Testing familial aggregation

Rosendaal, F.R.

BIOMETRICS 51, 1292–1301
December 1995

Testing Familial Aggregation

Jeanine J. Houwing-Duistermaat,[1] Bert H. F. Derkx,[2] Frits R. Rosendaal,[3] and Hans C. van Houwelingen[4]

[1]Department of Medical Statistics, Leiden University,
P.O. Box 9604, 2300 RC, Leiden;
[2]Department of Pediatrics, Children's Academical Medical Centre, Amsterdam;
[3]Department of Clinical Epidemiology, and Leiden University, Leiden; and
[4]Department of Medical Statistics, Leiden University,
Leiden, the Netherlands

SUMMARY

Likelihood calculation for pedigrees is complicated and often time-consuming. Testing correlation structures due to familial aggregation is therefore a preliminary procedure. A score statistic is given to check correlations between relatives of randomly chosen pedigrees. This statistic can be used for quantitative and dichotomous data. For both data types, the distribution of the statistic under the null hypothesis is derived. To demonstrate the performance of the statistic, results of simulations under various models are given. Finally, the test is applied to data on a continuous blood factor.

1. Introduction

The general genetic models of Elston and Steward (1971) are often used to model familial data. Since 1971, many other genetic models have been developed and introduced. Thompson (1986a) gives an overview of the range of genetic models available. A useful class of genetic models is the regressive models introduced by Bonney (1984, 1986). These models specify a major gene effect and a residual effect that represents the combined effects of environmental, familial, and polygenic factors. For both effects, the distribution is specified by conditioning each response on those of the preceding relatives. However, since these genetic models have many model parameters and the maximum likelihood estimates are highly interdependent, statistical conclusions are often difficult to make (Thompson, 1986b). Hence, these genetic models cannot be applied to small data sets concerning a trait for which the genetic effect is rather unclear.

For simplicity, we consider a set of pedigrees from a population of individuals who mate completely at random and where no natural selection exists and no mutations occur; hence the genotype probabilities for the founders of a pedigree are the Hardy-Weinberg equilibrium probabilities, and the conditional genotype probabilities for the nonfounders follow via the Mendelian laws from the genotypes of the parents. A discussion about these assumptions can be found in Thompson (1986a, chapter 1). Now, the likelihood is the weighted sum over all genotype combinations G:

$$L(\theta|Y) = \sum_{\substack{\text{genotype} \\ \text{combinations}}} P(Y|G,\ \theta)P(G|\theta),$$

where θ is a vector of model parameters and $P(G|\theta)$ can be computed using Hardy-Weinberg and the Mendelian laws.

In studies on familial aggregation of a certain trait, investigation for a possible correlation between relatives is an essential preliminary procedure. The genetic distance between individuals should determine the correlation between two individuals of the same pedigree. The aim of this paper is to develop a test for correlation structures between relatives that is less specified than the conventional genetic models. This test is a scaled version of the goodness-of-fit test based on models for random effects derived by le Cessie and van Houwelingen (1995).

Key words: Familial aggregation; Random effects; Score test.

In Section 2, a random effect model is introduced. In Section 3, the statistic to test the hypothesis of no correlation is given. In Section 4, the distribution of the statistic under the null hypothesis is derived. To investigate the performance of the test, simulations were carried out. The results of these simulations are given in Section 5. In Section 6, the test is applied to a study of familial aggregation of a blood variable. Section 7 discusses the implications of our findings.

2. The Model

Let $\mathbf{Y}$ be a response vector of members of randomly chosen pedigrees and Y_j the response variable of member j. The response variables of members of the same pedigree may be correlated because of genetic effects. These genetic effects are random effects and the following family of random effect models is proposed to model the data:

$$E(Y_j|u_j) = h^{-1}(\mu + u_j), \tag{2.1}$$

where h is a link function (McCullagh and Nelder, 1989) and the values of uj are correlated zero-mean genetic (random) effects with covariance matrix $\tau^2\mathbf{R}$ (the correlation matrix $\mathbf{R}$ will be specified below). The genetic effects u_j induce a correlation between the response variables Y_j and it can be assumed that Y_j are conditionally independent given the genetic effects u_j. The variance of Y_j given u_j represents the unpredictability of the response variable given the genetic effect.

For $\tau^2 = 0$, the response variables Y_j are independent and identically distributed. Testing whether the genetic effects are present in the data is equivalent to testing the hypothesis $\tau^2 = 0$ versus $\tau^2 > 0$. In the following section a test statistic is given.

Now consider a Mendelian dominant autosomal gene that influences a quantitative trait. Individuals who have the allele A have a larger mean response than individuals who do not have the allele A. Define G_j to be the genotype of person j. By taking the identity as link function, the model becomes

$$E[Y_j|G_j] = \mu + \beta([G_j \in \{AA, Aa\}] - P(G_j \in \{AA, Aa\})), \tag{2.2}$$

where $[.]$ is the indicator function and $\beta([G_j \in \{AA, Aa\}] - P(G_j \in \{AA, Aa\}))$ is the genetic effect u_j. Let p be the allele frequency of A, then $P(G_j \in \{AA, Aa\}) = p^2 + 2p(1-p) = p(2-p)$ and the variance of the genetic effect $\tau^2 = \beta^2 p(2-p)(1-p)^2$. Elandt-Johnson (1971, pp. 138–149) derives the correlation between individuals of a sibship as $(4-3p)/(8-4p)$ and between a parent and a sib as $(1-p)/(2-p)$. Therefore, as p approaches zero, the correlation of the genetic effects of a pedigree tends to the following natural correlation structure $\mathbf{R}$:

- individuals within a sibship have correlation 1/2.
- parent–offspring have correlation 1/2.
- grandchild–grandparent have correlation 1/4.
- aunt/uncle–niece/nephew have correlation 1/4, etc.

(2.3)

For a Mendelian autosomal gene with incomplete penetrance, the following model can be used:

$$E[Y_j|G_j] = \mu + \beta([G_j = AA] + b[G_j = Aa] - P(G_j = AA) - bP(G_j = Aa)), \tag{2.4}$$

where $0 < b < 1$. Here the genetic effect u_j is equal to

$$\beta([G_j = AA] + b[G_j = Aa] - P(G_j = AA) - bP(G_j = Aa)).$$

For $b = 0.5$, $P(G_j = AA) - bP(G_j = Aa) = p$, $\tau^2 = 0.5\beta^2 p(1-p)$, and for every allele frequency the correlation matrix is equal to correlation matrix $\mathbf{R}$ (2.3).

If the response variable is dichotomous, the link function h in model (2.1) is usually the logit function of $E[Y_j]$, and regression model (2.2) corresponds to a logistic regression model for a locus with dominance:

$$\text{logit } E[Y_j|G_j] = \mu + \beta([G_j \in \{AA, Aa\}] - P(G_j \in \{AA, Aa\})). \tag{2.5}$$

This model reduces to two probabilities:

$$P(Y_j = 1|G_j \in \{AA, Aa\})$$

$$P(Y_j = 1|G_j = aa). \tag{2.6}$$

In the following sections, we use the matrix $\mathbf{R}$ (2.3) as the working correlation matrix of the genetic effects:

$$\mathrm{COV}(u) = \tau^2\mathbf{R}.$$

As will be discussed in Sections 5 and 7, for testing dependency between relatives, it is not necessary that the working correlation structure agree with the correlation structure of the genetic effects completely.

3. The Statistic Q

Suppose the data are obtained from k pedigrees containing n persons. Le Cessie and van Houwelingen (1995) show that the score test for testing the hypothesis of $\tau^2 = 0$ is based on the quadratic form $\Sigma_{i=1}^{n}(Y_i - \mu\mathbf{1}_i)'\mathbf{R}_i(Y_i - \mu\mathbf{1}_i)$, where Y_i is the response vector of pedigree i, $\mathbf{1}_i$ is a vector of ones of same length as Y_i, μ is the mean of Y_{ij}, and $\mathbf{R}_i$ is the correlation matrix of pedigree i. We will use the following version of this statistic:

$$Q = \sum_{i=1}^{k} \frac{(Y_i - \mu\mathbf{1}_i)'\mathbf{R}_i(Y_i - \mu\mathbf{1}_i)}{\sigma^2},$$

where σ^2 is the variance of Y_{ij} under the null hypothesis. By defining

$$\mathbf{R} = \begin{pmatrix} R_1 & & \\ & \ddots & \\ & & R_k \end{pmatrix},$$

and $Y = (Y_1', \ldots, Y_k')$, Q can be written

$$Q = \frac{(Y - \mu\mathbf{1})'\mathbf{R}(Y - \mu\mathbf{1})}{\sigma^2},$$

where $\mathbf{1}$ is a vector of ones of length n. This last formula of Q will be used in the following sections.

To get an impression of the statistic Q, let Y_j be a dichotomous response variable of person j with known mean μ and variance $\sigma^2 = \mu(1 - \mu)$ and write

$$Q = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{(Y_i - \mu)\mathbf{R}_{ij}(Y_j - \mu)}{\sigma^2},$$

where $\mathbf{R}_{ij}$ is the natural correlation (2.3) between person i and person j of the same pedigree and $\mathbf{R}_{ij}$ is zero if i and j are members of different pedigrees. Now, for the pair of individuals i, j,

$$(Y_i - \mu)R_{ij}(Y_j - \mu) \begin{cases} = (1 - \mu)^2\mathbf{R}_{ij} & \text{if } Y_i = Y_j = 1 \\ = \mu^2\mathbf{R}_{ij} & \text{if } Y_i = Y_j = 0 \\ = -\mu(1 - \mu)\mathbf{R}_{ij} & \text{if } Y_i \neq Y_j \end{cases}$$

Q tends to be large, if $Y_i = Y_j$ for those individuals for which $\mathbf{R}_{ij}$ is large. Hence, Q measures familial aggregation.

Note that for $\mu = 1/2$:

$$Q = \sum_{\substack{\text{concordant} \\ \text{pairs}}} \mathbf{R}_{ij} - \sum_{\substack{\text{discordant} \\ \text{pairs}}} \mathbf{R}_{ij}$$

The value of Q is determined mainly by sibship and parent–child relations, but the other relationships also contribute to the value of Q.

4. The Distribution of Q under the Null Hypothesis

If Y follows a normal distribution then $E(Q) = \mathrm{trace}(\mathbf{R})$ and $\mathrm{VAR}(Q) = 2\mathrm{trace}(\mathbf{R}^2)$ under the null hypothesis of no correlation (Kendall and Stuart, 1963), and if Y follows a binomial distribution ($\sigma^2 = \mu(1 - \mu)$), $E(Q) = \mathrm{trace}(\mathbf{R})$ and

$$\mathrm{VAR}(Q) = \frac{1}{\sigma^4}\,\mathrm{VAR}((Y-\mu)'\mathbf{R}(Y-\mu))$$

$$= \frac{1}{\sigma^4}\left(\sum_{i=1}^{n} \mathbf{R}_{ii}^2\mu(1-\mu)(1-6\mu+6\mu^2) + 2\mu^2(1-\mu)^2\mathrm{trace}(\mathbf{R}^2)\right) \tag{4.1}$$

$$= n\,\frac{1-6\mu+6\mu^2}{\mu(1-\mu)} + 2\mathrm{trace}(\mathbf{R}^2)$$

(le Cessie and van Houwelingen, 1995).

The distribution of Q can be approximated by a χ^2 distribution with scale parameter $c = \mathrm{VAR}(Q)/2E(Q)$ and $\nu = 2E^2(Q)/\mathrm{VAR}(Q)$ degrees of freedom. By means of simulations le Cessie and van Houwelingen (1995) show that the performance of the scaled χ^2 is better than the straightforward approximation by a normal distribution.

The parameters μ and σ^2 are often unknown and have to be estimated from the data. Since the estimated mean will be closer to the observed data, it leaves less variance to the residuals. Moreover, when σ^2 is unknown, Q is a quotient of two quadratic forms, the numerator $N = (Y - \hat{\mu})'\mathbf{R}(Y - \hat{\mu})$ and the denominator $D = [1/(n-1)](Y - \hat{\mu})'(Y - \hat{\mu})$. These quadratic forms are positively correlated, and neglecting this correlation gives an overestimation of the variance of Q. Therefore, it is necessary to adjust for the estimation of these parameters.

Let $H = (1/n)\mathbf{1}\mathbf{1}'$, with $\mathbf{1}$ the n-dimensional vector $(1, \ldots, 1)'$, then H is the projection of Y on $\mathbf{1}$. Since $Y - \hat{\mu} = (I - H)(Y - \mu)$, the matrix $\mathbf{R}$ has to be replaced by $\tilde{\mathbf{R}} = (I - H)'\mathbf{R}(I - H)$ when we compute the expectation and variance of Q (see le Cessie and van Houwelingen, 1995). Now, we can write Q as follows:

$$Q = \frac{N}{D} = (n-1)\,\frac{(Y-\hat{\mu})'\mathbf{R}(Y-\hat{\mu})}{(Y-\hat{\mu})'(Y-\hat{\mu})} = (n-1)\,\frac{(Y-\mu)'\tilde{\mathbf{R}}(Y-\mu)}{(Y-\mu)'(I-H)(Y-\mu)}.$$

If Y_i are independent and normally distributed, the mean and variance of Q under the null hypothesis can be computed taking the dependency of the numerator N and the denominator D into account (see Appendix):

$$E(Q) = \mathrm{trace}(\tilde{\mathbf{R}})$$

and

$$\mathrm{VAR}(Q) = \frac{2}{n+1}\left((n-1)\mathrm{trace}\,(\tilde{\mathbf{R}}^2) - \mathrm{trace}^2(\tilde{\mathbf{R}})\right).$$

Indeed, the variance of Q is smaller than the variance computed without correction for the estimation of the variance of Y. Observe that the $E(Q)$ and $\mathrm{VAR}(Q)$ are constants, in contrast with the version of le Cessie and van Houwelingen (1995). Observe also, that $E(Q) = E(N)/E(D)$ and $\mathrm{VAR}(Q)$ agree with the following approximation except for a factor $(n-1)/(n+1)$:

$$\mathrm{VAR}\left(\frac{N}{D}\right) \approx \frac{\mathrm{VAR}(N)}{E^2(D)} + \frac{E^2(N)}{E^4(D)}\,\mathrm{VAR}(D) - 2\,\frac{E(N)}{E^3(D)}\,\mathrm{COV}(N, D). \tag{4.2}$$

This approximation follows by a first-order Taylor expansion of $Q = N/D$ around $(E(N), E(D))$ and by the fact that $E(Q) = E(N)/E(D)$.

For Y binomially distributed, the conditional expectation of Q given $\Sigma\, Y_i = s$ under the null hypothesis can be computed:

$$E(Q|\textstyle\sum Y_i = s) = \mathrm{trace}\,(\tilde{\mathbf{R}}).$$

Since the conditional expectation is independent of s, it is equal to the expectation of Q and it appears that also for Y binomially distributed $E(Q) = E(N)/E(D)$. However, the conditional variance of Q given $\Sigma\, Y_i = s$ depends on s. (Note that the conditional variance of Q given $\Sigma\, Y_i = s$ can be derived, since we can calculate $E(((Y-\mu)'\tilde{\mathbf{R}}(Y-\mu))^2)$ and the statistic $\Sigma\, Y_i$ is complete.) We preferred to use approximation (4.2) to compute the variance of Q under the null hypothesis, motivated by the facts that, for Y distributed normally and binomially, $E(Q)$ is equal to $E(N)/E(D)$ and that for Y distributed normally this approximation agrees with the variance of Q

except for a factor $(n - 1)/(n + 1)$. Now, VAR(N), VAR(D), and COV(N, D) can be computed using

$$\mathrm{COV}((Y-\mu)'A(Y-\mu),\ (Y-\mu)'B(Y-\mu)) = \sum_{i=1}^{n} A_{ii}B_{ii}\mu(1-\mu)(1-6\mu+6\mu^2) + 2\mu^2(1-\mu)^2\mathrm{trace}(A\cdot B) \tag{4.3}$$

and it follows from approximation (4.2) that the variance of Q under the null hypothesis can be approximated by

$$\mathrm{VAR}(Q) \approx K\left(\sum_{i=1}^{n}\tilde{\mathbf{R}}_{ii}^2 - \frac{\mathrm{trace}^2(\tilde{\mathbf{R}})}{n}\right) + \frac{2}{n-1}\left((n-1)\mathrm{trace}(\tilde{\mathbf{R}}^2) - \mathrm{trace}^2(\tilde{\mathbf{R}})\right),$$

where

$$K = \frac{1-6\mu+6\mu^2}{\mu(1-\mu)}.$$

From simulations of dichotomous data, it appears that this approximation performs well. The results of these simulations will be given in Section 5.

5. Performance of the Test Statistic

To study the performance of the test, 7 pedigrees (with sizes 10, 10, 5, 8, 5, 3, and 2) were simulated under different genetic models. This data set of 7 pedigrees has the same structure as the data of the example of Section 6. First the performance of the test for continuous data is studied. To check the significance level of the test under the null hypothesis of no familial aggregation, a simulation of 10,000 samples of 43 standard normal distributed response variables is performed. The formulae of Section 4 give $E(Q) = 39.96$ and $\mathrm{VAR}(Q) = 65.62$, giving a scale factor of $c = .82$ and $\nu = 48.67$ degrees of freedom. A nominal level of $\alpha = 0.05$ corresponds to a cut-off point of $65.95 \cdot .82 = 54.05$. The test rejects the null hypothesis of no correlation in 5.5% of the cases. The estimated mean and variance are 39.92 and 65.59, respectively. In Figure 1, the cumulative distribution function of the simulated Q and the cumulative scaled χ^2 distribution with $\nu = 48.67$ degrees of freedom are given. It is clear that the scaled χ^2 distribution is a good approximation of the distribution of Q under the null hypothesis of no correlation.

To study the power of the test, 10,000 samples of the genotypes of the data set (43 individuals) for the allele frequencies $p = .01$ and .1 are generated. For each group of 10,000 simulated gene patterns, response variables under 10 different genetic models are simulated. The models are

$$E(Y_j|G_j) = \mu + \beta([G_j = AA] + b[G_j = Aa] - p^2 - 2bp(1-p)),$$

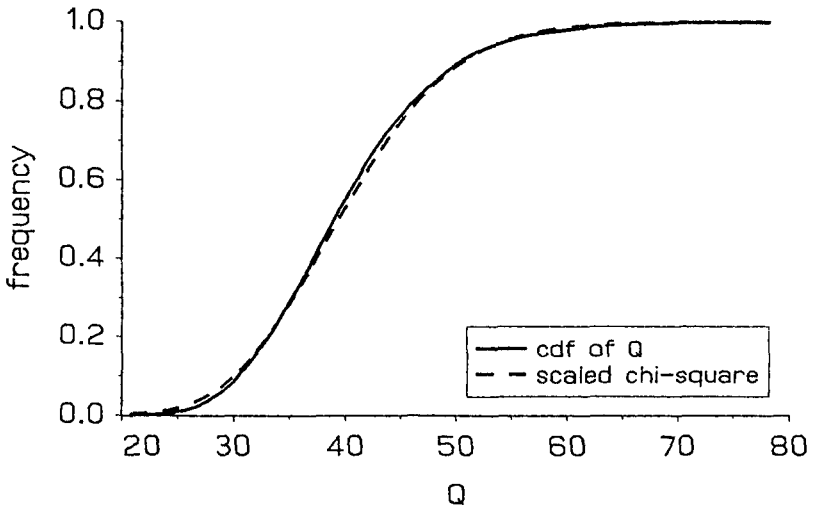

Figure 1. The cumulative distribution function of the under the null hypothesia simulated statistic Q and the cumulative scaled χ^2 distribution with $\nu = 18.67$ degrees of freedom.

Table 1

Ten thousand simulations under a dominant model with an allele frequency of .01 for continuous response variables: the variance of the random effect (τ^2), the power, and the expectation and variance of Q estimated from the simulated data and the computed expectation and variance of Q

			Estimated		Computed	
β	τ^2	Power	$\hat{E}(Q)$	$\hat{\text{VAR}}(Q)$	$\tilde{E}(Q)$	$\tilde{\text{VAR}}(Q)$
1	.02	.066	40.58	69.65	40.62	72.86
2	.08	.100	41.80	88.60	42.46	84.82
3	.18	.134	43.07	120.01	45.13	103.14
4	.31	.163	44.13	153.28	48.23	125.26
5	.49	.181	44.93	182.07	51.41	148.31

where b is equal to 1 (dominant model) or 0.5 (incomplete dominant model) and β varies from 1 to 5 (recall $\sigma^2 = 1$). The expectations and variances of Q under the different models are estimated from the simulated data. The results for the dominant model are given in Tables 1 and 2 and for the incomplete dominant model in Tables 3 and 4. It is clear that the power is larger for a stronger genetic effect (large β and/or p). The power is small for a weak genetic effect, because of the smallness of the data set.

The equality $E(Q) = E(N)/E(D)$ does not hold under alternative models, but for small τ^2, the expectation of Q can still be approximated by $E(N)/E(D)$. When the working correlation (2.3) is used as correlation matrix of the genetic effects, the expectation is

$$\tilde{E}(Q) \approx (n-1)\frac{\text{trace}(\tilde{\mathbf{R}}) + \frac{\tau^2}{\sigma^2}\text{trace}(\tilde{\mathbf{R}}^2)}{n - 1 + \frac{\tau^2}{\sigma^2}\text{trace}(\tilde{\mathbf{R}})}, \tag{5.1}$$

and formula (4.2) can be used to approximate the variance of Q, where $\text{VAR}(N)$, $\text{VAR}(D)$, and $\text{COV}(N, D)$ can be computed using

$$\tilde{\text{COV}}((Y-\mu)'A(Y-\mu), (Y-\mu)'B(Y-\mu)) = 2\text{trace}((\sigma^2 I + \tau^2\mathbf{R})A(\sigma^2 I + \tau^2\mathbf{R})B). \tag{5.2}$$

Table 2

Ten thousand simulations under a dominant model with an allele frequency of .1 for continuous response variables: the variance of the random effect (τ^2), the power, and the expectation and variance of Q estimated from the simulated data and the computed expectation and variance of Q

			Estimated		Computed	
β	τ^2	Power	$\hat{E}(Q)$	$\hat{\text{VAR}}(Q)$	$\tilde{E}(Q)$	$\tilde{\text{VAR}}(Q)$
1	.15	.137	43.87	88.74	44.57	99.26
2	.62	.349	50.97	135.90	53.30	161.99
3	1.39	.536	56.70	179.72	60.50	211.50
4	2.46	.644	60.51	211.10	65.28	240.70
5	3.85	.705	62.97	231.97	68.33	257.33

Table 3

Ten thousand simulations under an incomplete dominant model with an allele frequency of .01 for continuous response variables: the variance of the random effect (τ^2), the power, and the expectation and variance of Q estimated from the simulated data and the computed expectation and variance of Q

			Estimated		Computed	
β	τ^2	Power	$\hat{E}(Q)$	$\hat{\text{VAR}}(Q)$	$\tilde{E}(Q)$	$\tilde{\text{VAR}}(Q)$
1	.01	.058	40.18	66.20	40.30	70.84
2	.04	.066	40.60	69.80	41.27	77.04
3	.09	.082	41.18	77.49	42.78	87.01
4	.16	.101	41.83	89.44	44.69	100.08
5	.25	.119	42.49	104.57	46.84	115.30

Table 4

Ten thousand simulations under an incomplete dominant model with an allele frequency of .1 for continuous response variables: the variance of the random effect (τ^2), the power, and the expectation and variance of Q estimated from the simulated data and the computed expectation and variance of Q

			Estimated		Computed	
β	τ^2	Power	$\hat{E}(Q)$	$\hat{\text{VAR}}(Q)$	$\tilde{E}(Q)$	$\tilde{\text{VAR}}(Q)$
1	.09	.081	41.34	73.62	42.81	87.18
2	.36	.159	44.60	95.91	49.18	132.11
3	.81	.272	48.51	126.20	55.68	178.90
4	1.44	.389	52.19	156.35	60.84	213.72
5	2.25	.488	55.30	181.78	64.58	236.68

The computed expectations and variance under the alternative models are also given in Tables 1, 2, 3, and 4. From these tables it can be concluded that only for weak genetic effects the computed expectations and variances agree with the estimated expectations and variances.

Estimates of τ^2, the variance of the genetic effects, are obtained by a first-order approximation (based on the score statistic for $\tau^2 = 0$):

$$\left(\frac{\tau^2}{\sigma^2}\right) \approx 2 \cdot \frac{Q - E(Q)}{\text{VAR}(Q)} \tag{5.3}$$

These estimates and the true values are given in Table 5. Only for weak genetic effects are the estimates reasonable; for strong genetic effects the first-order estimates are too small.

Table 5

The variance and the estimated variance of the genetic effects for an allele frequency of .1 and continuous response variables

	$b = 1$		$b = 0.5$	
β	τ^2	$\hat{\tau}^2$	τ^2	$\hat{\tau}^2$
1	0.15	0.12	0.05	0.04
2	0.62	0.34	0.18	0.14
3	1.39	0.51	0.41	0.26
4	2.46	0.63	0.72	0.37
5	3.85	0.70	1.125	0.47

For various marginal probabilities of getting a dichotomous trait ($= \Sigma\, P(Y = 1|G)P(G)$), 10,000 samples of 43 dichotomous response variables are simulated under model (2.6). The marginal probabilities are .1, .2, .4, and .5. To study the power, dominant models are considered. Gene patterns are simulated under allele frequencies .01 and .1 of A and response variables are simulated for two different conditional probabilities of getting the disease given the genotype is AA or Aa: .7 and .99, whereas the marginal probability of getting the disease is kept on .1, .2, .4, or .5. If the response vector is zero (giving a zero denominator of Q), the null hypothesis of no correlation is not rejected. The results are given in Table 6. It appears that the four actual levels (4.3%, 5.4%, 5.3%, and 5.4%) agree with the nominal level of 5% and that the power is reasonable for an allele frequency of 0.1 and a large difference between the conditional probabilities of getting the disease given the genotype is AA or Aa and getting the disease given the genotype is aa. The power is small for an allele frequency of .01.

For weak genetic effects, τ^2 can be estimated using the score statistic, and the expectation and variance of Q under an alternative model can be computed using formulae similar to those for the continuous response variables (5.1) and (5.2).

To compare the correlation of the genetic effects under a dominant model with the natural working correlation matrix **R** (2.3), these correlations are calculated using the formulae of Section 2 for two members of a sibship and for a child and a parent (Table 7). Since the true correlation matrix hardly differs from our working correlation matrix **R** (2.3), we are confident that our test loses very little power when compared with a score test based on the correct correlations.

Table 6
Ten thousand simulations of dichotomous data under null models and alternative models: various marginal probabilities P, allele frequencies p, $P_A = P(Y_j = 1|G_j \in \{AA, Aa\})$, $P_{aa} = P(Y_j = 1|G_j = aa)$ *and the power*

P	p	P_A	P_{aa}	Power
.10	—	.10	.10	.043
.10	.01	.99	.08	.139
.20	—	.20	.20	.054
.20	.01	.99	.18	.097
.20	.10	.70	.08	.320
.20	.10	.99	.01	.776
.40	—	.40	.40	.053
.40	.01	.99	.39	.070
.40	.10	.70	.33	.110
.40	.10	.99	.26	.328
.50	—	.50	.50	.054
.50	.01	.99	.49	.063
.50	.10	.70	.45	.075
.50	.10	.99	.39	.219

Table 7
Correlations between two members of sibship and between a parent and a child for a dominant model with allele frequencies .1 *and* .1

Pair	$p = .01$	$p = .1$
Sibs	.499	.487
Child–parent	.497	.474

6. Example

In a study of familial aggregation of the response to endotoxin stimulation in whole blood (WB) and monocyte cultures (MO), seven pedigrees from volunteers (with sizes 10, 10, 6, 8, 5, 4, and 2) were collected randomly. The third and sixth families have one missing observation of the response in whole blood (WB). The median, mean, and standard deviation of WB are 15216.0, 14667.4, and 5577.7, respectively; the median, mean, and standard deviation of MO are 6910.0, 8024.5, and 3942.0, respectively; the correlation between WB and MO is .47.

The question of interest is whether the response is genetically influenced. To show the test for dichotomous variables, both response variables were dichotomized, using the medians as cut-off point. The statistics are given in Table 8. The linear combination of the quantitative factors, for which Q takes its maximum, is also calculated. The statistics of the quantitative factors are given in Table 9.

It is clear that the hypothesis of no family aggregation for these data cannot be rejected. The quantitative and dichotomized response variable WB gives a lower value of Q than the expectation of Q under the null hypothesis. It can be concluded that the response WB shows no familial correlation. The response MO shows some genetic effect, but the effect is not statistically significant. This may be due to the small size of the data set.

Table 8
Statistics of the dichotomized responses WB and MO

	Q	$E(Q)$	VAR(Q)	c	ν	P value
WB	9.81	9.99	4.40	.22	45.39	.51
MO	13.44	10.49	4.67	.22	46.06	.09

7. Discussion

A test is given to verify whether familial aggregation in data in randomly chosen pedigrees exists. From simulations with 43 people, it appears that the test performs well; with strong genetic effects the power is reasonable, but with weaker genetic effects this particular data set seems too small. When the variance of the genetic effect is small, an impression of the power can be obtained by using the approximations of the expectation and variance of Q (formulae (5.1) and (5.2)). These approx-

Table 9
Statistics of the quantitative responses WB, MO, and a linear combination for which Q takes its maximum

	Q	$E(Q)$	VAR(Q)	c	ν	P value
WB	35.81	39.96	65.62	.82	48.67	.68
MO	51.08	41.95	69.98	.83	50.30	.14
−WB+5.83MO	50.04	39.96	65.62	.82	48.67	.11

imations probably perform better in larger data sets. Note that the power depends not only on the number of individuals but also on the structure of the pedigrees. For weak genetic effects, τ^2 can be approximated using formulae (5.3), but for stronger genetic effects this approximation is poor.

Because of the complex structure of the model parameters and the large set of genotype combinations for a certain pedigree, genetic modeling is quite complicated. Moreover, incorrect modeling may have large effects on the estimates of the parameters, whereas an incorrect correlation matrix simply reduces power. Testing the correlation between relatives by means of our test is therefore a necessary preliminary procedure. If these correlations are significant, genetic models can then be fitted to the data to study the type of heritability.

Model (2.1) can be extended by incorporating covariates X:

$$E(Y_j|u_j) = h^{-1}(X'\gamma + u_j).$$

Let d be the rank of X and H the projection matrix on the d dimensional subspace spanned by the columns of X and $\tilde{\mathbf{R}} = (I - H)\mathbf{R}(I - H)$. Then if Y follows a normal distribution the expectation and variance of Q are

$$E(Q) = \text{trace}(\tilde{\mathbf{R}})$$

and

$$\text{VAR}(Q) = \frac{2}{n - d + 2}\left((n - d)\text{trace}(\tilde{\mathbf{R}}^2) - \text{trace}^2(\tilde{\mathbf{R}})\right).$$

For logistic regression, the variances of Y_i are not identical, and we propose to replace the denominator of Q, σ^2, by $1/(n - d)\,\Sigma\, v_{ii}$, where v_{ii} is equal to the variance of Y_i under the null hypothesis. To correct for estimation of $\mu = h^{-1}(X'\gamma)$ the following approximation can be used: $Y - \hat{\mu} \approx (I - H)(Y - \mu)$, where $H = \mathbf{V}X(X'\mathbf{V}X)^{-1}X'$, where $\mathbf{V}$ is the diagonal matrix with elements v_{ii} (le Cessie and van Houwelingen, 1991).

The test statistic can only be used for randomly chosen pedigrees. A suitable test for pedigrees that are selected because of the response of a proband should be based on the conditional likelihood given the response of the proband.

Résumé

Calculer la vraisemblance de généalogies est complexe et souvent consommateur de temps. Tester des structures de corrélation dues à la présence d'une agrégation familiale constitue en conséquence une procédure préliminaire. Une statistique du score est présentée pour tester les corrélations entre membres de généalogies choisies au hasard. Cette statistique peut être utilisée pour des données quantitatives et qualitatives. Pour ces deux types de données, la distribution de la statistique sous l'hypothèse nulle est dérivée. Pour démontrer la performance de la statistique, les résultats de simulations sous divers modèles sont exposés. Finalement, le test est appliqué à des données sur un facteur sanguin continu.

References

Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics*. Oakland, California: Holden-Day.

Bonney, G. E. (1984). On the statistical determination of major gene mechanisms in continuous human traits: Regressive models. *American Journal of Human Genetics* **18,** 731–749.

Bonney, G. E. (1986). Regressive logistic models for familial disease and other binary traits. *Biometrics* **42,** 611–625.

Elandt-Johnson, R. C. (1971). *Probability Models and Statistical Methods in Genetics*. New York: John Wiley.

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

le Cessie, S. and van Houwelingen, J. C. (1991). A goodness of fit test for binary regression models, based on smoothing methods. *Biometrics* **47,** 1267–1282.

le Cessie, S. and van Houwelingen, J. C. (1995). Testing the fit of a regression model via score tests in random effect models. *Biometrics* **51,** 600–614.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall.

Kendall, M. A. and Stuart, A. (1963). *The Advanced Theory of Statistics*, Vol. I. London: Griffin.

Thompson, E. A. (1986a). *Pedigree Analysis in Human Genetics*. Baltimore and London: Johns Hopkins University Press.

Thompson, E. A. (1986b). Genetic epidemiology: A review of the statistical basis. *Statistics in Medicine* **5,** 291–302.

Received April 1994; *revised February* 1995; *accepted March* 1995.

Appendix

Derivation of the expectation and variance of Q for Y independently and normally distributed

The distribution of Q under the null hypothesis when Y_t are normally distributed can be derived by defining

$$Q = \frac{N}{D} = (n-1)\frac{(Y-\hat{\mu})'\mathbf{R}(Y-\hat{\mu})}{(Y-\hat{\mu})'(Y-\hat{\mu})} = (n-1)\frac{(Y-\mu)'\tilde{\mathbf{R}}(Y-\mu)}{(Y-\mu)'(I-H)(Y-\mu)} = (n-1)\frac{\sum \lambda_t z_t^2}{\sum z_t^2},$$

where λ_t are the eigenvalues of the matrix $\tilde{\mathbf{R}} = (I-H)\mathbf{R}(I-H)$ and z_t are $(n-1)$ orthonormal transformations of the response variables $Y-\mu$. Since Y_t are independent and normally distributed, the values of z_t are independent and z_t^2 and $\Sigma_{j\neq t} z_j^2$ follow a χ^2 distribution with 1 and $(n-2)$ degrees of freedom, respectively, hence

$$x_t = \frac{z_t^2}{z_t^2 + \sum_{j\neq t} z_j^2}$$

follows a $\beta(1/2,\ 1/2(n-2))$ distribution. It follows that $x_1 \cdots x_{n-1}$ are identically distributed with correlation $-1/(n-2)$,

$$E(x_t) = \frac{1}{n-1}$$

and

$$\mathrm{VAR}(x_t) = 2\frac{n-2}{(n-1)^2(n+1)}$$

(Bickel and Doksum, 1977, p. 44). Hence, the mean and variance of Q are

$$E(Q) = \sum \lambda_t = \mathrm{trace}(\tilde{\mathbf{R}})$$

and

$$\mathrm{VAR}(Q) = \frac{2}{n+1}\left((n-1)\mathrm{trace}(\tilde{\mathbf{R}}^2) - \mathrm{trace}^2(\tilde{\mathbf{R}})\right).$$