameters for genetic transmission and genotype × environment interaction most subsequent data analysis has focused on purely cultural models and ignored the contribution of genetic differences to family resemblance (CAVALLI-SFORZA *et al.,* 1982). Thus these important ideas have been seen as opposed to, rather than as supplementing, existing models of segregation and linkage.

This paper outlines the principal issues relating to the joint analysis of genetic and environmental variables affecting segregation of discontinuous traits in a form sufficiently rigorous to make the modification of existing methods a matter of routine derivation, but sufficiently transparent to the substantive questions about where the environment comes from and how it works. We begin with the basic proposition of segregation and linkage analysis that at least one gene has effects large enough to stand out against the background of other effects, and then consider how models may be changed to incorporate the effects of the environment.

THE TWO-LOCUS TWO-TRAIT GENETIC MODEL

Many of the basic issues can be explored within the framework of a model in which either or both of two loci may contribute to liability for either or both of two traits. We begin by outlining a fairly general form of this model for the two locus case, and then consider how the model may be modified to allow for various kinds of environmental effect. The model distinguishes between "risk" ($0<\phi<1$) which is the probability that an individual will express a particular form of the disease or trait and "liability" ($-\infty<L<\infty$) which is the latent biological dimension which is supposed to be affected directly by genes and environment. The relationship between liability and risk is specified by the "penetrance function", $\phi = f(L)$. For our purpose we assume that the penetrance function is logistic, i.e.

$$\phi = 1/(1+\mathrm{e}^{-L})$$

but other functions may be used if desired. The basic model assumes that the liability of an individual for each trait may be due entirely to the additive, dominant and epistatic effects of the two loci, although in any particular application the model for gene action may not require the epistatic parameters. Indeed, in many applications even the effects of the second locus may not be required. The important point is that the model be sufficiently general to admit a number of informative variations including those which allow specification of non-genetic effects and their interaction with genotype.

Table 1 presents the full model for genetic effects on liability for the two locus model for each of nine genotypes at a pair of diallelic loci A/a and B/b. Following the convention employed by MATHER and JINKS (1983), we let the upper case letter denote the allele which increases liability. The genetic model involves parameters for the mid-point between homozygotes, m, additive deviations at both loci, d_a and d_b, heterozygous effects, h_a and h_b expressed as deviations from the mid-point of the homozygotes, and epistatic effects. The epistatic effects are represented in a general form, following MATHER and JINKS (1983), which recognises that four kinds of digenic interactions are possible: between homozygotes (i_ab); heterozygotes (l_ab); and homozygote-heterozygote interactions (j_ab and j_ba). It is not supposed that all these genetic parameters will be tractable individually in human kinships, but the advantage of this general model is that it yields all the segregation patterns of classical

Table 1. Model for effects of two loci on disease liability

			Liability					
Genotype	Additive		Dominant		Epistatic			
	d_a	d_b	h_a	h_b	i_{ab}	j_{ab}	j_{ba}	l_{ab}
AABB	+1	+1	0	0	+1	0	0	0
AABb	+1	0	0	+1	0	+1	0	0
AAbb	+1	−1	0	0	−1	0	0	0
AaBB	0	+1	+1	0	0	0	+1	0
AaBb	0	0	+1	+1	0	0	0	+1
Aabb	0	−1	+1	0	0	0	−1	0
aaBB	−1	+1	0	0	−1	0	0	0
aaBb	−1	0	0	+1	0	−1	0	0
aabb	−1	−1	0	0	+1	0	0	0

epistasis, such as complementary and duplicate gene action, as specific constraints on the *d, h, i, j* and *l* (Mather and Jinks, 1982). In the two-trait case, different gene effects may be specified for each trait. For example, if the first trait is a disease both loci may contribute but if the second is some biological marker only one of the two loci may have non-zero effects. Conversely, the disease may be affected by one gene only, but the second locus may create clinical heterogeneity among affected individuals.

Given a particular model, the necessary step for data analysis is the formulation of the likelihood of observing a particular set of informative families. Typically, the likelihood is then maximized with respect to the parameters of a particular model. We consider only nuclear families in this treatment, though recognize that other kinds of family, including extended pedigrees, kinships of twins and adoptions, may be more informative for many purposes. We do not consider the problems of ascertainment and censoring because our models do not require any new principles of correction for these processes. Let γ_i be the frequency of the *i*th genotype. Let ϕ_i *and* Ψ_i be the probability that an individual will be affected by the first and second trait respectively (obtained by substitution the expression for the corresponding liability in the penetrance function). Let X_1 and X_2 be the maternal phenotypes (0,1) for the two traits in a given nuclear family, Y_1 and Y_2 be the paternal phenotypes, and U_{1i} and U_{2i} be the phenotypes of the *i*th offspring for the first and second trait. The likelihood of the family for the two variables is then:

$$l = \sum_{ij}\sum \gamma_i \gamma_j \phi_i^{X_1} \phi_j^{Y_1} \Psi_i^{X_2} \Psi_j^{Y_2} (1-\phi_i)^{(1-X_1)} (1-\phi_j)^{(1-Y_1)} (1-\Psi_i)^{(1-X_2)} (1-\Psi_j)^{(1-Y_2)}$$

$$\times \prod_{m=1}^{m=s} \sum_k \pi_{ijk} \phi_k^{U_{1m}} \Psi_k^{U_{2m}} (1-\phi_k)^{(1-U_{1m})} (1-\Psi_k)^{(1-U_{2m})} \qquad (1)$$

where the product is taken over *s* offspring (*c.f.* Elston and Stewart, 1971). The probability that parents of genotypes *i* and *j* produce a child of genotype *k* is π_{ijk}. These probabilities can be derived simply from the so-called "transmission probabilities", τ (*c.f.* Elston and Stewart, 1971) the probabilities that the genotypes AA,Aa and aa generate the A allele. In the classical Mendelian case these probabilities are 1, 0.5 and 0 respectively.

TYPES OF ENVIRONMENTAL MODEL

The above model shares with most conventional models for linkage and segregation the assumption that all the effects of the environment on determining whether or not an individual develops a disorder can be consigned to the same kinds of random and unspecifiable processes which affect whether the toss of a coin generates a head or a tail. This basic model can be modified in several distinct ways to reflect different aspects of environmental causation. Some of these have been described in recent papers by Eaves (1984) and Kendler and Eaves (1986). However, these papers do not exhaust all the major issues of environmental causation. We first consider, in words, the conceptual distinctions it is helpful to make in thinking environmental effects then show how our simple model for segregation in nuclear families may be changed to reflect each of these processes in turn.

(a) *Is the environmental factor latent or indexed?*

The distinction between "indexed" (or "measured") and "latent" environment is parallel to that made in genetic analysis between genes whose contribution to liability can be measured directly because the genotypes of individuals can be determined (as in the case of blood-group polymorphisms) and those genes whose contribution has to be inferred from their effects on segregation of the disease. Examples of environmental effects which may be indexed are salt intake, and life events such as loss of a family member or the threat of unemployment. An unidentified virus, by contrast, would appear in the model as a latent environmental effect.

(b) *Is the environment uncorrelated with genotype?*

Models for the joint biological and cultural inheritance of continuous traits have long recognized that the effects of genes and environment may be correlated because an individual's environment may depend, directly or indirectly, on his genotype or the genotypes of his relatives. Such "genotype–environment correlations" may be divided formally into three types. The first class ("genotype–environment auto-correlations") arise because an individual's environment depends on his own genotype, for example, in creating adverse life events. The second type of genotype–environment correlation may arise as a result of "vertical cultural inheritance" (Cavalli-Sforza and Feldman, 1981) when the environment of an offspring depends on the genotype or phenotype of parents. These effects may include, but are not restricted to, maternal effects. In such cases, the parents who provide the genotype of a child also generate the environment in which development takes place. An example of this type of correlation might be the correlated effects of a depressed or anxious parent providing a child with increased genetic liability for anxiety or depression as well as an adverse psychological environment. The third kind of genotype–environment correlation important for our theoretical purposes is that which arises because of "horizontal" transmission between siblings. Such effects might prove important in "infectious" models of transmission in which a person's chance of disease is a function of the disease status of family members. Liability may also depend on the particular configuration of genotypes in the family because some individuals may be genetically more liable to infection than others.

(c) *Is there genotype–environment interaction?*

An *additive* model for the effects of genes and environment on liability must be distinguished from an *interactive* model. In the additive model, the effect on liability of a particular change in the environment (from high to low sodium intake, for example) is assumed to be the same for all genotypes. Thus, all genotypes are equally sensitive. "Genotype × environment interaction" (G×E) arises when different genotypes display different sensitivity to the same environmental change. G×E interaction would arise if a genotype at high risk for depression is especially sensitive to adverse life events, or if a child with an increased genetic risk for anti-social personality is also less sensitive to the modifying effects of the normal social environment. EAVES (1984) developed a theoretical model for such interactions and KENDLER and EAVES (1986) showed they may be detectable by analysis of kinship data conditional on exposure to specified protective and predisposing environments.

OUTLINE OF MATHEMATICAL FORMULATION OF ENVIRONMENTAL MODELS

(a) *Unmeasured environment independent of genotype*

The first and simplest way the conventional genetic model for segregation may fail is that the genotype alone is inadequate to specify variation in liability because some major (unidentified) environmental risk factor is creating heterogeneity in the population. In this case we can employ some version of the strategy proposed by LALOUEL *et al.* (1983) in order to test for the presence of segregating non-Mendelian factor in the "unified mixed model" and allow the transmission probabilities (τ_i) at the second "locus" to depart from their Mendelian values. If the frequency of the environmental factor is not to change between generations it is necessary that the $0 < \tau_i < 1$ also satisfy the constraint $p - \gamma\tau' = 0$, where P is the frequency of the risk increasing factor. In addition to the Mendelian values of the transmission probabilities, the constraint is satisfied for $\tau \cong (p, p, p)$, in which case there would be no evidence of intergenerational transmission (equivalent to a purely "sporadic" environmental factor). Other constraints may be imposed to specify particular models of vertical inheritance of an independent non-genetic factor.

(b) *The measured environment and genotype–environment interaction*

If a particular environmental factor can be measured, the regression of liability on the values of the environmental variable can be incorporated in the model. Thus, the liability of the *i*th genotype, having some (independent) environmental index value, *E,* is a function of the average effect of the *i*th genotype on liability, g_i, and the environmental covariate:

$$L_i = g_i + \beta_i E.$$

If the regression coefficient, β_i, is the same for all genotypes (EAVES, 1984; KENDLER and EAVES, 1986) the effects of genes and environment are additive on the scale of liability, since sensitivity to the environment is not under genetic control. However, each genotype may have a different sensitivity to the environment so that a different regression coefficient needs to be estimated for each genotype. Eaves shows how the sensitivity parameters, β_i, can also be expressed in terms of additive and heterozygous effects on sensitivity to the environment (c.f. MATHER and JINKS, 1982). Control of sensitivity to the environment may be mediated by quite different genes from those which affect average liability in just the same way that there may be genes which specifically affect the age of onset of a disorder. Detection of a second locus in a two locus system (i.e. "genetic heterogeneity") is made easier when one of the two genes affects a different aspect of the phenotype. This arises when the first locus controls average liability and the second controls sensitivity to the environment.

(c) *The effects of genes on life events*

The effects of "environment" on liability to psychiatric disease are not necessarily expected to be independent of genotype if premorbid expression of a high risk genotype leads to disruption of life style which, in turn, increases risk for the disease still further. The two-locus two-trait model permits us to explore this possibility by coding the hypothesized environmental variable as one of the traits. The first phenotype may be the presence or absence of a life event and the second may be the presence or absence of the disease itself. The two phenotypes may be controlled by the same genes or different genes, one may be controlled by Mendelian factors and the other by non-Mendelian factors, etc. We may distinguish two simple possibilities in theory which are practically important. The first case, which we may call "classical pleiotropy", allows both phenotypes to be a function of genotype at one or more loci, but does not allow for any direct effect of the "life event" on disease liability. In this case, then likelihood of a nuclear family is identical to that presented for the conventional two-locus model above. The only cause of correlation between life event and disease is thus assumed to be the underlying effect of the genotype. There is no direct effect of the life event on liability to the disease itself. The case in which the *phenotype* for the life event has a direct effect on the liability to the disease is represented by changing the definition of Ψ_i to allow for the regression of liability to disease (X_2) to depend on *phenotype* for the life event (X_1) thus:

$$(\Psi_i \mid X_1) = \frac{1}{1+e} - (g_i + \beta_i X_1).$$

As in the earlier case of the dependence of liability on the measured environment, the regression of liability on life event phenotype may be the same for all genotypes, in which case the β_i are constant, or they may depend on genotype if there is assumed to be genotype × environment interaction. This model is really an extension of the previous model for the effect of covariates on liability. In this model, we allow the covariate also to have a genetic component and, since the genetic component of the covariate is inferred from the pattern of segregation, we compute the joint likelihood of the disease *and* the covariate. In the former case, the likelihood of the disease was computed conditional upon the values of the covariate.

(d) *The environment depends on relatives*

In the previous case, the environment was considered to be an aspect of the individual's phenotype which may be influenced by his own genotype. However, the salient features of the environment may be the phenotypes of other people. If the environment is created by the phenotype of a relative, then the genetic correlation between relatives will generate a correlation between the genotype of an individual and the environment provided by this relative. Under these circumstances, the transmission of a disease phenotype may appear to be non-Mendelian, even though a major gene is actually segregating. The model for liability of an offspring can now be expressed as a function of its own genotype and the phenotypes of the parents, X and Y, thus:

$$L_i = g_i + \beta_i X + \beta_i' Y.$$

The β_i and β_i' may differ between mothers and fathers if, for example, there are maternal effects. Also, if there is genotype × environment interaction, the regression of offspring liability on parental disease phenotype may depend on the genotype of the offspring. In theory, it would also be possible to allow the β's to depend on the genotypes of the parents but our experience with complex models for continuous traits suggests that such subtle effects may be beyond resolution.

The original expression for the likelihood of a nuclear family has to be modified when there is vertical cultural inheritance because the probability that a child will develop the disease is conditional on his own genotype and the phenotypes of his parents, X and Y. We denote this probability for the kth genotype by ϕ_{kXY}. Furthermore, when the phenotypes of the *grandparents* are unknown, the probability that a parent of genotype k will have the disease is the weighted average of the corresponding ϕ_{kXY}'s over all possible combinations of grandparental phenotypes. This average penetrance is denoted by ξ_k. The modified likelihood for a nuclear family is thus:

$$l = \sum_{ij}\sum \gamma_i \gamma_j \xi_i^X \xi_j^Y (1-\xi_i)^{1-X} (1-\xi_j)^{1-Y} \prod_{m=1}^{m=s} \sum_k \pi_{ijk} \phi_{kXY}^{U_m} (1 - \phi_{kXY})^{(1-U_m)}.$$

Now, given the genotype frequencies, values for the ξ_k in the parental generation, parameters for the genotypic effects and the environmental impact of parents on children, the average penetrance of each genotype in the offspring generation, ξ_k^*, may be obtained:

$$\xi_k^* = \sum_{ij}\sum \gamma_i \gamma_j \pi_{ijk} \alpha_{ijk} = f(\xi_k)$$

where

$$\alpha_{ijk} = \sum_{X=0}^{X=1} \sum_{Y=0}^{Y=1} \xi_i^X \xi_j^Y (1 - \xi_i)^{(1-X)} (1 - \xi_j)^{(1-Y)} \phi_{kXY}.$$

In the absence of selection, and intergenerational change in the intensity of vertical cultural inheritance, we expect the penetrances in the two generations to approach equilibrium values $\hat{\boldsymbol{\xi}} - f(\hat{\boldsymbol{\xi}}) = 0$ must be satisfied. Such constraints are tedious to solve algebraically, but they may be imposed numerically by use of Lagrange multipliers when the model is fitted to actual data. These conditions for the single gene model are analogous to the equilibrium constraints on the correlation between genetic and environmental effects implied by path models for biological and cultural inheritance in human populations (e.g. RICE *et al.*, 1978).

The case of "vertical cultural inheritance" in the presence of genetic effects is one way the phenotype of one person may influence the risk of another. A final example is that of "sibling interaction", exemplified by the case of infectious disease. The case of infection in the presence of genetic variation in sensitivity to a pathogen is important theoretically, but turns out that the likelihood even for a nuclear family, requires the computation of a very large number of terms when there are genetic differences in sensitivity. The simple case in which everyone in a family of size s is equally sensitive to infection, however, is tractable. We then assume that the individuals in a family initially uninfected each have a probability, P, of becoming infected from "outside" the family. The chance of initial infection of i members of the family from "outside" is thus:

$$h_{si} = \binom{i}{s} p^i (1-p)^{s-i}.$$

Once the disease is established in the family, the greater contact between family members leads to an increased risk of infection, q for each hitherto uninfected member. The likelihood of obtaining a total of $r>0$ affected individuals in a family of size s is thus:

$$t_{sr} = \sum_{i=1}^{i=r} h_{si} \binom{r-i}{s-i} q^{(r-i)} (1-q)^{(s-r)} \qquad r>0$$

An obvious consequence of this mechanism of transmission is the dependence of risk on family size, since an individual in a large family has a higher chance of being exposed to someone who has acquired the disease from outside.

DISCUSSION

The search for major gene effects on disease has been pursued in isolation from epidemiological studies of environmental variables. Genetic epidemiology has a wide range of models for the effects of environment in multifactorial systems but these have not been extended consistently to the discontinuous case in the presence of genetic differences. This paper presents a number of different ways in which the environment may add to, and interact with, the genetic component of disease liability, and outlines ways in which such effects may be specified within a framework of more conventional genetic analysis. It remains to be seen how far such effects are tractable with actual data, and whether it is necessary to take account of them in the genetic analysis of psychiatric disorders.

Acknowledgement—Supported by grants GM30250, HL28922, GM32732, MH40828 and AA06781 from the National Institutes of Health.

REFERENCES

Cavalli-Sforza, L. L. and Feldman, M. W. (1973) Cultural versus biolcgical inheritance: phenotypic transmission from parents to children. (A theory of the effect of parental phenotypes on children's phenotypes.) *Am. J. hum. Genet.* **25**, 618–637.

Cavalli-Sforza, L. L. and Feldman, M. W. (1981) *Cultural Transmission and Evolution: A Quantitative Approach.* Princeton University, Princeton, N.J.

Cavalli-Sforza, L. L., Feldman, M. W., Chen, K. H. and Dornbusch, S. M. (1982) Theory and observation in cultural transmission. *Science* **218**, 19–27.

Eaves, L. J. (1976a) The effect of cultural transmission on continuous variation. *Heredity* **37**, 41–57.

Eaves, L. J. (1976b) A model for sibling effects in man. *Heredity* **36**, 205–214.

Eaves, L. J. (1984) The resolution of genotype × environment interaction in segregation analysis of nuclear families. *Genet. Epidemiol.* **1**, 215–228.

Elston, R. C. and Stewart, J. (1971) A general model for the genetic analysis of pedigree data. *Hum. Hered.* **21**, 523–542.

Fisher, R. A. (1918) On the correlation between relatives on the supposition of Mendelian inheritance. *Trans. Roy. Soc. Edin.* **52**, 399–433.

HEATH, A. C., KENDLER, K. S., EAVES, L. J. and MARKELL, D. (1985) The resolution of cultural and biological inheritance: informativeness of different relationships. *Behav. Genet.* **15**, 39–466.

KENDLER, K. S. and EAVES, L. J. (1976) Models for the joint effect of genotype and environment on liability to psychiatric illness. *Archs gen. Psychiat.* **143**, 279–289.

LALOUEL, J. M., RAO, D. C., MORTON, N. E. and ELSTON, R. C. (1983) A unified model for complex segregation analysis. *Am. J. hum. Genet.* **35**, 816–826.

MATHER, K. and JINKS, J. L. (1982) *Biometrical Genetics: The Study of Continuous Variation.* Chapman and Hall, London.

MORTON, N. E. (1974) Analysis of family resemblance I. Introduction. *Am. J. hum. Genet.* **36**, 318–330.

RAO, D. C., MORTON, N. E. and YEE, S. (1976) Resolution of biological and cultural inheritance by path analysis. *Am. J. hum. Genet.* **28**, 228–242.

RICE, J., CLONINGER, C. R. and REICH, T. (1978) Multifactorial inheritance with cultural transmission and assortative mating. I. Description and basic properties of the unitary models. *Am. J. hum. Genet.* **30**, 618–643.