Testing Association between Candidate-Gene Markers and Phenotype in Related Individuals, by Use of Estimating Equations

David-Alexandre Trégouët, Pierre Ducimetière, and Laurence Tiret

INSERM Unité 258, Hôpital Broussais, Paris

Summary

Association studies are one of the major strategies for identifying genetic factors underlying complex traits. In samples of related individuals, conventional statistical procedures are not valid for testing association, and maximum likelihood (ML) methods have to be used, but they are computationally demanding and are not necessarily robust to violations of their assumptions. Estimating equations (EE) offer an alternative to ML methods, for estimating association parameters in correlated data. We studied through simulations the behavior of EE in a large range of practical situations, including samples of nuclear families of varying sizes and mixtures of related and unrelated individuals. For a quantitative phenotype, the power of the EE test was comparable to that of a conventional ML test and close to the power expected in a sample of unrelated individuals. For a binary phenotype, the power of the EE test decreased with the degree of clustering, as did the power of the ML test. This result might be partly explained by a modeling of the correlations between responses that is less efficient than that in the quantitative case. In small samples (<50 families), the variance of the EE association parameter tended to be underestimated, leading to an inflation of the type I error. The heterogeneity of cluster size induced a slight loss of efficiency of the EE estimator, by comparison with balanced samples. The major advantages of the EE technique are its computational simplicity and its great flexibility, easily allowing investigation of gene-gene and gene-environment interactions. It constitutes a powerful tool for testing genotype-phenotype association in related individuals.

Introduction

Association studies based on candidate genes are one of the major strategies used to identify genetic factors underlying complex traits. Compared with linkage studies, this approach has far greater power to detect genes of modest effect, such as those involved in the susceptibility to complex human diseases (Risch and Merikangas 1996). Owing to increasing availability of candidate genes for many diseases, this approach is likely to develop in the future (Lander 1996).

Association studies rely on the direct identification of variants predisposing to disease or on the linkage disequilibrium existing, at a populational level, between measured markers and unknown functional variants at candidate loci. They are usually performed in samples of unrelated individuals, and, in that case, the estimating and testing of association parameters are straightforward by use of conventional statistical procedures. The major advantage of this approach lies in its simplicity and its flexibility, easily allowing investigation of gene-gene and gene-environment interactions that constitute the underlying architecture of complex traits.

Although use of unrelated individuals is a priori the most efficient way of studying the association between measured markers and phenotype, in some situations one could be interested in using familial data for testing association. For example, it is more and more frequent to combine populational and familial approaches—as, for example, a case-control study and a sib-pair study—in which the same candidate genes are investigated (Jeunemaitre et al. 1992; Bonnardeaux et al. 1994; Lindpaintner et al. 1996). It would then be tempting to use all the available information for testing marker-phenotype association. Likewise, with increasing availability of family data sets, one may wish to screen data for possible association between phenotype and candidate-gene markers and to identify potentially relevant interactions, before embarking on more time-consuming segregation/linkage analysis. Last, the large-scale samples currently collected for linkage analysis (e.g., affected sib pairs) of complex diseases could also be used for association analysis (Risch and Merikangas 1996). However, as far as related individuals are concerned, conventional statistical methods are no longer valid and may lead to incorrect inferences. Maximum likelihood (ML) methods specifying the joint family distribution of the trait have to be used, but these methods are computationally demanding and are not necessarily robust to violations of

Received November 15, 1996; accepted for publication April 23, 1997.

Address for correspondence and reprints: Dr. Laurence Tiret, INSERM U258, Hôpital Broussais, 96 rue Didot, 75674 Paris Cedex 14, France. E-mail: tiret@hbroussais.fr

© 1997 by The American Society of Human Genetics. All rights reserved.
0002-9297/97/6101-0026$02.00

their assumptions. This problem is crucial, since it is generally not possible to check the validity of these assumptions.

The EE approach offers an alternative to ML methods, for studying a genotype-phenotype association in samples of related individuals. This technique was initially proposed, independently, by Godambe (1960) and Huber (1964). Liang and Zeger (1986) contributed to popularization of this technique, through successful application of it to longitudinal data analysis (for a review, see Godambe 1991). The EE method is a general approach for estimating regression parameters for correlated data that makes no distributional assumption, unlike ML methods, but it only models the expectation of the marginal moments of the data as functions of covariates. The EE method is robust in the sense that consistent estimates of regression parameters and their standard errors are obtained even though correlations between responses are partially misspecified. As pointed out by Zhao et al. (1992*b*), for the partly exponential family the proposed EE has a form identical to that of the score-estimating equation, establishing an equivalence between the EE and the ML approaches. In fact, EE encompasses ML, since the score-estimating equation under ML is a particular case of EE, in which a specific distribution is assumed.

The EE technique has recently been introduced to human genetics, and several methods have been developed for analyzing familial data (Liang and Beaty 1991; Zhao et al. 1992*a;* Grove et al. 1993; Olson 1994*a;* Hsu and Zhao 1996; Liang and Pulver 1996). Applications to nonparametric linkage analysis (Olson and Wijsman 1993; Olson 1994*b*) and segregation analysis (Lee et al. 1993; Stram et al. 1993; Whittemore and Gong 1994; Zhao 1994; Zhao and Grove 1995; Lee and Stram 1996) have also been proposed.

In this paper, we consider an application of the EE technique to the problem of association between measured markers and phenotypes, either quantitative or binary. The EE properties of robustness and efficiency have been shown to be asymptotically valid, but less is known about the technique's behavior in small samples or when cluster sizes are unequal. We studied, through simulations, the behavior of EE in a large range of practical situations, including small samples of families and mixtures of related and unrelated individuals.

Material and Methods

EE

Consider a sample of K families with family k consisting of n_k individuals. Let $y_k^t = (y_{k1}, \ldots, y_{kn_k})$ (t denotes "transposition") denote the vector of phenotypes of the kth family, with expected mean $\mu_k^t = (\mu_{k1}, \ldots, \mu_{kn_k})$, and let $x_k^t = (x_{k1}, \ldots, x_{kn_k})$ denote a $p \times n_k$ matrix of p covariates. The EE approach (Liang and Zeger 1986) requires no assumption about the joint distribution of the y_{ki} but assumes only that the marginal distribution of y_{ki} ($i = 1, \ldots, n_k$) has a mean correctly specified by a known function, referred to as the "link function," of a linear combination of the covariates x_{ki} with a vector of regression coefficients $\beta^t = (\beta_1, \ldots, \beta_p)$ to be estimated. A consistent estimate of β is obtained by solving the following EE:

$$\sum_{k=1}^{K} \frac{\partial\mu_k^t}{\partial\beta} \mathrm{Var}(y_k)^{-1}(y_k - \mu_k) = 0 \; . \tag{1}$$

Provided that only the mean is correctly specified, the solution of equations (1) is such that $K^{1/2}(\hat{\beta} - \beta)$ is asymptotically normally distributed with mean 0 and covariance matrix consistently estimated by

$$V(\hat{\beta}) = KW^{-1}\left(\sum_{k=1}^{K} \frac{\partial\mu_k^t}{\partial\beta} \mathrm{Var}(y_k)^{-1}(y_k - \mu_k) \times (y_k - \mu_k)^t \mathrm{Var}(y_k)^{-1} \frac{\partial\mu_k}{\partial\beta}\right)W^{-1} \; , \tag{2}$$

where

$$W = K^{-1} \sum_{k=1}^{K} \frac{\partial\mu_k^t}{\partial\beta} \mathrm{Var}(y_k)^{-1} \frac{\partial\mu_k}{\partial\beta} \; ,$$

the quantity (2) being evaluated at $\hat{\beta}$. The robustness property of EE relies on the fact that consistent estimates of the parameters and of their variances are obtained even if the dependency between familial phenotypes is not correctly specified. In other terms, when $\mathrm{Var}(y_k)$ is written as $\mathrm{diag}(\mathrm{Var}[y_{ki}])^{1/2}R_k\mathrm{diag}(\mathrm{Var}[y_{ki}])^{1/2}$, where R_k is a "working correlation matrix" specifying the correlations between individuals of family k, the solutions obtained by EE are robust to any misspecification of R_k. This is particularly true when individuals within a family are taken to be independent; that is, when R_k is taken as the identity matrix. However, the better that R_k specified, the more efficient are the estimates $\hat{\beta}$ and $\mathrm{V}(\hat{\beta})$. As mentioned by Rotnizsky and Jewell (1990), misspecification of the working correlation matrix may have a greater impact on the efficiency of the EE estimate when cluster size is not constant, which is generally the case in family studies. Note that equation (1) is the score equation under models of the partly exponential family (Zhao et al. 1992*b*), including the multivariate normal distribution, provided that $\mathrm{Var}(y_k)$ is correctly specified. In this case, EE and ML parameter estimates are equivalent. However, such equivalence for the variance-covariance matrix holds only asymptotically (Park 1993).

Application to Genotype-Phenotype Association

In the following applications, the phenotype y_{ki} of the ith individual in the kth family is assumed to result from a genotype effect measured at a diallelic locus A/a and from an independent residual component e_{ki}, which can also be the source of family resemblance. In the most general form, the genotype is a set of two indicator variables associated with the genotypes Aa and AA, respectively, with the genotype aa being taken as the reference. Under specific genetic models (additive, recessive, or dominant), this set reduces to only one variable. The extension to a multiallelic locus or to several loci is straightforward.

Quantitative phenotype.—For a quantitative phenotype, the link function "identity" is used to relate the mean to genotype; that is, $E(y_{ki}|x_{ki}) = \alpha + \beta x_{ki}$, where $\beta = (\beta_1,\beta_2)$ is the vector of association parameters. Note that β_1 (or β_2, respectively) represents the mean difference of the trait, between Aa (or AA, respectively) and aa subjects. The residual variances and the working residual correlations based on the interclass and intraclass pairwise correlations (Donner and Eliasziw 1991) are

$$\mathrm{var}(e_{kt}) = \frac{1}{K_1}\sum_{k=1}^{K_1} n_{kt}^{-1}\sum_{j=1}^{n_{kt}} e_{kj}^2 \ ,$$

and

$$\mathrm{corr}(e_{ks}, e_{kt}) = \left(\frac{1}{K_2}\sum_{k=1}^{K_2} n_{ks}^{-1}n_{kt}^{-1}\sum_{j=1}^{n_{ks}} e_{kj}\sum_{m=1}^{n_{kt}} e_{km}\right)\Bigg/ \sqrt{\mathrm{var}(e_{ks})\mathrm{var}(e_{kt})} \ ,$$

and

$$\mathrm{corr}(e_{kt}, e_{kt}) = \left(\frac{1}{K_3}\sum_{k=1}^{K_3} n_{kt}^{-1}(n_{kt} - 1)^{-1} \times \sum_{j=1}^{n_{kt}}\sum_{\substack{m=1\\ m\neq j}}^{n_{kt}} e_{kj}e_{km}\right)\Bigg/ \mathrm{var}(e_{kt}) \ ,$$

where s and t refer to a subclass of relatives (fathers, mothers, sons, daughters, grandfathers, . . .); n_{ks} (or n_{kt}, respectively) is the number of individuals within subclass s (or t, respectively) in family k; and K_1, K_2, and K_3 are the number of families over which these parameters are estimated. In the case of a fixed family structure (e.g., nuclear families of equal size), $K_1 = K_2 = K_3 = K$. In this special case, the interclass and intraclass pairwise estimators are identical to the ML estimators under the multivariate normal distribution. This is no longer true when the cluster size varies or when the distribution is not normal.

Binary phenotype.—For a binary phenotype (e.g., a disease), the link function "logit" is used; that is, $E(y_{ki}|x_{ki}) = \pi_{ki} = [\exp(\alpha + \beta x_{ki})/[1 + \exp(\alpha + \beta x_{ki})$, where $\beta = (\beta_1,\beta_2)$ is the vector of association parameters, e^{β_1} and e^{β_2} being the odds ratios for disease associated with genotypes. In that case, residual variances are those of a Bernoulli variable; that is, $\mathrm{var}(e_{ki}) = \hat{\pi}_{ki}(1 - \hat{\pi}_{ki})\forall i = 1, . . . , n_k$. The best way of modeling the association between a pair of binary responses is not obvious. With pairwise correlations as originally proposed by Prentice (1988), interclass and intraclass working residual correlations were estimated by

$$\mathrm{corr}(e_{ks}, e_{kt}) = \frac{1}{K_2}\sum_{k=1}^{K_2} n_{ks}^{-1}n_{kt}^{-1} \times \sum_{j=1}^{n_{ks}}\frac{e_{kj}}{\sqrt{\hat{\pi}_{kj}(1 - \hat{\pi}_{kj})}}\sum_{m=1}^{n_{kt}}\frac{e_{km}}{\sqrt{\hat{\pi}_{km}(1 - \hat{\pi}_{km})}}$$

and

$$\mathrm{corr}(e_{kt}, e_{kt}) = \frac{1}{K_3}\sum_{k=1}^{K_3} n_{kt}^{-1}(n_{kt} - 1)^{-1} \times \sum_{j=1}^{n_{kt}}\sum_{\substack{m=1\\ m\neq j}}^{n_{kt}}\frac{e_{kj}e_{km}}{\sqrt{\hat{\pi}_{kj}(1 - \hat{\pi}_{kj})\hat{\pi}_{km}(1 - \hat{\pi}_{km})}} \ ,$$

where s, t, K_2, and K_3 are defined as above. Note that in the binary situation there is no obvious ML method for analyzing correlated data.

Simulated Data

Simulation studies were performed to study the performances of EE in terms of power, bias, and type I error. In all simulations, the A-allele frequency was fixed to .3. Each individual was assigned a genotype, under Hardy-Weinberg laws (for founders) and Mendelian transmission laws (for offspring). The genetic model was considered to be strictly codominant; that is, $\beta_1 = \beta_2/2 = \beta$. For a quantitative phenotype, this model corresponds to an additive model, β being the mean effect associated with allele A, whereas, for a binary phenotype, this model corresponds to a multiplicative model, e^{β} being the odds ratio associated with allele A. In the quantitative situation, the vector of residual familial phenotypes was generated from a standardized multivariate normal distribution. Spouses were uncorrelated, and the correlation between parent and offspring was identical to the correlation between sibs. In the binary situation, the trait (e.g., a disease) was generated from a truncated underlying multinormal distribution with the same family correlation structure as in the quantitative situation. An individual was considered as affected if his or her phenotype value was greater than a given threshold, this latter being a function of the prevalence of the disease in the population (fixed to .25), the allele frequency in unaffected individuals (fixed to .3), and the allelic odds ratio e^{β}.

We first considered samples composed of nuclear families of equal size (fixed clusters). Then we considered samples composed of mixtures of clusters of different type (nuclear families of varying size, sibships, and unrelated individuals). Unrelated individuals, in that case, are considered as families of size 1.

Analysis Methods

For implementing the EE method, we developed our own program in C language. The performances of the EE method were assessed in terms of power, relative bias (mean of the parameter estimate minus the true value divided by the true value), coverage probability (probability that the observed 95% confidence interval includes the true value β), mean square error of the parameter, and type I error.

In simulations performed on fixed clusters, we also compared the power and type I error of the EE test with those of the conventional test of association, used for unrelated individuals, which does not take into account the family structure. This test will be referred to as the "naive" test. We also analyzed the data by using a conventional ML method based on a measured genotype analysis (Boerwinkle et al. 1986). For this purpose, we used, for quantitative phenotype, our own program, which is based on a regressive model assuming that the penetrance function within a family is the multinormal density function (for a detailed description of the model, see Georges et al. 1996). For the binary phenotype, we used the REGRESS program (Demenais and Lathrop 1994), in which the penetrance is modeled by a logistic function depending on the genotype-dependent baseline risk and on residual family dependencies. These residual family dependencies are modeled by specifying a regression relationship between a person's phenotype, the phenotypes of antecedents, and the genotype.

All EE simulations were conducted on 1,000 replicates. Because the ML methods are extremely computing-time demanding, the corresponding simulations were conducted on only 200 replicates. The null hypothesis $\beta = 0$ was tested in EE analyses by a Wald test using the statistic $\hat{\beta}^2/\mathrm{var}(\hat{\beta})$ and in ML analyses by a likelihood-ratio test. In both cases, the statistics follow, under the null hypothesis, a χ^2 distribution with 1 df. The significance level was taken to be .05. The following section presents a summary of the results, but all detailed results are available on request.

Results

Fixed Clusters: Simulations with $\beta \neq 0$

Quantitative phenotype.—Three different values of β—.2, .3, and .4—were successively considered, corresponding to a proportion of variance explained by the marker (h^2) of 1.6%, 3.6%, and 6.3%, respectively. The h^2 value first determined the total number of individuals to be sampled if these individuals were unrelated; for example, for h^2 = 1.6%, the required number of unrelated individuals would be 626, in order to detect the marker effect with 90% power at a nominal level of .05. To examine the influence of clustering, this total number was then divided successively into *K* families of sibship size 1 (*K* = 209), sibship size 2 (*K* = 157), and sibship size 4 (*K* = 104). For h^2 = 3.6% and 6.3%, the numbers of families are shown in figure 1*A*. The residual family correlation was successively fixed to .0, .1, .3, and .5.

The power of the EE test was first compared with that of the naive and ML tests. A mean power was calculated

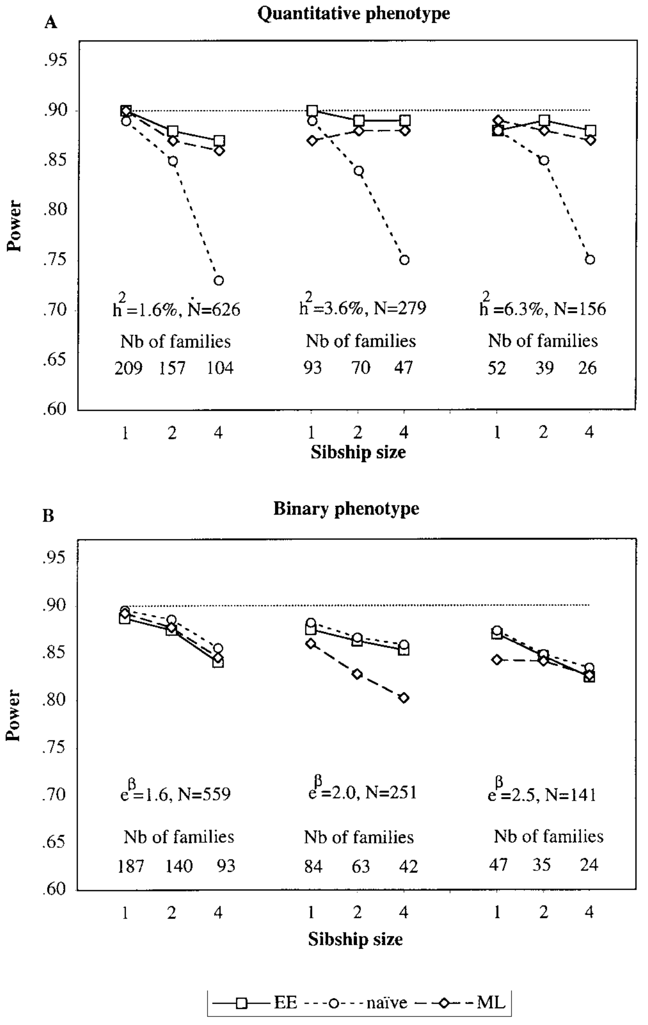

Figure 1 Fixed clusters: power of the different tests to detect association, according to marker effect and sibship size (mean power over the four different values of residual family correlation). *A,* Quantitative phenotype. *B,* Binary phenotype. The horizontal line indicates power of .90, which would be expected if unrelated individuals had been sampled. N = total number of individuals.

Table 1

Quantitative Trait, Fixed Clusters: Power, Relative Bias, and Coverage Probability of EE Estimate of Association Parameter in Small Samples

	Sibship Size 1, $K = 52$			Sibship Size 2, $K = 39$			Sibship Size 4, $K = 26$		
Residual Correlation	Power	Bias (%)	Coverage Probability	Power	Bias (%)	Coverage Probability	Power	Bias (%)	Coverage Probability
.0	.910	−1.0	.920	.912	−.8	.922	.899	.7	.887
.1	.891	−.6	.928	.888	1.5	.906	.862	−.1	.905
.3	.844	−2.5	.925	.872	.8	.917	.856	.0	.911
.5	.897	−.4	.924	.900	−.4	.912	.903	.0	.918

Note.—Data are for nuclear families of equal size; $h^2 = 6.3\%$; total number of individuals is 156, giving a 90% power in a sample of unrelated individuals; and nominal level $\alpha = .05$.

over the four different values of residual family correlation (fig. 1*A*). In all cases, the power of the ML and the EE tests was close to .90, the expected power if unrelated individuals, rather than families, had been sampled. This first result indicates that sampling relatives yields only a slight loss of power, compared with the use of unrelated individuals, provided that the dependency between individuals is correctly specified. The power of the two methods was little influenced by the extent of clustering—that is, the sibship size—even when the number of families was relatively small ($K < 50$). As expected, ignoring the within-family correlation (the naive test) induced a loss of power that dramatically increased with the extent of clustering. Whatever the sibship size, the power of the EE and ML tests followed a U-shaped curve with a minimum power observed for a correlation of .3. By contrast, the loss of power of the naive test increased with the magnitude of the within-cluster correlation (data not shown).

The bias of the EE estimate was never >3% of the true value of β. The coverage probability was close to the designed value of .95, in large samples. In small samples ($K < 50$), the coverage probability was lowered, although the power and the relative bias remained within acceptable ranges (table 1). Actually, in small samples, the EE variance of the parameter tended to be underestimated, compared with the mean square error that provides an estimate of the true variance. As a consequence, the confidence interval was smaller than it should be.

Binary phenotype.—When the frequency of the *A* allele was set to .3 in unaffected individuals, three different values of the allelic odds ratio—1.6, 2.0, and 2.5—were successively considered, corresponding to an *A*-allele frequency, in affected individuals, of .406, .461, and .517, respectively. In a manner similar to that in the quantitative situation, the allelic odds ratio determined the total number of unrelated individuals and the subsequent number of families of sibship size 1, 2, and 4, respectively (see fig. 1*B*). The different sample sizes were chosen to be roughly similar to those of the simulations for a quantitative trait. Again, four different values of the residual family correlation (0, .1, .3, and .5) in the underlying liability distribution were considered.

The effect of sibship size on the power of the EE, ML, and naive tests is shown in figure 1*B*. For the three tests, the power decreased with increasing sibship size, whatever the sample size. This result was at variance with that of the quantitative case, in which the power of the EE and ML tests seemed rather insensitive to the degree of clustering. The power also decreased with increasing correlation, the decrease appearing more pronounced in larger sibships (data not shown). It should be stressed that, unlike the quantitative case, in which data were analyzed under the true model of generation, in the binary case the phenotype was generated from a truncated continuous variable but was analyzed as a dichotomous variable, which probably induced a loss of power. Moreover, whereas in the quantitative situation all families contribute to estimation, the most informative families for a binary trait are those with several affected members, whereas families with no affected member poorly contribute to the estimation of β. Last, the way of modeling the dependency between binary observations (by pairwise correlations or by odds ratios) might affect the efficiency of estimation in both ML and EE analyses. Unlike the quantitative case, the naive test appeared to have a power identical to that of the EE test, in all situations, but these results should be tempered by the fact that the type I error of the naive test was largely inflated (see below).

In large samples, as for a quantitative trait, the EE estimate behaves well in terms of bias and coverage probability (data not shown). In small samples ($K < 50$), the bias of the EE estimate was slightly larger than that in the quantitative situation, and the coverage probability again was lower than the designed probability of .95 (table 2).

Table 2

Binary Trait, Fixed Clusters: Power, Relative Bias, and Coverage Probability of EE Estimate of Association Parameter in Small Samples

	Sibship Size 1, K = 47			Sibship Size 2, K = 35			Sibship Size 4, K = 24		
Residual Correlation	Power	Bias (%)	Coverage Probability	Power	Bias (%)	Coverage Probability	Power	Bias (%)	Coverage Probability
.0	.889	2.2	.941	.882	2.8	.924	.871	0.8	.919
.1	.877	3.8	.929	.851	1.7	.919	.843	3.3	.912
.3	.854	2.5	.931	.842	3.8	.932	.801	1.7	.911
.5	.860	4.6	.926	.810	2.3	.925	.783	2.7	.919

Note.—Data are for nuclear families of equal size; allelic odds ratio associated with the marker is 2.5; total number of individuals is 141, giving a 90% power in a sample of unrelated individuals; and nominal level α = .05.

Fixed Clusters: Simulations with β = 0

The total sample size was fixed successively to 600, 300, and 120, in order to be of the same order of magnitude as that considered in the power simulations. The number of families of sibship size 1, 2, and 4 was deduced from this total number. The residual family correlations varied over the same values as have been reported above.

Quantitative phenotype.—As for power, the type I error of the three tests was compared, for varying sibship sizes, the mean error being calculated over the four different within-family correlations (fig. 2*A*). The observed type I error of the ML test was close to the nominal value of .05. By contrast, as already observed for power, the asymptotic properties of the EE method did not seem to hold for small samples, in which the observed type I error was substantially inflated. For $K \leq 30$, it was even >.10. This inflation was due to an underestimation of the EE estimate of the variance, an underestimation already observed in the simulations with $\beta \neq 0$ for small samples. Looking more deeply into our simulation results revealed that an underestimation of the residual correlations for small samples might explain this inflation, as already reported by Hendricks et al. (1996). As expected, the naive test yielded a type I–error inflation that increased both with the sibship size and with the within-family correlation. However, for low within-family correlation, this inflation appeared to be smaller than that for the EE test, especially in small samples (data not shown). Actually, in small samples, the advantage of using a presumably more accurate working correlation matrix might be offset by the need to estimate more nuisance parameters, which may create finite-sample instability (Rotnitzky and Jewell 1990; Liang and Pulver 1996).

Binary phenotype.—The behaviors of the three tests were quite similar to those observed in the quantitative simulations (fig. 2*B*). However, the inflation of the type I error in the EE test was lower than that in the quantita-

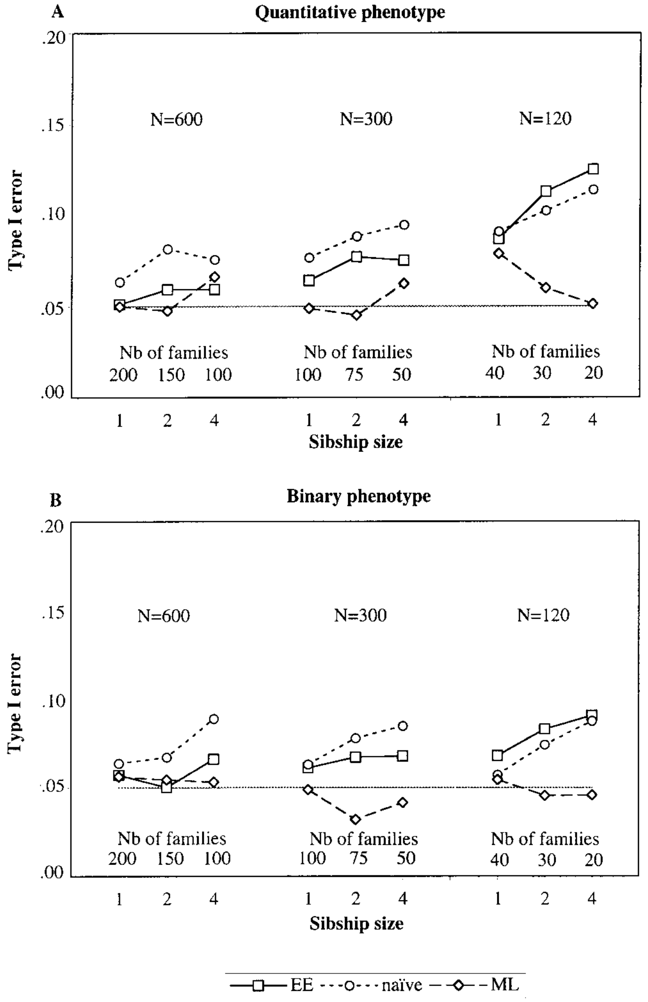

Figure 2 Fixed clusters: observed type I error of the different tests, according to number of families and sibship size (mean power over the four different values of residual family correlation). *A*, Quantitative phenotype. *B*, Binary phenotype. The horizontal line indicates the nominal type I error, .05. N = total number of individuals.

Table 3

Quantitative Phenotype, Varying Clusters: Power, Relative Bias, and Coverage Probability of EE Estimate, According to Sample Structure

	h^2 = 1.6% (N = 626)			h^2 = 3.6% (N = 279)			h^2 = 6.3% (N = 156)		
Residual Correlation	Power	Bias (%)	Coverage Probability	Power	Bias (%)	Coverage Probability	Power	Bias (%)	Coverage Probability
	K = 144			K = 64			K = 36		
Structure 1 (NF1 25%, NF2 25%, NF4 50%):									
.1	.839	−.4	.941	.848	2.0	.921	.833	−2.1	.878
.5	.851	−2.0	.923	.831	2.9	.901	.755	−4.6	.908
	K = 271			K = 121			K = 68		
Structure 2 (NF1 40%, SP 60%):									
.1	.822	.2	.942	.892	.2	.938	.889	1.2	.916
.5	.903	.2	.944	.885	−.7	.923	.846	−1.3	.878
	K = 250			K = 112			K = 62		
Structure 3 (NF2 40%, SP 60%):									
.1	.866	.6	.913	.860	1.7	.887	.855	.3	.846
.5	.853	1.4	.899	.831	−.7	.894	.826	1.2	.849
	K = 501			K = 243			K = 125		
Structure 4 (SP 40%, UI 60%):									
.1	.886	1.4	.950	.897	−.08	.941	.880	−.2	.923
.5	.893	.2	.919	.893	−.3	.937	.876	−2.1	.934

NOTE.—N = total number of individuals; K = number of clusters; NFi = nuclear families of sibship size i; SP = sib pairs; and UI = unrelated individuals.

tive case. This might be explained by the fact that nuisance parameters to be estimated are fewer, since the residual variance for a binary trait is a bijective function of the mean whereas it is not so for a quantitative trait.

Varying Clusters: Simulations with $\beta \neq 0$

For both quantitative and binary phenotypes, four different sample structures were considered: (1) a mixture of nuclear families with different sibship sizes (1, 2, and 4); (2) a mixture of sib pairs and nuclear families with sibship size 1; (3) a mixture of sib pairs and nuclear families with sibship size 2; and (4) a mixture of sib pairs and unrelated individuals. The proportion of individuals in each type of cluster is given in tables 3–5. All simulations were performed for two contrasted values of the residual family correlation, .1 and .5. The total sample sizes considered were the same as for the fixed sample structure.

Quantitative phenotype.—Detailed results are given in table 3. When the sample was composed of clusters similar in size (structures 2 and 4), the power was close to that observed for a fixed sample structure. By contrast, in samples composed of clusters more heterogeneous in size (structures 1 and 3), a slight loss of power of the EE test was observed, especially for high residual correlation in small samples. Looking more deeply into the results suggested that this loss of power was due to a higher mean square error of the association parameter than was seen in the fixed structure. As for fixed clusters, the bias was not >3%, with one exception (K = 36). Again, a decrease of the coverage probability was observed in smaller samples. For comparable sample sizes, the decrease was more marked in small samples composed of unequal clusters than in small samples composed of equal clusters. The lower coverage probabilities were observed in structure 3, composed of a mixture of clusters of size 2 and 4.

Binary phenotype.—Quite similar results were observed for a binary phenotype (table 4). The loss of power of the EE test was negligible when the clusters were similar in size (structures 2 and 4) but increased with the heterogeneity of cluster sizes, especially for high residual correlation. As for the quantitative phenotype,

Table 4

Binary Phenotype, Varying Clusters: Power, Relative Bias, and Coverage Probability of EE Estimate According to Sample Structure

	Allelic Odds Ratio 1.6 (*N* = 559)			Allelic Odds Ratio 2.0 (*N* = 251)			Allelic Odds Ratio 2.5 (*N* = 141)		
Residual Correlation	Power	Bias (%)	Coverage Probability	Power	Bias (%)	Coverage Probability	Power	Bias (%)	Coverage Probability
	K = 128			*K* = 58			*K* = 32		
Structure 1 (NF1 25%, NF2 25%, NF4 50%):									
.1	.860	.4	.933	.861	1.6	.927	.839	.4	.936
.5	.856	2.3	.934	.827	4.7	.915	.797	.8	.928
	K = 241			*K* = 109			*K* = 61		
Structure 2 (NF1 40%, SP 60%):									
.1	.895	.4	.952	.859	−.2	.938	.849	2.4	.943
.5	.867	.2	.944	.864	.8	.945	.838	3.4	.943
	K = 224			*K* = 100			*K* = 57		
Structure 3 (NF2 40%, SP 60%):									
.1	.866	.8	.910	.839	−.1	.898	.854	3.8	.884
.5	.822	−.3	.914	.814	1.2	.895	.825	4.1	.912
	K = 447			*K* = 201			*K* = 114		
Structure 4 (SP 40%, UI 60%):									
.1	.879	−.3	.938	.907	2.9	.939	.866	2.4	.931
.5	.888	.3	.952	.887	.7	.943	.865	3.4	.939

Note.—Abbreviations are as in table 3.

the lower coverage probabilities were observed in structure 3.

Varying Clusters: Simulations with $\beta = 0$

Quantitative and binary phenotypes.—Simulations were performed for residual correlations of .1 and .5, but results are reported in table 5 only for correlation of .1, since similar findings were obtained for correlation of .5. For large samples, in almost all situations, except for the case in which samples were composed of clusters of size 2 and 4 (structure 3), the observed type I error of the EE test only slightly exceeded the nominal value of .05. As already observed in the case of fixed clusters, the type I error was substantially inflated in small samples, except in structure 4, probably because the sample was mainly composed of unrelated individuals.

Discussion

In this paper, we have been interested in the application of the EE technique to the problem of association between genetic markers and a trait in related individuals. It is important to recognize that, although dealing with family data, the EE application proposed here does not test for linkage but only for association and then does not overcome the risk of spurious association due to uncontrolled stratification of the population, one of the main pitfalls of association studies.

The EE technique offers several advantages over ML methods, including flexibility of the model (it is easily extended to several markers and gene-environment interactions), computational rapidity, and the possibility of handling incomplete family data or a mixture of related and unrelated individuals. Another major advantage of EE is that it does not require any assumption regarding the joint family distribution. However, it must be stressed again that, when a specific distribution is assumed, the score EE under the ML method is a particular form of EE, in which the covariance matrix is fully parametrized. By contrast, the EE method proposed here could be viewed as an “empirical” EE method, in the sense that the covariance matrix is an empirical one. From this perspective, the comparison between the EE and ML methods that is performed in

Table 5

Quantitative and Binary Phenotypes, Varying Clusters: Observed Type I Error of EE Test, According to Sample Structure and Sample Size

	N = 600	N = 300	N = 120
	K = 138	K = 69	K = 28
Structure 1 (NF1 25%, NF2 25%, NF4 50%):			
Quantitative	.061	.064	.127
Binary	.056	.066	.093
	K = 260	K = 130	K = 52
Structure 2 (NF1 40%, SP 60%):			
Quantitative	.062	.068	.117
Binary	.055	.059	.058
	K = 240	K = 120	K = 48
Structure 3 (NF2 40%, SP 60%):			
Quantitative	.101	.109	.161
Binary	.085	.103	.103
	K = 480	K = 240	K = 96
Structure 4 (SP 40%, UI 60%):			
Quantitative	.066	.064	.060
Binary	.042	.044	.054

NOTE.—Nominal level α = .05; and residual family correlation is .1. Abbreviations are as in table 3.

the present paper is a comparison between two different forms of EE.

Several conclusions can be made from our simulations. For a quantitative trait and clusters of equal size, the power of the EE test based on a completely specified correlation matrix was comparable to that of the ML test and was similar to the power expected in a sample of unrelated individuals. The mean bias of the association parameter was negligible. In large samples, the full efficiency of the EE estimator was thus demonstrated when normality holds and asymptotic conditions are valid (Liang and Zeger 1986; Zhao et al. 1992*b*). However, in small samples (<50 families), the variance of the EE association parameter tended to be underestimated. This underestimation led to a decrease of the coverage probability for $\beta \neq 0$ and to an inflation of the type I error for $\beta = 0$, as already noted by several authors (Emrich and Piedmonte 1992; Olson 1994*b;* Hendricks et al. 1996). The inflation of the type I error could be quite substantial in the presence of a strong clustering effect. The small sample size might explain the relatively high rate of false-positive associations reported by Bull et al. (1995) when they analyzed the Genetic Analysis Workshop 9 data on 23 extended families.

For a binary trait and clusters of equal size, the power of the EE test tended to decrease as within-cluster effect increased, but this phenomenon was also observed for the ML test. Several reasons might explain this result. First, as already stressed above, the way of modeling the dependency between two binary responses is not obvious. In ML analyses, we used a regressive approach that models the family dependencies, by regressing a person's phenotype on those of preceding relatives (Demenais 1991); but other formulations of the familial dependency have been proposed (Bonney 1992; Abel et al. 1993). In EE analyses, we chose to model the pairwise association in terms of marginal correlations, as originally proposed by Prentice (1988); but other authors have proposed use of the marginal odds ratios, which have desirable properties and are easier to interpret than the correlation coefficients (Lipsitz et al. 1991; Carey et al. 1993; Lipsitz and Fitzmaurice 1996). Second, the efficiency of regression estimates has been shown to depend on the covariate distribution and to be quite sensitive to the between- and within-cluster variation of the covariate (Mancl and Leroux 1996). Actually, in the presence of a strong clustering effect, the genotype distribution is expected to be quite different between families. Last, both the ML and EE methods assume a correct specification of the mean vector of phenotypes. Using a logistic parametrization for a binary variable obtained from a truncated liability distribution is a priori not correct, and this could also affect the efficiency. As already observed for a quantitative trait, the variance of the parameter was underestimated in small samples, and the main consequence was an inflation of the type I error.

When the cluster size was varying, we observed a slight loss of efficiency of the EE estimator, by comparison with balanced samples. Similar loss of efficiency has been reported elsewhere (Lipsitz and Fitzmaurice 1996; Mancl and Leroux 1996). The loss of efficiency seemed to depend on the degree of heterogeneity of cluster size, but this finding needs to be studied in more detail. One explanation could be that in unbalanced samples the pairwise interclass- and intraclass-correlation estimators are no longer the ML estimators and have been shown to be less efficient than these latter (Donner and Eliasziw 1991).

A final issue concerns ascertainment. In this paper, we have considered families randomly sampled, a situation more often encountered in the analysis of quantitative phenotypes (Tiret et al. 1992; Rice et al. 1994; Georges et al. 1996). By contrast, when a disease is studied, families are usually ascertained through a proband or on the basis of familial patterns of phenotypes. Except if selection is also made on marker status or if the marker interacts with other factors contributing to familial aggregation, one should expect the ascertainment not to affect the estimate of the association parameter but only the second-order moments (Liang and Beaty 1991). Not correcting for ascertainment would then yield a loss of efficiency of the working correlation matrix, which could have greater impact in small samples. The properties of the EE method in different schemes of ascertainment would need further investigation.

In conclusion, the EE approach appears to be of potential usefulness for in the study of association between candidate-gene markers and phenotype in related individuals. However, it should be kept in mind that this approach is anticonservative in small samples. In such samples, it could be wise to lower the level of significance of the test, in order to maintain an acceptable type I error, or to compute an empirical level of significance by use of Monte Carlo methods.

Acknowledgments

We wish to thank R. Rakotovao for computational assistance, F. Demenais for having provided the REGRESS program, and two anonymous reviewers for their helpful comments.

References

Abel L, Golmard J-L, Mallet A (1993) An autologistic model for the genetic analysis of familial binary data. Am J Hum Genet 53:894–907

Boerwinkle E, Chakraborty R, Sing CF (1986) The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. Ann Hum Genet 50:181–194

Bonnardeaux A, Davies E, Jeunemaitre X, Féry I, Charru A, Clauser E, Tiret L, et al (1994) Angiotensin II type 1 receptor gene polymorphsims in human essential hypertension. Hypertension 24:63–69

Bonney G (1992) Compound regressive models for family data. Hum Hered 42:28–41

Bull SB, Chapman NH, Greenwood CMT, Darlington GA (1995) Evaluation of genetic and environmental effects using GEE and APM methods. Genet Epidemiol 12:729–734

Carey V, Zeger SL, Diggle P (1993) Modelling multivariate binary data with alternating logistic regressions. Biometrika 80:517–526

Demenais FM (1991) Regressive logistic models for familial diseases: a formulation assuming an underlying liability model. Am J Hum Genet 49:773–785

Demenais F, Lathrop M (1994) REGRESS: a computer program including the regressive approach into the LINKAGE programs. Genet Epidemiol 11:291

Donner A, Eliasziw M (1991) Methodology for inferences concerning familial correlations: a review. J Clin Epidemiol 44: 449–455

Emrich LJ, Piedmonte MR (1992) On some small sample properties of generalized estimating equation estimates for multivariate dichotomous outcomes. J Stat Comput Simulations 41:19–29

Georges J, Régis-Bailly A, Salah D, Rakotovao R, Siest G, Visvikis S, Tiret L (1996) Family study of lipoprotein lipase gene polymorphisms and plasma triglyceride levels. Genet Epidemiol 132:179–192

Godambe VP (1960) An optimum property of regular maximum likelihood estimation. Ann Math Stat 31:1208–1212

——— (1991) Estimating function. Oxford University Press, Oxford

Grove JS, Zhao LP, Quiaoit F (1993) Correlation analysis of twin data with repeated measures based on generalized estimating equations. Genet Epidemiol 10:539–544

Hendricks SA, Wassell JT, Collins JW, Sedlak S (1996) Power determination for geographically clustered data using generalized estimating equations. Stat Med 15:1951–1960

Hsu L, Zhao LP (1996) Assessing familial aggregation of age at onset, by using estimating equations, with application to breast cancer. Am J Hum Genet 58:1057–1071

Huber P (1964) Robust estimation of a location parameter. Ann Math Stat 35:73–101

Jeunemaitre X, Soubrier F, Kotelevtsev YV, Lifton RP, Williams CS, Charru A, Hunt SC, et al (1992) Molecular basis of human hypertension: role of angiotensinogen. Cell 71: 1–20

Lander ES (1996) The new genomics: global views of biology. Science 274:536–539

Lee H, Stram DO (1996) Segregation analysis of continuous phenotypes by using higher sample moments. Am J Hum Genet 58:213–224

Lee H, Stram DO, Thomas DC (1993) A generalized estimating equations approach to fitting major gene models in segregation analysis of continuous phenotypes. Genet Epidemiol 10:61–74

Liang KY, Beaty TH (1991) Measuring familial aggregation by using odds-ratio regression models. Genet Epidemiol 8: 361–370

Liang KY, Pulver AE (1996) Analysis of case-control/family sampling design. Genet Epidemiol 13:253–270

Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22

Lindpaintner K, Lee M, Larson MG, Rao VS, Pfeffer MA, Ordovas JM, Schaefer EJ, et al (1996) Absence of association or genetic linkage between the angiotensin-converting-enzyme gene and left ventricular mass. N Engl J Med 334: 1023–1028

Lipsitz S, Fitzmaurice G (1996) Estimating equations for measures of association between repeated binary responses. Biometrics 52:903–912

Lipsitz SR, Laird NM, Harrington DP (1991) Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. Biometrika 78:153–160

Mancl LA, Leroux BG (1996) Efficiency of regression estimates for clustered data. Biometrics 52:500–511

Olson JM (1994*a*) Robust estimation of gene frequency and association parameters. Biometrics 50:665–674

——— (1994*b*) Some empirical properties of an all-relative-pairs linkage test. Genet Epidemiol 11:41–49

Olson JM, Wijsman EM (1993) Linkage between quantitative trait and marker loci: methods using all relative pairs. Genet Epidemiol 10:87–102

Park T (1993) A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements. Stat Med 12:1723–1732

Prentice R (1988) Correlated binary regression with covariates specific to each binary observation. Biometrics 44:1033–1048

Rice T, Province M, Pérusse L, Bouchard C, Rao DC (1994) Cross-trait familial resemblance for body fat and blood pressure: familial correlations in the Québec family study. Am J Hum Genet 55:1019–1029

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Rotnitzky A, Jewell NP (1990) Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. Biometrika 77:485–497

Stram DO, Lee H, Thomas DC (1993) Use of generalized estimating equations in segregation analysis of continuous outcomes. Genet Epidemiol 10:575–579

Tiret L, Rigat B, Visvikis S, Breda C, Corvol P, Cambien F, Soubrier F (1992) Evidence, from combined segregation and linkage analysis, that a variant of the angiotensin I–converting enzyme (ACE) gene controls plasma ACE levels. Am J Hum Genet 51:197–205

Whittemore AS, Gong G (1994) Segregation analysis of case-control data using generalized estimating equations. Biometrics 50:1073–1087

Zhao LP (1994) Segregation analysis of human pedigrees using estimating equations. Biometrika 81:197–209

Zhao LP, Grove J (1995) Identifiability of segregation parameters using estimating equations. Hum Hered 45:286–300

Zhao LP, Grove JS, Quiaoit F (1992*a*) A method for assessing patterns of familial resemblance in complex human pedigrees, with an application to the nevus-count data in Utah kindreds. Am J Hum Genet 51:178–190

Zhao LP, Prentice RL, Self SG (1992*b*) Multivariate mean parameter estimation by using a partly exponential model. J R Stat Soc [B] 54:805–811