Am. J. Hum. Genet. 64:871–885, 1999

Analysis of Affected Sib Pairs, with Covariates—With and Without Constraints

Celia M. T. Greenwood[1] and Shelley B. Bull[2]

[1]Department of Human Genetics, McGill University, Montreal; and [2]Samuel Lunenfeld Research Institute of Mount Sinai Hospital and Department of Public Health Sciences, University of Toronto, Toronto

Summary

Covariate models have previously been developed as an extension to affected-sib-pair methods in which the covariate effects are jointly estimated with the degree of excess allele sharing. These models can estimate the differences in sib-pair allele sharing that are associated with measurable environment or genes. When there are no covariates, the pattern of identical-by-descent allele sharing in affected sib pairs is expected to fall within a small triangular region of the potential parameter space, under most genetic models. By restriction of the estimated allele sharing to this triangle, improved power is obtained in tests for genetic linkage. When the affected-sib-pair model is generalized to allow for covariates that affect allele sharing, however, new constraints and new methods for the application of constraints are required. Three generalized constraint methods are proposed and evaluated by use of simulated data. The results compare the power of the different methods, with and without covariates, for a single-gene model with age-dependent onset and for quantitative and qualitative gene-environment and gene-gene interaction models. Covariates can improve the power to detect linkage and can be particularly valuable when there are qualitative gene-environment interactions. In most situations, the best strategy is to assume that there is no dominance variance and to obtain constrained estimates for covariate models under this assumption.

Received September 4, 1998; accepted for publication January 8, 1999; electronically published February 16, 1999.

Address for correspondence and reprints: Dr. Celia Greenwood, Division of Medical Genetics, Montreal General Hospital, L10-109, 1650 Cedar Avenue, Montreal, Quebec H3G 1A4, Canada. E-mail: celia@utstat.utoronto.ca

0002-9297/99/6403-0025$02.00

Introduction

Affected-sib-pair models are a popular approach for detection of genetic loci linked to a disease gene when the mode of inheritance is unknown. Methods for analysis of affected-sib-pair data generally estimate a function of the expected allele or haplotype sharing identical by descent (IBD) at a marker locus in the affected pairs. Although the genetic model is unknown for most complex diseases, there is often epidemiological evidence showing that measurable environmental factors affect disease risk, and it is plausible that the presence of such factors may change the ratio of disease penetrances and hence affect the evidence for linkage. In the classic model-based LOD-score linkage models, covariate effects can be incorporated by alteration of the disease penetrances in liability classes (Beaty 1997). In affected-sib-pair linkage studies, examination of known environmental modifiers of disease risk can help in providing an understanding of the disease etiology and can identify subpopulations in which the evidence for linkage is stronger. Khoury et al. (1987) and Yang and Khoury (1997) have examined allele sharing and relative risks in the presence of an exposure variable, by stratifying a sample of affected sib pairs by exposure status. More-general models for sib pairs have been proposed by Dawson et al. (1990) and Flanders and Khoury (1991); each developed a method for modeling a delayed age at onset and the effects of other covariates in mixtures of affected and unaffected sibs.

Explicit covariate models have also been developed as an extension to affected-sib-pair methods (Greenwood and Bull 1997; Greenwood 1998), in which the covariate effects are jointly estimated with the degree of excess allele sharing. Although, for categorical covariates, the results will be similar to those obtained by estimation of the allele sharing in each subgroup, these general models enjoy the benefit of being able to include continuous covariates and more than one covariate at once. Covariate models can also be used to evaluate different definitions of phenotype or severity and to assess how the evidence for linkage changes with the phenotype.

The expected proportions of sib pairs sharing zero,

one, or two alleles IBD at a genetic locus have been expressed in terms of genetic-model parameters and population prevalences for a single gene (Suarez et al. 1978; Motro and Thomson 1985). Louis et al. (1987) demonstrated that these expected proportions will fall within a subregion of the potential parameter space defined by $z_1 \leqslant .5$, and $z_1 \geqslant 2z_0$, where z_j is the expected proportion of sib pairs sharing j alleles IBD. In addition, Risch (1990) showed that, for many multigene models without strong epistasis, the expected allele-sharing pattern for affected sib pairs falls within the same bounded triangle. Holmans (1993) showed that constraining the estimated allele sharing to this "possible triangle" could increase the power to detect linkage. Craddock et al. (1995) delineated the constraints appropriate for oligogenic models. Louis et al. (1987) and Whittemore and Tu (1998) examined the case of three affected sibs and developed an appropriate constraint region for this case.

In the presence of environmental effects, however, the allele-sharing estimates will not necessarily fall within the triangle applicable for simple genetic models. We first explore potential patterns of allele sharing for one gene and one covariate. Subsequently, the affected-sib-pair model with covariates (Greenwood and Bull 1997) is briefly reviewed, and three approaches for the fitting of constrained models in this context are proposed. Appropriate constrained models have not been presented previously for affected- sib-pair models with covariates, and such constraints can lead to tests for linkage that have better performance. The power of the models with covariates and of the various methods for constraining these models are evaluated by use of data simulated under several different models of gene-environment and gene-gene interaction.

Allele-Sharing Patterns with One Gene and One Covariate

Let A denote that an individual is affected with a disease, and assume that one binary covariate x and one gene g affect the risk of disease. Under the assumption that g is biallelic, with a high-risk and a low-risk allele, let μ denote the probability of disease for unexposed individuals carrying two copies of the low-risk allele, and let γ be the risk difference associated with exposure for such individuals. Let ξ_j, $j = 1,2$ denote the risk difference associated with one or two copies of the high-risk allele when the individual is unexposed ($x = 0$), and let δ_j, $j = 1,2$ represent the gene's differential impact among the exposed ($x = 1$). Define $\xi_0 = \delta_0 = 0$. A linear model for an individual's disease risk, conditional on his or her gene and covariate and including a gene-environment interaction, can then be written as

$$P(A|g,x) = \mu + \xi_j + \gamma x + \delta_j x \ . \quad (1)$$

This model assumes additivity of the genetic and environmental factors for disease risk. Alternatively, a logistic function (Lio and Morton 1997) for the probability of being affected could be used and would lead to probabilities that always fall within the bounds of zero and one, for any values of the parameters.

Assume that there is a specific chromosomal location being examined, very close to a disease-susceptibility locus, so that the recombination fraction (θ) is 0. For an affected sib pair (individuals A_1,A_2) with known covariate values $X = (x_1,x_2)$, let the IBD status at this location (measured by data from one or more markers) be denoted by IBD = k, $k = 0,1,2$. The expected allele sharing, $z_k(X)$, $k = 0,1,2$, at this location can be expressed as a function of the covariates. The dependence on X implies that the observed allele sharing in the sample of affected sib pairs can be expected to vary with the pair's covariate values, if the covariates affect the disease probability. For example, suppose that a sample of affected sib pairs with breast cancer was collected and that the sample was divided two groups: pairs in which both sibs had onset at age <45 years and pairs in which both sibs had onset at age >75 years. Since the pairs with the earlier onset are more likely to be carrying a mutation at BRCA1 or BRCA2, allele sharing (near one of these loci) in the earlier-onset group would be expected to deviate further from the null values. In fact, it might be reasonable to expect a continuum for the expected allele-sharing values, which approaches the null-hypothesis allele-sharing values as the sib pair's mean age at onset increases.

The development of expressions for the expected allele sharing resembles the approach of Suarez et al. (1978). By Bayes's rule,

$$\begin{aligned} z_k(X) &= P(\text{IBD} = k|A_1,A_2,x_1,x_2) \\ &= \frac{P(A_1,A_2|x_1,x_2,\text{IBD} = k)P(\text{IBD} = k)}{P(A_1,A_2|x_1,x_2)} \ , \end{aligned} \quad (2)$$

where $P(A_1,A_2|x_1,x_2)$ is the probability that an affected sib pair with a given set of covariates will be observed. We assume that $P(\text{IBD} = k|x_1,x_2) = P(\text{IBD} = k)$, so the IBD status at an unlinked marker should not be dependent on the covariates. The probability that an affected sib pair will be observed, given the marker IBD status k, can be written as

$$\begin{aligned} &P(A_1,A_2|x_1,x_2,\text{IBD} = k) \\ &\quad = P(A_1|x_1)P(A_2|A_1,x_1,x_2,\text{IBD} = k) \\ &\quad = \sum_{g_1,g_2:k} P(A_1|g_1,x_1)P(A_2|g_2,x_2)P(g_1,g_2|\text{IBD} = k) \ , \end{aligned} \quad (3)$$

where the summation is over all values of g_1 and g_2, the genotypes of the first and second members of the pair, which are possible, given the IBD status k. Given their genotypes, the disease probabilities of the two sibs are assumed to be independent, and the penetrances are modeled by equation (1). As shown in Appendix A, explicit expressions for the expected allele sharing can be obtained for this model with one covariate and one biallelic gene, and they are functions of the parameters μ, ξ_j, γ, and δ_j.

The joint probability that an affected sib pair will be observed is a function of the covariate values of the pair. When the covariate has no effect, the expected allele sharing is exactly that specified by Suarez (1978), and, under the hypothesis that there is no disease gene linked to the studied marker, the expected proportions of pairs sharing (0,1,2) alleles IBD are (.25, .5, .25) for Mendelian segregation patterns of the marker. For a single binary covariate and a single gene, there could be as many as three different patterns of allele sharing—for sib-pair covariate values (0,0) (i.e., both sibs are unexposed), (0,1) or (1,0) (i.e., one sib is exposed, and the other is not), and (1,1) (i.e., both sibs are exposed). For a continuous covariate, the allele sharing will vary in a way that depends on the penetrance function. Note that random ascertainment of affected sib pairs will preferentially sample high-risk exposure patterns.

The kinds of effects that measurable variables can have on the allele-sharing patterns are illustrated in table 1. In model A, the exposure increases the disease penetrance so that the probability of disease is higher for exposed individuals carrying one or two high-risk alleles. The expected allele sharing is shown for pairs in which both sibs are unexposed, for pairs in which both sibs are exposed, and for pairs in which only one sib is exposed. Evidently, when both sibs are exposed and the high-risk allele has a greater effect, the deviations from the null-hypothesis allele sharing are more marked. For mixed pairs, which have one exposed sib and one unexposed sib, the pattern of allele sharing is intermediate between those for the unexposed and the exposed pairs. In model B, the exposure is necessary in order for the gene to have a deleterious effect. Pairs in which one sib is exposed, like pairs in which both sibs are unexposed, will, in this case, show absolutely no evidence for linkage. Therefore, in unselected samples of sib pairs, under this genetic model, the evidence for linkage will be greatly diluted by the contribution of the two kinds of sib pairs that will never demonstrate linkage. The third model, C, is one with no gene-environment interaction on an additive scale, although the exposure alters the risk for all genotypes. In addition, overdominance is assumed, so that the risk to heterozygotes is greater than that to the two kinds of homozygotes. Again, all three exposure patterns for the pair lead to allele-sharing estimates that fall within the boundaries of the plausible triangle.

Table 1

Expected Patterns of Allele Sharing in Sib Pairs, with One Binary Covariate and One Gene, Shown as a Function of Exposure-Specific Penetrances f_j

Model and Exposure	q	f_0	f_1	f_2	z_0	z_1	z_2
A:							
Neither sib	.05	.05	.10	.30	.24	.50	.27
Both sibs		.05	.20	.60	.19	.50	.31
One sib					.22	.50	.28
B:							
Neither sib	.05	.05	.05	.05	.25	.50	.25
Both sibs		.05	.20	.60	.19	.50	.31
One sib					.25	.50	.25
C:							
Neither sib	.01	.05	.40	.05	.15	.49	.37
Both sibs		.45	.80	.45	.23	.50	.27
One sib					.20	.49	.30
D:							
Neither sib	.20	.424	.0424	.0042	.21	.50	.30
Both sibs		.10	1.00	1.00	.17	.49	.34
One sib					.37	.51	.12

NOTE.—The population frequency of the exposure does not affect the results shown in this table but will, of course, affect the estimated allele-sharing proportions in a sample of sib pairs with various exposures.

In all the models described so far (i.e., A–C), deviations from the null-hypothesis allele-sharing values are in the direction of excess allele sharing, so that >25% of affected sib pairs are expected to share two alleles IBD, and the expected sib-pair allele sharing falls within the possible triangle designated by Louis et al. (1987) and Holmans (1993). In fact, these possible-triangle boundaries will always hold in the presence of quantitative gene-environment interactions in which the size of a genetic effect may be modified by the environment, but the direction of the effect remains the same. However, the last model in table 1, model D, shows a strong gene-environment interaction, in which the gene is protective in unexposed individuals but confers risk in the exposed individuals. Although both the allele-sharing patterns for unexposed pairs and those for exposed pairs show excess allele sharing, the mixed pair demonstrates less allele sharing than is expected under the null hypothesis. Therefore, allele-sharing patterns outside the possible triangle can occur when (*a*) an exposure changes the direction of effect of the disease gene and (*b*) the sample contains pairs in which the two sibs have different exposures.

Note that any heterogeneity model that assigns whole families to different risk groups will lead to allele-sharing estimates that are within the possible triangle, because two sibs will never differ in their risk grouping. Note also that, for all examples in the table, the expected

proportion sharing one allele IBD is close to .5. Risch (1990) and Holmans (1993) have also noted that the dominance variance component is small in most plausible genetic models.

Affected-Sib-Pair Linkage Models with Covariates

If z_j, $j = 0,1,2$, denotes the expected allele-sharing proportions for affected sib pairs, a test for linkage can be obtained by taking the log ratio of (*a*) the likelihood of the data when the expected allele-sharing proportions z_j have been estimated divided by (*b*) the likelihood under the null-hypothesis allele-sharing values of (.25, .5, .25),

$$\text{LOD}(\hat{z}_0,\hat{z}_1,\hat{z}_2) = \log_{10}\left[\frac{L(\hat{z}_0,\hat{z}_1,\hat{z}_2)}{L(.25,.5,.25)}\right]$$

(Risch 1990). Multiplied by 2ln(10), this "LOD score" has an expected χ^2_2 distribution under no linkage, provided that no constraints have been applied to the allele-sharing estimates.

An extended specification of allele sharing, one that allows the inclusion of covariates, leads to an extension of this 2-df test for linkage (Greenwood and Bull 1997). Let $z_j(x)$, $j = 0,1,2$, be the generalized expected IBD allele sharing as a function of some covariates x. Let x_i be a vector with $P + 1$ rows for P covariates from pair i and with an intercept. A multinomial logistic model for the allele-sharing proportions,

$$z_j(x_i) = \frac{\exp(\beta_j' x_i)}{1 + \exp(\beta_0' x_i) + \exp(\beta_1' x_i)} \ ,$$

for $j = 0,1,2$ and where $\beta_2 = 0$, constrains the allele sharing z_j to add to 1 for any x_i.

Let m_i denote relevant marker data for the ith affected sib pair. Define $\pi_{ij} = P(\text{IBD}_i = j|m_i)$ to be the probability of j alleles IBD for pair i at a particular location, given the marker data, and let $\alpha_j = P(\text{IBD}_i = j)$ be the probability of j alleles being IBD under the null hypothesis of no linkage. Conditional on the sampling of only affected sib pairs and on their observed covariates, the likelihood for the marker data for the ith sib pair can be written, with use of Bayes's rule, as (Kruglyak and Lander 1995)

$$L(\beta_0,\beta_1) = \sum_{j=0}^{2} \frac{\pi_{ij} P(m_i)}{\alpha_j} z_j(x_i) \ .$$

Under the null hypothesis of no linkage, $\beta_{jp} = 0$, ($j = 0,1$, $p = 1,\ldots,P$), and $\beta_{j0} = \log(\alpha_j/\alpha_2)$, $j = 0,1$. Hence, the expected allele sharing will be the null values (.25, .5, .25) for all covariate values. Let $\rho_{ij} = \pi_{ij}/\alpha_j$. Then the likelihood under the null hypothesis is $P(m_i)$, and the LOD score can then be written as

$$\text{LOD}(\beta_0,\beta_1) = \sum_{i=1}^{n} \log_{10}\left[\sum_{j=0}^{2} \rho_{ij} z_j(x_i)\right] \ ,$$

for n affected sib pairs. When there is no linkage, 2ln(10) LOD will be approximately χ^2 with $2P + 2$ df. This is a two-tailed test that detects any departures from the null-hypothesis allele-sharing values.

The probabilities π_{ij} can be estimated at any specific location, from data on a set of markers, by a multipoint method based on the work of Lander and Green (1987) and implemented in Mapmaker/Sibs (Kruglyak and Lander 1995). Estimation of the remaining parameters is performed by an extended E-M algorithm. In the E step, the expected allele-sharing proportions for pair i, $z_{ij} = P(\text{IBD}_i = j|m_i)$, are calculated for each pair, as

$$z_{ij} = \frac{\rho_{ij} z_j(x_i)}{\sum_j \rho_{ij} z_j(x_i)} \ .$$

In the M step, updated estimates of β_j are obtained from a multinomial logistic regression, with IBD status as the outcome. The expected values z_{ij} are used as relative-frequency weights that correspond to the contribution of each sib-pair i to each possible IBD outcome $j = 0,1,2$. When there are more than two affected sibs in a family, an adjustment can be performed that equalizes the contribution of each family. For example, each pair could be treated as equivalent to $2/N_i$ independent observations, where N_i is the number of affected sibs (Suarez and Van Eerdewegh 1984). Sib pair i would then contribute to IBD outcome j, with weight $2z_{ij}/N_i$. We have not used this or any other weighting function for multiple sib pairs, because under the null hypothesis such weights lead to a biased distribution for likelihood-ratio tests (Abel and Müller-Myhsok 1998; Greenwood and Bull 1999).

Constraints on the Allele-Sharing Estimates in Covariate Models

By evaluation of the expressions in Appendix A, in the absence of gene-environment interaction (i.e., when $\delta_1 = \delta_2 = 0$), it can be seen that the allele-sharing estimates are bounded by $z_1 = .5$ and $z_1 = 2z_0$ (fig. 1) (Louis et al. 1987; Risch 1990; Holmans 1993). Suarez et al. (1978) gave expressions for allele sharing that were based on the additive and dominant components of genetic variance, and, under the assumption that these variances must be nonnegative, these two boundaries also follow. Assuming that $z_1 = .5$ corresponds to assuming

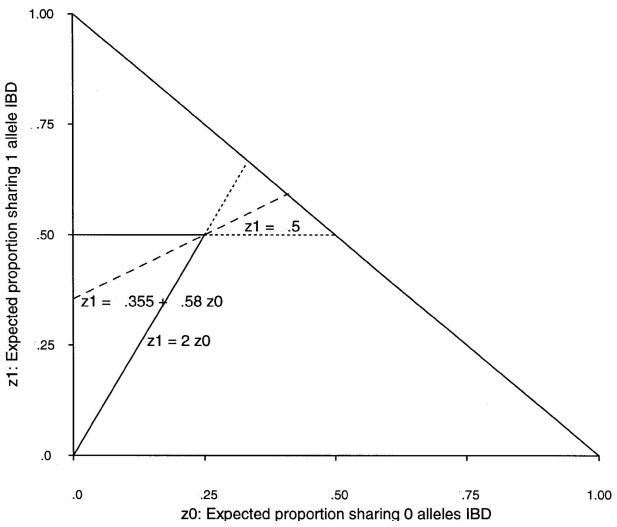

Figure 1 The parameter space for allele sharing in affected sib pair models. the possible triangle of Holmans (1993) is bounded by unbroken lines. The horizontal line at $z_1 = .5$ corresponds to the assumption of no dominance variance. The diagonal line at $z_1 = 2z_0$ corresponds to no additive variance. The oblique dashed line corresponds to the minmax-optimal test of Whittemore and Tu (1998).

that there is no dominance variance, and assuming that $z_1 = 2z_0$ is equivalent to assuming that there is no additive variance (when $\theta = 0$). Holmans (1993) showed that an estimation method that constrains the z_i's to lie within this possible triangle leads to increased power to detect linkage, compared with a general model that left the estimated allele-sharing proportions unconstrained. Whittemore and Tu (1998) showed that each of these boundaries corresponds to an efficient score test under a particular genetic model: the means test, which is based on $z_2 + 0.5z_1$, is optimal under an additive genetic model (i.e., along the line where $z_1 = .5$). Similarly, the proportions test, which is based on z_2, is optimal when there is a large dominance variance component, as is seen in rare recessive traits.

The arguments that Risch (1990) and Holmans (1993) made in favor of constraints are based on the genetic model at a population level—that is, the risk to sibs in the population, or the genetic variance in the population. It follows from this line of argument that constraints in models with covariates should apply to the population from which families are selected. For example, when a covariate such as an environmental factor is considered, especially if there is a gene-environment interaction, it may be more reasonable to apply the constraint to the whole sample, since the assumptions about the expected variances in the population might not apply in a particular exposure subgroup. On the other hand, if a covariate indicates membership in ethnic groups, then the constraints could be applied within each ethnic group separately, since different genes may confer risk in different groups, and the ethnic groups could be considered different populations.

Therefore, new approaches to constrained estimation are required for models with covariates, and the power of such approaches requires investigation. Three approaches are proposed for investigation.

1. *Average Constraints*

The population-based arguments for constraints can be interpreted to mean that the expected value of the constrained allele-sharing estimates must fall within the plausible region, where expectation is taken over the covariate distribution of the population of affected sib pairs. The average-constraint method is proposed to satisfy this requirement. However, since the covariate distribution is usually unknown, a feasible way to apply such a constraint is to sum over the observed covariate distribution in the sample of affected sibs. Therefore, the two bounds on the allele-sharing estimates become $\frac{1}{n}\Sigma_{i=1}^{n} z_1(x_i) = 0.5$ and $2\Sigma_{i=1}^{n} z_0(x_i) = \Sigma_{i=1}^{n} z_1(x_i)$. Lagrange multipliers are added to the LOD-score equations, to estimate the constrained IBD allele-sharing values. For example, the average-constrained estimates for no dominance variance would be obtained by maximization of

$$\mathrm{LOD}^*(\beta_0,\beta_1) = \sum_{i=1}^{n} \log_{10}\left[\sum_{j=0}^{2} \rho_{ij} z_j(x_i)\right] + \lambda \sum_{i=1}^{n} [z_1(x_i) - 0.5] \ .$$

The algorithm of Holmans (1993) can be adapted to decide when the average constraints should be applied, by examining the mean allele sharing over all the sib pairs, $\Sigma_i z_j(x_i)/n$, for $j = 0,1$, and then forcing these means to lie either on one of the boundaries or at the null hypothesis, by use of Holmans's sequential method for the application of constraints. The df of the test for linkage would be reduced by 1 if one average constraint is used and would be reduced by 2 if the mean allele sharing is constrained to the null hypothesis. Note that, with covariate models, the mean could be constrained to the null hypothesis, although, for any specific covariate value, the allele-sharing estimates might not be (.25, .5, .25). Average constraints can be applied to a model containing any number of covariates of any type.

2. *Subgroup-Triangle Constraints*

If all covariates are categorical, then the usual constraints for models without covariates can be applied in each of the subgroups defined by the covariates, and the LOD scores can then be summed across the subgroups.

Suppose that there are S subgroups. There would be $2S$ df associated with unconstrained allele sharing in a model with covariates (including all appropriate interaction terms), and in a constrained model the df would be between 0 and $2S$. Subgroup constraints can be applied to models with continuous covariates if the continuous covariate is categorized into a small number of disjoint groups.

3. *Simultaneous-Boundary Constraints*

A third approach can be thought of as constraining the allele sharing to one of the boundaries, for each value of the covariates. In essence, this corresponds to a covariate model that has only $P + 1$ df, instead of $2P + 2$ df. In figure 1, no dominance variance corresponds to the horizontal line at $z_1 = .5$. A boundary constraint can be defined that forces $z_1(x_i) = .5$ always, and the binomial "log-likelihood" used for estimation of the covariate effects (inside the E-M algorithm) becomes

$$\mathrm{LOD}^{**}(\beta_0) = \sum_{i=1}^{n} \{z_{i0}\log[z_0(x_i)] + z_{i2}\log[0.5 - z_0(x_i)]\} ,$$

where

$$z_0(x_i) = \frac{\exp(\beta_0' x_i)}{2 + 2\exp(\beta_0' x_i)}$$

and $z_1(x_i)$ is fixed at .5. The E step remains unchanged.

Another simultaneous-boundary constraint can be defined by assuming that there is no additive variance; this model will force all allele-sharing estimates to fall on the line described by $z_1(x_i) = 2z_0(x_i)$ (for details, see Appendix B). Furthermore, a third simultaneous-boundary constraint can be defined by use of the minmax-optimal test of Whittemore and Tu (1998), where $z_1(x_i) = 0.335 + 0.58z_0(x_0)$ (Appendix B; see the dashed line in fig. 1). Although this line falls within the possible triangle and not on an edge, the concept and the estimation approach are very similar to those for the other two simultaneous-boundary constraints, in that, for any value of x_i, the allele-sharing estimates are forced to fall on this line. These simultaneous-boundary approaches can be used for any number or combination of covariates. Unlike the other two constraint methods, no decision is made, during the estimation process, about which boundary is appropriate. A single boundary constraint is chosen in advance and is then applied to all the data and for all covariates.

Although the unconstrained LOD scores, when multiplied by $2\ln(10)$, have an asymptotic χ^2 distribution under the null hypothesis, the constrained test statistics may not be χ^2 distributed, because the allele-sharing estimates have different df values, depending on the constraints applied. In particular, the estimates obtained by either the average-constraint method or the subgroup-constraint method will not have a χ^2 distribution. Holmans (1993) discussed the distribution of affected-sib-pair tests in constrained models without covariates and showed that the distribution (which is a mixture of χ^2 distributions with 1 and 2 df) depends on the probability that the estimates will fall into different regions of figure 1. In models with covariates, the theoretical distribution will depend on the unknown distributions of the covariates. Therefore, we recommend the use of Monte Carlo (simulated) P values to assess the significance of linkage tests under these constrained models.

The simultaneous boundary–constraint methods, however, will lead to a LOD-score test that does have an asymptotic χ^2 distribution. As long as the estimated allele sharing is allowed to fall anywhere on the chosen boundary, even outside the possible triangle, these tests for linkage are expected to follow a χ^2_P distribution for P covariates. For families consisting of no more than one affected sib pair, asymptotic significance values could be obtained for these models. However, since multiple affected sibs per family can lead to dependence between sib pairs, especially for incompletely informative markers (Kong et al. 1997), it may be preferable to use simulated P values for all significance testing.

When subgroups are of intrinsic interest, then the LOD scores can be examined for each subgroup separately. Within each subgroup, constraints can be applied by use of an appropriate method. For example, if the covariate is ethnicity, then Holmans's constraint method can be applied within each ethnic group. Suppose that there are two covariates, ethnicity and mean age at onset. Then, within each ethnic group, a choice can be made as to which constraint method would be appropriate for a model containing mean age at onset. Testing for linkage within each subgroup, however, will make the probability that at least one subgroup test is significant, in the absence of linkage, higher than the nominal type I error rate. In this situation, simulation methods can be used to control type I error and to obtain adjusted P values for the linkage tests.

Simulation Study

Design/Methods

Simulations were undertaken to evaluate the performance of the extended affected-sib-pair model with covariates and constraints. Nuclear families with at least two children were generated under the assumption of random mating and no segregation distortion. The number

of sibs was based on the truncated geometric distribution $P(s = k) = \delta(1 - \delta)^{2-k}$, for $k = 2,3,\ldots,9$, and δ was chosen to be .45. The family was ascertained through a proband of age 20 years. This individual's birth order in the family was randomly chosen from the numbers 1 through *s,* where *s* is the number of children in the family. Then ages for the older sibs were created by consecutive addition of 2 years to the proband's age; ages for younger sibs were calculated by consecutive subtraction of 2. This method of assignment of ages led to a minimum possible age of 4 years and to a maximum possible age of 36 years. One fully informative marker was created in the parents, and the alleles were segregated randomly to all offspring. A disease gene was assumed to be linked to this marker, with $\theta = 0$. For various models for the probability of disease (described below), the affection status was determined for all children in a family. Then the affected offspring in families with at least two affected children were retained for analysis. The models that were used to generate the data included a single-major-gene model, a quantitative gene-environment interaction, and a qualitative gene-gene interaction.

Simulated Model 1: A Single Major Gene Affecting Age at Onset, with No Environmental Factors

A single gene acting in a dominant manner was assumed to increase the risk of disease and to lower the age at onset, compared with noncarriers of the gene. The lifetime risk of disease for carriers was assumed to be .90, and that for noncarriers was assumed to be .12, and the disease allele was assumed to have a population prevalence of .003 (similar to the lifetime disease probability and allele frequency of BRCA1, in linkage studies). Given that an individual was randomly assigned to be susceptible to disease, the age at onset was generated from a normal distribution with an SD of 4 years and a mean of 23 years in carriers and 28 years in noncarriers. (These ages are reasonable, in light of the observed ages of the sibs in the generated data, and are not meant to be realistic for breast cancer data). Individuals who were younger than their generated age at onset were considered to be unaffected. For each family with at least two affected children, two continuous covariates were calculated: the mean age at onset and the maximum age at onset, of all affected children in the family. In addition, a categorical covariate was created that was 1 when the mean age at onset was $\leq$23 years and that was 0 otherwise. Fifty families with at least two affected children were ascertained for each data set.

Simulated Model 2: Quantitative Gene-Environment Interaction

Sibs were randomly "exposed," with probability .5, to an environmental agent; no intrafamilial clustering was assumed for the exposure. Within this structure, two different models for gene-environment interaction were assumed. In each case, the age at onset among susceptible individuals was assumed to be normally distributed, with a mean of 25 years and an SD of 4 years. One hundred families with at least two affected children were ascertained, and the frequency of the high-risk allele was assumed to be .2.

Model 2a.—In unexposed individuals the linked disease-susceptibility gene has no effect, but in exposed individuals the gene acts in a mainly recessive manner. The lifetime risks of disease were assumed to be .05 for noncarriers and unexposed gene carriers, .20 for exposed heterozygous carriers, and .60 for exposed homozygous carriers (equivalent to model B in table 1).

Model 2b.—The effect of the linked gene is present in both exposed and unexposed individuals, but the genetic effect is stronger among the exposed. The lifetime risks of disease were assumed to be .05 for noncarriers (either exposed or unexposed), .10 for unexposed heterozygous carriers, .20 for exposed heterozygous carriers, .30 for unexposed homozygous carriers, and .60 for exposed homozygous carriers (equivalent to model A in table 1).

For each model, two binary covariates were created. One of them indicated whether the sibs were concordantly exposed, and the other indicated whether only one sib was exposed.

Simulated Model 3: Qualitative Gene-Gene Interaction

A measured (unlinked) common gene lowers age at onset but does not affect lifetime risk, and an unknown rare (linked) gene lowers age at onset and increases the disease risk substantially. However, no individuals carrying both genes are in the sample. This might occur if, for example, the presence of both genes was lethal. The rare gene has an allele frequency of .05 and lifetime penetrances of .05, .30, and .80 for carriers of zero, one, and two copies, respectively. The common gene is assumed to have an allele frequency of .30. If a sib were randomly chosen to get the disease during his or her lifetime (with these penetrance values being assumed), then the age at onset was generated from a normal distribution with a mean of 28 years (SD of 4 years) for an individual carrying neither gene but with a mean of 24 years for an individual carrying one of the two genes. One binary covariate was created that indicated whether the sib pair was discordant or concordant for the presence of the measured common gene. One hundred families were generated.

For each simulated data set, linkage tests using the likelihood-ratio LOD scores presented above were calculated with and without covariates, under each of the possible constraint methods. To evaluate the distribution

of the linkage tests under the null hypothesis, for each model a set of 5,000 simulations was undertaken in which the disease-susceptibility locus of interest was unlinked to the marker. Percentiles of the distributions of the LOD scores from these unlinked runs were used to evaluate the performance of the tests under linkage, for both 5% type I error and 1% type I error. The power of the different constraint methods was assessed by comparison of the percentage of the linked data sets in which the tests for linkage exceeded the chosen percentiles of the null distribution. For test statistics that should have a χ^2 distribution (for independent sib pairs), the asymptotic power is also given.

Results

Tables 2–4 present the results of the simulations for the three generating models. In each case, 500 simulated data sets were created in which there was linkage to the marker locus, and 5,000 data sets without linkage were generated to evaluate the empirical power. For some estimation methods, the E-M algorithm did not converge in every case; data sets for which this occurred were then excluded from the summaries for all estimation methods. Good starting values specific to the generating model and the estimation method increased the convergence rates, but it was not feasible to alter starting values individually for each simulated data set to attempt to find a converged solution. For model 1, the age-at-onset model, <7% of the data sets had to be excluded from the summaries. For model 2a, 16% of the linked data sets were excluded; for model 2b, 2%; for model 3, almost 30%. In all cases, it was in the models with covariates that the convergence problems occurred, and it was the simultaneous boundary–constraint methods that were the most difficult. Convergence with the minmax-optimal method proved to be particularly difficult, and so the summary statistics were recalculated, with this method's results being ignored. Then, only 9% of the data sets were excluded from model 2a, and 16% were excluded from model 3. The average LOD scores and empirical powers were in close agreement, between the summaries that did or did not include the minmax-optimal results.

In all tables, the df value refers to the number of parameters estimated. For the models using the possible-triangle constraints, different numbers of parameters are estimated in different data sets, and the df value reported is the average over the simulations. The results discussed focus on the estimates obtained by use of the empirical null distribution and 5% power.

Effects of Inclusion of Covariates in Models

Table 2 shows LOD scores and empirical power estimates for the single-gene model in which age at onset

Table 2

LOD Scores and Power for Simulated Model 1, a Single-Gene Model with Earlier Age at Onset in Carriers

Constraint Method and Covariate(s)[a]	Mean LOD Score	df	5% Power[b]		1% Power[b]	
			Empirical	χ^2	Empirical	χ^2
Unconstrained models:						
No covariates	2.20	2	.70	.71	.42	.50
Age at onset ⩽23 years	3.40	4	.72	.77	.40	.57
Maximum age at onset	3.22	4	.73	.75	.42	.54
Mean age at onset	3.50	4	.77	.79	.50	.59
Boundary-constrained tests with no dominance variance:						
No covariates	1.97	1	.74	.76	.49	.60
Age at onset ⩽23 years, simultaneous-boundary constraint with no dominance variance	2.94	2	.82	.85	.56	.69
Maximum age at onset, simultaneous-boundary constraint with no dominance variance	2.79	2	.83	.84	.57	.69
Mean age at onset, simultaneous-boundary constraint with no dominance variance	3.08	2	.87	.89	.67	.73
Triangle constraint:						
No covariates	1.94	1.4	.72	NA	.46	NA
Age at onset ⩽23 years, subgroup-triangle constraint	2.93	2.3	.80	NA	.53	NA
Age at onset ⩽23 years, average-triangle constraint	2.88	3.2	.67	NA	.43	NA
Mean age at onset, average-triangle constraint	2.99	3.2	.71	NA	.49	NA
Other constraints:						
No covariates, no additive variance	1.26	1	.53	.56	.30	.33
Mean age at onset, no additive variance, simultaneous-boundary constraint	1.88	2	.58	.58	.32	.38
No covariates, minmax-optimal constraint	1.78	1	.71	.73	.45	.53
Mean age at onset, minmax-optimal simultaneous-boundary constraint	2.61	2	.82	.80	.53	.57

[a] Estimates did not converge, in at least one estimation method, for 7/500 linked simulated data sets when the binary covariate for age at onset ⩽23 years was used; these data sets were excluded from the summaries. For maximum age at onset, 21/500 linked data sets were excluded because of nonconvergence; for mean age at onset, 34/500 linked data sets were excluded.

[b] NA = not applicable.

Table 3

LOD Scores and Power for Simulated Model 2

Constraint Method and Covariates	Mean LOD Score	df	5% Power		1% Power	
			Empirical	χ^2	Empirical	χ^2
Model 2a:[a]						
Unconstrained models:						
No covariates	.99	2	.25	.28	.09	.11
Two covariates for pair exposure	2.21	6	.27	.27	.11	.13
Constrained models:						
No covariates, no dominance variance	.77	1	.32	.33	.11	.17
Two covariates, simultaneous-boundary constraint with no dominance variance	1.56	3	.36	.36	.18	.18
No covariates, triangle constraint	.87	1.2	.34	NA	.14	NA
Two covariates, average-triangle constraint	1.89	5.0	.29	NA	.11	NA
Two covariates, subgroup-triangle constraint	1.50	2.7	.36	NA	.13	NA
No covariates, minmax-optimal constraint	.76	1	.30	.33	.10	.17
Two covariates, minmax-optimal constraint	1.45	3	.32	.32	.13	.14
Model 2b:[b]						
Unconstrained models:						
No covariates	1.47	2	.42	.48	.13	.26
Two covariates for pair exposure	2.42	6	.30	.34	.12	.16
Constrained models:						
No covariates, no dominance variance	1.25	1	.55	.57	.27	.34
Two covariates, simultaneous-boundary constraint with no dominance variance	1.73	3	.42	.43	.20	.23
No covariates, triangle constraint	1.32	1.4	.56	NA	.22	NA
Two covariates, average-triangle constraint	2.05	5.2	.35	NA	.16	NA
Two covariates, subgroup-triangle constraint	1.93	3.5	.51	NA	.22	NA
No covariates, minmax-optimal constraint	1.23	1	.54	.54	.26	.33
Two covariates, minmax-optimal constraint	1.72	3	.44	.42	.20	.20

[a] Estimates did not converge for at least one estimation method, for 82/500 data sets simulated under linkage. When the minmax-optimal method was not estimated, only 45/500 data sets experienced convergence problems.

[b] Estimates did not converge for 12/500 simulated linked data sets.

is reduced in carriers. Three different covariate models were fitted to the same data—one with the mean age at onset in the affected sibs in each family, one with maximum age at onset, and one with a dichotomous variable for mean age at onset $\leqslant$23 years. The disease gene for the age-at-onset model had a large effect, and power was quite good in all models. Including either the maximum age-at-onset covariate or the dichotomous age-at-onset covariate for this model improved power nonsignificantly (2%–3%) for the unconstrained model, but the covariate for mean age at onset significantly ($P <$ .01, for comparison of the two proportions) increased the power to detect linkage. When no dominance variance was assumed, the covariate for mean age at onset again increased power more than did the other covariates. In fact, the covariate for mean age at onset was the best choice for age-at-onset coding for any constraint method. This was probably due to the model used to generate the data, in which the age at onset was assumed to be normally distributed with a downward shift in the mean for carriers of the disease-susceptibility gene. The analysis of a different generating model might require a different parameterization of the age-at-onset covariate.

Table 3 shows results for the quantitative gene-exposure interaction in simulated models 2a and 2b (these correspond, respectively, to models B and A in table 1). For a necessary exposure (model 2a in the top half of table 3), the power is low. In the section on expected allele sharing with one covariate and one gene, it was shown that, for a necessary exposure, only the sib pairs in which both individuals are exposed will have expected allele sharing that deviates from that of the null hypothesis (table 1). The simultaneous-boundary constraint, which assumes that there is no dominance variance in any of the three covariate groups, gives an empirical power of 36%, which is not significantly higher than the power of 32% for a model with no covariates and with no dominance variance assumed. The use of covariates, however, will estimate allele sharing in the three subgroups and could therefore show which subgroup is contributing most to the evidence for linkage.

In the bottom of table 3, the disease-susceptibility gene has an effect in both exposure groups, but the impact is more pronounced among the exposed individuals. Here the models with covariates have lower power than is seen in the models without covariates, when the constraint method is the same. Essentially, the gene-environment interaction is only a small quantitative one, and all groups show similar evidence for linkage. Under these

Table 4

LOD Scores and Power for Simulated Model 3, with a Covariate for Concordance at the Measured Common Gene

Constraint Method and Covariates[a]	Mean Lod Score	df	5% Power		1% Power	
			Empirical	χ^2	Empirical	χ^2
Unconstrained models:						
No covariates	1.57	2	.51	.50	.27	.27
With covariate	2.62	4	.58	.61	.34	.41
Constrained models:						
No covariates, no dominance variance	1.37	1	.64	.66	.35	.39
With covariate, simultaneous-boundary constraint with no dominance variance	2.18	2	.68	.72	.45	.50
No covariates, triangle constraint	1.42	1.4	.68	NA	.38	NA
With covariate, average-triangle constraint	2.28	3.3	.62	NA	.38	NA
With covariate, subgroup-triangle constraint	1.89	1.8	.64	NA	.38	NA
No covariates, no additive variance	.93	1	.43	.44	.23	.23
With covariate, no additive variance	1.58	2	.49	.52	.25	.30

[a] When the minmax-optimal method was not estimated, estimates did not converge for at least one of the other estimation methods, for 78/500 data sets simulated under linkage.

conditions, the tests with covariate effects and additional df generally do not perform as well as tests in models with no covariates.

In table 4, a model with a strong qualitative gene-gene interaction was used to generate the sib-pair data. Many data sets experienced convergence problems; the results in table 4 are based on a simulation in which the minmax-optimal method was not estimated. However, power estimates were very similar when data sets were excluded when the minmax-optimal method did not converge, so the data sets that were excluded do not strongly affect the power relationships. For an unconstrained model, there is an increase in power, from 51% to 58% ($P = .04$), when the covariate is included. When no dominance variance is assumed, there is a nonsignificant increase in power, from 64% to 68%, when the covariate for concordance is included. So the covariate for concordance at the measured common gene has some effect, but it does not explain a large proportion of the allele-sharing variability.

Subgroup Testing

Table 5 shows the results for linkage tests conducted in specific covariate-defined subgroups of the data. Although the apparent power to detect linkage can be higher in certain subgroups than in the overall sample, this power may be artificially inflated if all subgroups are tested individually for linkage, since the probability of a type I error increases when several groups are examined. The second to the last column of table 5, therefore, shows empirical power estimates that have been corrected for the expected inflation in type I error. From the unlinked simulations, the maximum LOD score was calculated across the subgroups defined by the covariates. Then the linked-subgroup LOD scores were compared with the empirical distribution of the maxima. For each of the simulated models, the corrected empirical power estimates are 7%–10% lower than the uncorrected estimates. Therefore, if the subgroup with the strongest linkage were known in advance, better power could be obtained by examination of only that subgroup. However, if all subgroups are examined during the search for linkage, better or equivalent power can be obtained by use of a covariate model with an appropriate constraint than by a search of all subgroups one by one (see the last column of table 5).

Effect of Constraints on Models with Covariates

In all the simulations performed here, the simultaneous boundary–constraint methods that either assumed no additive variance or used the minmax optimal–constraint line performed quite poorly and had power lower than that of the boundary-constraint methods that assumed no dominance variance. For this reason, the tables present only a few results from these two constraint methods. For illustration, the bottom sections of tables 2–4 show a few of the power estimates obtained by these methods.

In table 2, application of the triangle constraint had a negligible effect on power in the model with no covariates, probably because the gene had a strong effect, and hence the allele-sharing estimates were almost always within the possible triangle. However, the average triangle–constraint method did not perform well and had power lower than that of the corresponding unconstrained covariate model. The subgroup triangle–constraint method improved power significantly and performed very well (with the covariate for age at onset $\leq$23 years, power [at 5%] increased from 72% to 80%), but the best power was obtained by application of the simultaneous-boundary constraint of no dominance variance, which improved the power by $\geq$12%–17% (depending

Table 5

Subgroup Tests for Group with Strongest Evidence for Linkage, from Simulated Models 1–3

Constraint Method	Mean LOD Score	df	5% Subgroup Power		5% Power with Covariates
			Usual	Corrected for Multiple Testing	
Model 1, with mean age at onset ≤23 years:					
Unconstrained	2.88	2	.80	.72	.72
No dominance variance	2.65	1	.88	.80	.82
Triangle constraint	2.66	1.4	.85	.79	.80[a]
Model 2a, with exposure necessary and both sibs exposed:					
Unconstrained	1.36	2	.37	.30	.27
No dominance variance	1.12	1	.47	.38	.36
Triangle constraint	1.20	1.4	.50	.38	.36[a]
Model 3, with sibs concordant for second gene:					
Unconstrained	2.03	2	.66	.56	.58
No dominance variance	1.82	1	.77	.65	.68
Triangle constraint	1.83	1.4	.76	.64	.64[a]

The last column shows the power from an overall model with covariates and the appropriate constraint.

[a] Estimates obtained by use of subgroup-triangle constraints with covariates.

on the covariate) over that of the unconstrained model with no covariates ($P < .0001$).

The simultaneous boundary of no additive variance tended to give very poor power. This is not surprising, since, in the unconstrained model with no covariates, the allele sharing for z_1 was estimated to be .497, almost .5. It follows, therefore, that the minmax optimal–constraint method had power intermediate between those of the two other simultaneous-boundary methods.

In table 3, for model 2a, again the best power is obtained by fitting of the simultaneous-boundary method assuming no dominance variance (with covariates for pair exposure), but the subgroup triangle–constraint method does just as well. However, for model 2b, the triangle-constraint method without covariates performs as well as the model assuming no dominance variance and without covariates. As discussed above, the covariates do not help the power for this example, because they have such a small effect. However, the appropriate constraint method can dramatically improve power, compared with that of the unconstrained models.

The allele-sharing patterns estimated in simulated models 2a and 2b (shown in table 6) are very similar to the expected patterns seen, in table 1, for a single gene and a single exposure variable, despite the introduction of a range of ages and an age-at-onset distribution. For example, for simulated model 2a, when the exposure is necessary for increased risk, when only one sib is exposed, the mean allele-sharing estimates are (.25, .49, .25), which are very close to the expected, null-hypothesis values. In all subgroups, $\hat{z}_1 \approx .5$, so assuming that there is no dominance variance is a better strategy than either constraining the estimates to the minmax-optimal line or assuming that there is no additive variance. Also, the low power to detect covariate effects in simulated model 2b could have been deduced from table 1 by comparison of the expected allele-sharing values for exposed individuals versus those for unexposed individuals.

In table 4, assuming that there is no dominance variance leads to a dramatic improvement in power of the model with no constraints and no covariates (from 51% to 64%, for 5% type I error). However, the power for the triangle constraints of Holmans (1993) is even better (although not significantly better), at 68%, which is as good as the simultaneous-boundary method of no dominance variance with the covariate for concordance. In the subgroup of sib pairs that are discordant for their exposure to the second gene, the unconstrained allele-sharing estimates are $(\hat{z}_0,\hat{z}_1,\hat{z}_2) = (.35,.46,.19)$, well outside the possible triangle, although allele sharing for the concordant sib pairs falls within the triangle (.13, .49, .37). The discordant pairs are, however, relatively rare in the sample (~12% of all pairs). This was responsible for the occasional difficulties when the parameters of the covariate models were estimated, and it also means that the overall LOD scores are mainly based on allele-sharing estimates for concordant pairs. If these discordant pairs were more common in the sample, the models with covariates could be expected to have much better power than is seen for the models without covariates. In both subgroups, the estimated value of z_1 is near .5, and therefore the simultaneous-boundary constraint assuming no dominance variance is again a good choice of constraint method.

Although empirical significance levels were used to compare all the results reported here, it is worth noting again that the simultaneous-boundary methods lead to linkage tests that are asymptotically χ^2 for independent sib pairs and that, therefore, asymptotic significance levels can be obtained for these tests when there are only

Table 6

Unconstrained Allele-Sharing Estimates from the Simulations Based on Simulated Models 2a and 2b, with Quantitative Gene-Environment Interaction

Subgroup	$\hat{z}_0$	$\hat{z}_1$	$\hat{z}_2$
Simulated model 2a: exposure necessary for increased risk:			
Both sibs unexposed	.245	.481	.274
One sib exposed	.253	.493	.254
Both sibs exposed	.170	.488	.342
Simulated model 2b: exposure-increased risk:			
Both sibs unexposed	.208	.481	.311
One sib exposed	.191	.490	.319
Both sibs exposed	.172	.487	.341

NOTE.—These models correspond to models B and A, respectively, in table 1.

two affected sibs per family. In these simulations, however, there were multiple affected sibs per family. The dependence between familial sib pairs leads to the inflated power estimates seen when the χ^2 distribution is assumed (Kong et al. 1997; Abel and Müller-Myhsok 1998). The inflation of the test statistics becomes more severe in the tail of the distribution, and therefore the agreement between the empirical and asymptotic power estimates is poorer when 1% type I error is used.

Discussion

Models for affected-sib-pair analyses with covariates, with and without constraints, have been presented, and three new constraint methods have been proposed and evaluated. As in any multiple-regression model, covariates must be chosen and formulated, and this requires some knowledge of the epidemiology of the disease under study, in order to be fully effective. In general, any covariates that assign whole families into groups will always lead to either null or excess allele sharing between affected sibs, because such covariates will never differ between sibs in the same family. Models with no covariates, in this case, will often do as well as or better than covariate models, because they have fewer df. Nevertheless, in some situations, such as that in the first simulated example, covariates will improve power, and likelihood-ratio tests can be used to identify sib pairs with especially strong evidence for linkage. If there are good a priori reasons to expect the penetrances to vary with a continuous covariate, then a model that allows allele sharing to vary, in a smooth fashion, with a parameter such as age at onset (i.e., a model with no intercept) can have substantially improved power to detect linkage.

In general, using a constrained model improves the power of the analysis. The results here agree with those in Holmans's (1993) paper, which showed that a constrained model should be more powerful than one without any constraints. In fact, this is always true for models without covariates, although, in the models that we have investigated, application of the constraint of no dominance variance gave better power than was seen for the triangle constraints. For models with covariates, the effect of constraints is a function of the allele sharing in the subgroups defined by the covariates. When all subgroups of sib pairs show either excess allele sharing or null allele sharing, constraints continue to improve the power to detect linkage.

When there is more than one covariate, some combination of the three constraint methods (average, triangle subgroup, and simultaneous boundary) would be possible. For example, with one continuous covariate and one categorical covariate, the average constraint could be applied within the subgroups defined by the second covariate. Alternatively, different boundary constraints could be used if it is anticipated that a gene might act recessively in one group but additively or dominantly in another group. This might be plausible if the grouping factor were a candidate gene thought to interact with the gene under study. However, caution must be used before assumptions about the mode of inheritance of the disease within subgroups are made. Affected-sib-pair methods are attractive because little needs to be assumed about the genetic model. Even if there are some data to support a dominant or recessive mode of inheritance, it is unlikely that there are good-quality data to support different modes of inheritance in covariate subgroups.

Whittemore and Tu (1998) defined a minmax-optimal 1-df test for linkage in affected sib pairs. They obtained allele-sharing estimates that fall on a line going through both the null-hypothesis point and point (0, .355); this line of constraint is closer to the no-dominance-variance line in figure 1 than to the other, no-additive-variance boundary. We implemented this boundary constraint, and it performed adequately, but, in the simulations examined, we found that it did not perform quite as well as the boundary constraint with $z_1 = .5$. All the simulations that we examined gave unconstrained allele-sharing estimates for z_1 that are very close to .5 and that therefore had large additive variance relative to their dominance variance. Hence, for these models, the first

boundary constraint is optimal. This minmax-optimal line can be expected to perform adequately for rare recessive traits as well as in additive models, and our results confirm that the cost, in power, to achieve good properties overall is not too high. However, real genetic models with small additive variance may be rare. Lunetta and Rogus (1998) showed that, in the full-parameter space, there is only a very small region where a model with triangle constraints would have better power than is provided by a model with no dominance variance; the boundary constraint with no dominance variance tends to do better than the triangle constraints, in almost all the models examined in the present study.

Covariate models can be especially useful when there is gene-environment interaction and when some sibs differ in their environmental exposures. In this case, the allele sharing between sibs who have different exposures can fall outside the possible triangle, and the covariate models will find this heterogeneity, if such discordant pairs are sufficiently frequent. An advantage of the simultaneous-boundary constraints is that the estimates obtained are allowed to fall outside the possible triangle. Therefore, the allele sharing will be appropriately estimated even in the presence of gene-environment interactions producing unusual patterns of allele sharing.

For a real data set, simulated tests of significance for these LOD scores can be obtained by random reassignment of the marker alleles to the sibs in the family. If the parents are typed, then the random allele segregation simply involves choosing one allele to be transmitted from each parent. For untyped parents, the information in the sibship would have to be used to reconstruct parental marker data as far as possible; this partial information, together with marker-allele frequencies, could then be used to generate marker data on the sibs, conditional on the number of distinct alleles observed, so that the simulated data remain just as informative as the real data. Significance levels can be estimated by repeating this random process many times and then counting the number of times that the simulated test statistic exceeds the test statistic obtained from the real data.

The model-based LOD-score method of analysis, because it assumes a particular genetic model, implicitly leads to allele sharing that satisfies the constraints. The software Mapmaker/Sibs, by Kruglyak and Lander (1995), fits either a triangle constraint or a boundary with $z_1 = .5$ that is further constrained so that $z_0 \leq .25$, and the parameterization of Olson (1995) constrains the genetic variances to be positive. However, there is some evidence that routine application of constraints may not be ideal. Knapp (1996) found a genetic model in which a deficit of allele sharing is seen in affected sib pairs. Clerget-Darpoux and Babron (1997) found that the allele sharing in one subgroup can fall outside the triangle. They indicated that covariates, stratification, or increased death rates can lead to apparent contradictions in the allele sharing. Automatic application of subgroup constraints can obscure the fact that the allele sharing has occurred outside the triangle. The unconstrained estimates of allele sharing should be carefully examined before constrained models are used, and any unusual patterns should be noted and explored.

Availability of SAS/IML Macro

An SAS/IML macro (SAS Institute 1988) capable of estimating the constrained and unconstrained affected-sib-pair models with covariates is available from the corresponding author.

Acknowledgments

C.M.T.G was supported, during her doctoral studies at the University of Toronto, by the Health Research Personnel Development Program of the Ontario Ministry of Health and currently is supported by a postdoctoral fellowship from the Medical Research Council of Canada. S.B.B. is a National Health Research Scholar of the National Health Research Development Program. This work was also supported by a Collaborative Project Grant from the Natural Sciences and Engineering Research Council of Canada.

Appendix A

Expressions for IBD Allele Sharing, with One Covariate

For a biallelic disease-susceptibility gene, under the assumption of random mating and no sex-specific effects, there are six distinct mating types for a single gene (Suarez 1978). Let q be the frequency of the high-risk disease allele. For example, the probability of a mating when one parent carries one high-risk disease allele and the other carries none occurs with probability $4q(1 - q)^3$. Let the subscripts "1" and "2" to x and g denote the covariate and genotypic values for sib 1 and sib 2, respectively. However, let the subscripts to the parameters ξ_j and δ_j, $j = 1,2$, refer to the effect that one or two high-risk alleles have on the disease probability. After summation over the mating types, the probability that the sib-pair disease genotypes, conditional on IBD status, will be observed, $P(g_1,g_2|\text{IBD} = k)$, is shown in table A1, for a marker with $\theta = 0$. The ordered genotype (g_1,g_2) shows the number of high-risk disease alleles in sibs 1 and 2, respectively. Because the covariate values can differ between the members of the pair, the probability that a pair with ordered genotype (0,1) will be observed is distinguished from the probability that (1,0) will be ob-

served, etc. The ordering has no biological relevance, but it is necessary to distinguish both the genotype and covariate patterns within the sib pair. For example, ($g_1 = 0$, $g_2 = 1$, $x_1 = 0$, $x_2 = 1$) is distinct from ($g_1 = 0$, $g_2 = 1$, $x_1 = 1$, $x_2 = 0$).

On the basis of equation (3), the probability that an affected sib pair will be observed, given IBD status and covariates x_1,x_2, can be calculated by summation of the entries in the columns of table A1, multiplied by the two relevant disease probabilities from equation (1). Then the allele-sharing probabilities are obtained by use of equation (2). Let $D = \mu + \xi_1 + \gamma x_1 + \delta_1 x_1$, $C = \mu + \xi_1 + \gamma x_2 + \delta_1 x_2$, $B = \mu + \xi_2 + \gamma x_1 + \delta_2 x_1$, and $A = \mu + \xi_2 + \gamma x_2 + \delta_2 x_2$. In fact, D is the probability of disease for the first sib with one high-risk disease allele, and C is the probability for sib 2 with one high-risk allele. Similarly, B and A are the disease probabilities for sibs 1 and 2, respectively, with two high-risk alleles. Furthermore, define $E = \mu + \gamma x_2$ and $F = \mu + \gamma x_1$, the disease probabilities for sibs 2 and 1, respectively, with no high-risk alleles. Let

$$\begin{aligned} DEN = {} & (1-q)^4 EF + 2q(1-q)^3(FC + DE) \\ & + 2q^2(1-q)^2(AF + BE) \\ & + 4q^2(1-q)^2 CD + 2q^3(1-q)(AD + BC) \\ & + q^4 AB + 2(1-q)^3 EF \\ & + 2q(1-q)^2(CF + DE) + 4q(1-q)CD \\ & + 2q^2(1-q)(AD + BC) + 2q^3 AB \\ & + (1-q)^2 EF + q^2 AB \ . \end{aligned}$$

Hence, the expected allele sharing can be written, in terms of the penetrance-model parameters, as

$$\begin{aligned} z_0 = {} & [(1-q)^4 EF + 2q(1-q)^3(CF + DE) \\ & + q^2(1-q)^2(FA + BE) \\ & + 4q^2(1-q)^2 CD + 2q^3(1-q) \\ & \times (AD + BC) + q^4 AB]/DEN \ , \\ z_1 = {} & 2[(1-q)^3 EF + q(1-q)^2(CF + DE) \\ & + q(1-q)CD \\ & + q^2(1-q)(AD + BC) + q^3 AB]/DEN \ , \\ z_2 = {} & [(1-q)^2 EF + 2q(1-q)CD + q^2 AB]/DEN \ . \end{aligned}$$

When $\delta_1 = 0$, $\delta_2 = 0$, and $\gamma = 0$, the estimated allele sharing is equivalent to that reported by Suarez et al. (1978), although differently parameterized; in their work, $f_0 = \mu$, $f_1 = \mu + \xi_1$, and $f_2 = \mu + \xi_2$. If $\gamma > 0$ but $\delta_j = 0$, $j = 1,2$, then there is an environmental effect, but no gene-environment interaction, on the additive scale.

Table A1

$P(g_1,g_2|\text{IBD} = k)$, for Disease Gene or Nearby Marker

	P, Given IBD =		
(g_1,g_2)	2	1	0
(0,0)	$(1-q)^2$	$(1-q)^3$	$(1-q)^4$
(0,1)	0	$q(1-q)^2$	$2q(1-q)^3$
(1,0)	0	$q(1-q)^2$	$2q(1-q)^3$
(0,2)	0	0	$q^2(1-q)^2$
(2,0)	0	0	$q^2(1-q)^2$
(1,1)	$2q(1-q)$	$q(1-q)$	$4q^2(1-q)^2$
(1,2)	0	$q^2(1-q)$	$2q^3(1-q)$
(2,1)	0	$q^2(1-q)$	$2q^3(1-q)$
(2,2)	q^2	q^3	q^4

Appendix B

Simultaneous-Boundary Constraints

Under the assumption of no additive variance, the allele-sharing estimates can be constrained to fall on the line $z_1 = 2z_0$, by use of the following altered multinomial likelihood within the M step of the E-M algorithm:

$$\begin{aligned} \text{LOD}^{**}(\beta_0) = \sum_{i=1}^{n} \{ & z_{i0} \log[z_0(x_i)] \\ & + z_{i1} \log[2z_0(x_i)] + z_{i2} \log[1 - 3z_0(x_i)]\} \ , \end{aligned}$$

where

$$z_0(x_i) = \frac{\exp(\beta_0' x_i)}{3 + 3\exp(\beta_0' x_i)} \ .$$

The line formed by $z_1 = 0.355 + 0.58z_0$ corresponds to the minmax-optimal test of Whittemore and Tu (1998) and is depicted as the dashed line in figure 1. A third simultaneous-boundary constraint can be defined by constraining all allele-sharing estimates to fall on this line. The altered likelihood for the M step becomes

$$\begin{aligned} \text{LOD}^{**}(\beta_0) = \sum_{i=1}^{n} \{ & z_{i0} \log[z_0(x_i)] \\ & + z_{i1} \log[0.335 + 0.58z_0(x_i)] \\ & + z_{i2} \log[0.645 - 1.58z_0(x_i)]\} \ , \end{aligned}$$

where

$$z_0(x_i) = \frac{0.645\exp(\beta_0' x_i)}{1.58[1 + \exp(\beta_0' x_i)]} \ .$$

In both cases, the E step of the E-M algorithm remains unchanged from that in the unconstrained models.

References

Abel L, Müller-Myhsok B (1998) Robustness and power of the maximum-likelihood–binomial and maximum-likelihood–score methods, in multipoint linkage analysis of affected-sibship data. Am J Hum Genet 63:638–647

Beaty TH (1997) Evolving methods in genetic epidemiology. I. Analysis of genetic and environmental factors in family studies. Epidemiol Rev 19:14–23

Clerget-Darpoux F, Babron MC (1997) Identity by descent homogeneity test as a simultaneous test of linkage and allelic association. Paper presented at the 6th annual meeting of the International Genetic Epidemiology Society, Baltimore, October 27–28

Craddock N, Khodel V, Van Eerdewegh P, Reich T (1995) Mathematical limits of multilocus models: the genetic transmission of bipolar disorder. Am J Hum Genet 57:690–702

Dawson DV, Kaplan EB, Elston RC (1990) Extensions to sib-pair linkage tests applicable to disorders characterized by delayed onset. Genet Epidemiol 7:453–466

Flanders WD, Khoury MJ (1991) Extensions to methods of sib-pair linkage analyses. Genet Epidemiol 8:399–408

Greenwood CMT (1998) Models for variable and age-dependent penetrances in genetic linkage analysis. PhD thesis, Graduate Department of Community Health, University of Toronto, Toronto

Greenwood CMT, Bull SB (1997) Incorporation of covariates into genome scanning using sib-pair analysis in bipolar affective disorder. In: Goldin LR, Bailey-Wilson JE, Borecki IB, Falk CT, Goldstein AM, Suarez BK, MacCluer JW (eds) Genetic Analysis Workshop 10: detection of genes for complex traits. Genet Epidemiol 14:635–640

——— (1999) Down-weighting of multiple affected sib pairs leads to biased likelihood-ratio tests under no linkage. Am J Hum Genet 64 (in press)

Holmans P (1993) Asymptotic properties of affected-sib-pair linkage analysis. Am J Hum Genet 52:362–374

Khoury MJ, Stewart W, Beaty TH (1987) The effect of genetic susceptibility in causal inference in epidemiologic studies. Am J Epidemiol 126:561–567

Knapp M (1996) Even a deficit of shared marker alleles in affected sib pairs can yield evidence for linkage. Am J Hum Genet 59:485–486

Kong A, Frigge M, Bell GI, Lander ES, Daly MJ, Cox NJ (1997) Diabetes, dependence, asymptotics, selection and significance. Nat Genet 17:148

Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84: 2363–2367

Lio P, Morton NE (1997) Comparison of parametric and nonparametric methods to map oligogenes by linkage. Proc Natl Acad Sci USA 94:5344–5348

Louis EJ, Payami H, Thomson G (1987) The affected sib method. V. Testing the assumptions. Ann Hum Genet 51: 75–92

Lunetta KL, Rogus JJ (1998) Strategy for mapping minor histocompatibility genes involved in graft-versus-host disease: a novel application of discordant sib pair methodology. Genet Epidemiol 15:595–607

Motro U, Thomson G (1985) The affected sib method. I. Statistical features of the affected sib-pair method. Genetics 110:525–538

Olson JM (1995) Multipoint linkage analysis using sib pairs: an interval mapping approach for dichotomous outcomes. Am J Hum Genet 56:788–798

Risch N (1990) Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229–241

SAS/IML user's guide, release 6.03 ed. SAS Institute, Cary, NC

Suarez BK (1978) The affected sib pair IBD distribution for HLA-linked disease susceptibility genes. Tissue Antigens 12: 87–93

Suarez BK, Rice J, Reich T (1978) The generalized sib pair IBD distribution: its use in the detection of linkage. Ann Hum Genet 42:87–94

Suarez BK, Van Eerdewegh P (1984) A comparison of three affected-sib-pair scoring methods to detect HLA-linked disease susceptibility genes. Am J Med Genet 18:135–146

Whittemore AS, Tu I-P (1998) Simple, robust linkage tests for affected sibs. Am J Hum Genet 62:1228–1242

Yang Q, Khoury MJ (1997) Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. Epidemiol Rev 19:33–43