THE AFFECTED SIB METHOD. I. STATISTICAL FEATURES OF THE AFFECTED SIB-PAIR METHOD

UZI MOTRO[1] AND GLENYS THOMSON[2]

Genetics Department, Mulford Hall, University of California, Berkeley, California 94720

Manuscript received July 7, 1983
Revised copy accepted March 21, 1985

ABSTRACT

The distribution of the number of HLA haplotypes shared by sibs affected with the same HLA-linked disease can be used to obtain information on the genetics of the disease. Since the inception of the use of sib-pair methods for the analysis of the HLA-associated diseases, the question has been raised of how to include families with more than two affected sibs in the sib-pair analysis. This paper presents appropriate weighting schemes. A procedure for estimating the frequency of the disease allele in the general population, under the assumptions of single-allele recessive, additive, dominant and intermediate models, with negligible recombination ($\theta = 0$) between the disease-predisposing gene and the HLA region, and no selective disadvantage of the trait, is also given. Cluster-sampling techniques are used in the analysis.

MULTIPLE case family studies have provided a method for detecting the presence of a human leukocyte antigen (HLA)-linked disease-predisposing gene (CUDWORTH and WOODROW 1975; BOBROW *et al.* 1975). These initial studies investigated the distribution of sharing of HLA haplotypes among sibs affected with a disease. Deviations of the sharing of HLA haplotypes from random expectations were taken as evidence of the existence of an HLA-linked "disease" gene (CUDWORTH and WOODROW 1975). The use of such data has been extended by theoretical studies so that now affected sib data are used to detect the presence of HLA-linked "disease" genes, as well as to try and determine the mode of inheritance of the "disease" genes. The affected sib-pair method has been applied to data on a number of diseases, including multiple sclerosis, hemochromatosis, insulin-dependent diabetes mellitus (IDDM), celiac disease, juvenile rheumatoid arthritis, adult rheumatoid arthritis and Graves' and Hashimoto's diseases (THOMSON and BODMER 1977a,b; KIDD *et al.* 1977; SUAREZ, HODGE and REICH 1979; SPIELMAN, BAKER and ZMIJEWSKI 1980; SVEJGAARD, PLATZ and RYDER 1980; WALKER and CUDWORTH 1980; GREENBERG and ROTTER 1981; STEWART *et al.* 1981; WEITKAMP 1981; GREENBERG, HODGE and ROTTER 1982; KHAN, KUSHNER and WEITKAMP 1983).

The theory of the affected sib-pair method has been investigated by a number of workers, who have made different assumptions about the penetrance

[1] Present address: Department of Statistics, The Hebrew University, Jerusalem, Israel.
[2] To whom reprint requests should be addressed.

parameters for the genotypes at the "disease" locus and the value of the recombination fraction between the "disease" gene and the HLA loci. The different models that have been investigated include consideration of "quasirecessive" and "quasidominant" models (DAY and SIMONS 1976), strict recessive and dominant models (THOMSON and BODMER 1977a,b), a model with general penetrance values for the three "disease" genotypes (SUAREZ 1978), extension of the general model to also include recombination (SUAREZ, RICE and REICH 1978; SPIELMAN, BAKER and ZMIJEWSKI 1980; PAYAMI, THOMSON and LOUIS 1984) and models that take account of family size information (GREEN, HENG and WOODROW 1983; EWENS and CLARKE 1984).

It is an implicit assumption in all but the GREEN, HENG and WOODROW (1983) and EWENS and CLARKE (1984) approaches cited above that affected sib pairs are obtained by simple random sampling. Obviously, this is not the case in practice, since the sampling units are families and not affected sib pairs. Sib pairs from families with more than two affected sibs are not statistically independent. Also, such sibships do not have the same distribution of parental genotypes at the disease locus as sibships with only two affected offspring. They are biased in the direction that the frequency of the disease-predisposing allele is higher among their parents than it is among parents of families with only two affected sibs (RUBINSTEIN, GINSBERG-FELLNER and FALK 1981). This implies that affected sib trios, quartets, etc. cannot be excluded from the sib-pair analysis, since the use in the analysis of only families with two affected sibs would not satisfy the conditions under which the probabilities of affected sib pairs sharing two, one, or zero HLA haplotypes were derived.

In this paper we will deal with the question of how to include families with more than two affected sibs in the sib-pair analysis. The question of ascertainment bias enters into our considerations since the probabilities with which families having two, three, etc. affected sibs enter our sample are pertinent to the calculations.

A method will be developed for estimating the frequency of the disease-predisposing allele in the general population, under the assumptions of single-allele recessive, additive, dominant and intermediate models, zero recombination and no selective disadvantage of the trait. Cluster-sampling techniques will be used in our analysis.

THE DISTRIBUTION OF THE NUMBER OF SHARED HLA HAPLOTYPES AMONG AFFECTED SIB PAIRS

We assume that there is an HLA-linked gene, with alleles D and d, which is involved in predisposing individuals to disease. We consider an intermediate model (SPIELMAN, BAKER and ZMIJEWSKI 1980; SVEJGAARD, PLATZ and RYDER 1980) where the genotypes DD and Dd are susceptible to the disease. We denote the penetrance of the DD genotype by x (that is, x denotes the probability that a DD individual will be affected by the disease, and this probability is determined by environmental and possibly other genetic factors). The penetrance of the heterozygotes Dd is denoted by λx ($0 \leq \lambda \leq 1$). In this model, individuals with the genotype dd do not contract the disease.

The case $\lambda = 0$ corresponds to a strict recessive mode of inheritance for the disease susceptibility allele D, $\lambda = 1$ corresponds to a strict dominant mode of inheritance and $\lambda = 1/2$ corresponds to an additive model.

The distribution of the number of shared haplotypes among affected sibs has been derived, assuming Hardy-Weinberg equilibrium at the disease susceptibility locus, and that sib pairs are chosen at random, by SUAREZ (1978) for the general model of disease predisposition, where all three genotypes DD, Dd and dd are disease susceptible. It is assumed that the disease-predisposing gene is very tightly linked to the HLA region such that recombination is negligible ($\theta = 0$). For the intermediate model to be considered here, the probabilities that a randomly chosen affected sib pair will share two, one or zero haplotypes denoted X, Y and Z, respectively, are given in (1a)–(1c).

$$X = P(2) = \frac{(1 - 2\lambda^2)p_D + 2\lambda^2}{(1 - 2\lambda)^2 p_D^3 + 2(1 - 2\lambda)\,(1 + 2\lambda)p_D^2 + (1 + 4\lambda)p_D + 4\lambda^2} \quad (1a)$$

$$Y = P(1) = \frac{2(1 - 2\lambda)p_D^2 + 2\lambda\,(2 - \lambda)p_D + 2\lambda^2}{(1 - 2\lambda)^2 p_D^3 + 2(1 - 2\lambda)(1 + 2\lambda)p_D^2 + (1 + 4\lambda)p_D + 4\lambda^2} \quad (1b)$$

$$Z = P(0) = \frac{(1 - 2\lambda)^2 p_D^3 + 4\lambda(1 - 2\lambda)p_D^2 + 4\lambda^2 p_D}{(1 - 2\lambda)^2 p_D^3 + 2(1 - 2\lambda)(1 + 2\lambda)p_D^2 + (1 + 4\lambda)p_D + 4\lambda^2} \quad (1c)$$

where p_D is the frequency of the disease susceptibility allele D.

The mean of S, the number of shared haplotypes per affected sib pair, denoted μ ($\mu = 2X + Y$), is given by

$$E(S) = \mu = \frac{2(1 - 2\lambda)p_D^2 + 2(1 - \lambda)(1 + 3\lambda)p_D + 6\lambda^2}{(1 - 2\lambda)^2 p_D^3 + 2(1 - 2\lambda)(1 + 2\lambda)p_D^2 + (1 + 4\lambda)p_D + 4\lambda^2} \quad (2)$$

The variance of S, denoted σ^2 is given by

$$\mathrm{Var}(S) = \sigma^2 = 2A/B^2 \quad (3)$$

where

$$\begin{aligned} A = {} & (1 - 2\lambda)^3 p_D^5 + (1 - 2\lambda)^2(2 + 6\lambda - 5\lambda^2)p_D^4 \\ & + (1 - 2\lambda)(1 + 8\lambda + 15\lambda^2 - 30\lambda^3)p_D^3 \\ & + \lambda(2 + 9\lambda + 20\lambda^2 - 58\lambda^3)p_D^2 + \lambda^2(1 + 4\lambda + 16\lambda^2)p_D + 2\lambda^4 \end{aligned}$$

and

$$B = (1 - 2\lambda)^2 p_D^3 + 2(1 - 2\lambda)(1 + 2\lambda)p_D^2 + (1 + 4\lambda)p_D + 4\lambda^2$$

THE INCLUSION OF PAIRS FROM FAMILIES WITH MORE THAN TWO AFFECTED SIBS IN THE SIB PAIR ANALYSIS

The formulas for X, Y and Z, the probabilities that affected sib pairs share two, one or zero parental haplotypes, given in (1a)–(1c), are derived under the assumption that affected sib pairs are obtained by simple random sampling from the population of all available sib pairs. Unless we restrict our analysis to only families of size two in which both sibs are affected, this condition is not satisfied. In the actual collection of sib-pair data, sampling is of families who have two or more affected children. Thus, affected sib pairs are sampled by cluster sampling (COCHRAN 1977), the clusters being the families. Each cluster size is given by the number of affected sib pairs in the family (one for a family with two affected sibs, three for a family with three affected sibs, $\binom{k}{2}$ for a family with k affected sibs.) In this section we address the question of how to include pairs from families with more than two affected sibs in the sib-pair analysis.

The weighting scheme we propose derives from a moments approach. (Maximum likelihood solutions cannot be used in this case due to the nonindependence of sib pairs in families with more than two affected sibs.) Our approach in determining the appropriate weighting scheme is that pairs from families with more than two affected sibs are included in the analysis in such a way as to be equivalent to random sampling, the condition under which the X, Y and Z values in (1a)–(1c) were derived. This procedure is the only one that will give X, Y and Z values that are independent of the family size distribution and the absolute penetrance x of the DD genotype (see APPENDIX). The appropriate weighting scheme to use will depend on the ascertainment procedure by which families were obtained.

Let us first consider the case in which ascertainment of families is by *incomplete truncate selection* (CAVALLI-SFORZA and BODMER 1971, p. 853), that is, all families with at least two affected children have the same probability (regardless of family size or the number of affected sibs) of entering our sample. Thus, families with a particular number of affected children are represented in our sample in proportion to the number of families of this type in the general population. Now, families with, for example, three affected sibs contribute three sib pairs to the total population of all affected sib pairs (families with two affected sibs contribute one pair, and families with four affected contribute six pairs, etc.), under our hypothetical scheme of random sampling from the total population of all available sib pairs. Thus in the case of incomplete truncate selection, in which families are represented in the sample in proportion to their number in the general population, the equivalent procedure is that every sib pair from all families is given equal weight in the contribution to the X, Y and Z values, that is, families with three affected sibs contribute all three sib pairs etc. as they do under random sampling. For other ascertainment schemes a weighting of the sib-pair contributions must be made (see APPENDIX).

We consider a general case in which we assume that the probability of selection of a family is a function of the number of affected sibs in the family. For any two families, one with j affected and another with k affected children, we assume that the ratio of these probabilities is b_j/b_k. Denote by m_i ($i = 1, 2, 3, \ldots, n$) the cluster size of the ith family in our sample of n families. This family has m_i affected sib pairs, and each sib pair is characterized by the number of shared haplotypes (2, 1 or 0). Let u_i, v_i and w_i ($u_i + v_i + w_i = m_i$) be the number of affected sib pairs in the ith family that share 2, 1 and 0 haplotypes, respectively. Also, let $s_i = 2u_i + v_i$, that is, s_i is the sum, over all m_i affected sib pairs of the ith family, of the number of shared haplotypes.

The appropriate estimates of the proportion of affected sib pairs sharing two, one and zero haplotypes are, respectively, given by

$$\hat{X} = \frac{\sum_{i=1}^{n} \frac{u_i}{b_i}}{\sum_{i=1}^{n} \frac{m_i}{b_i}}, \quad \hat{Y} = \frac{\sum_{i=1}^{n} \frac{v_i}{b_i}}{\sum_{i=1}^{n} \frac{m_i}{b_i}}, \quad \hat{Z} = \frac{\sum_{i=1}^{n} \frac{w_i}{b_i}}{\sum_{i=1}^{n} \frac{m_i}{b_i}} \tag{4}$$

and the corresponding estimate of μ ($= 2X + Y$), the number of shared haplotypes per sib pair is

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \frac{s_i}{b_i}}{\sum_{i=1}^{n} \frac{m_i}{b_i}} \tag{5}$$

$(s_i = 2u_{,i} + v_i)$ and

$$\begin{aligned} \widehat{\text{Var}}(\hat{\mu}) &= \frac{n}{(n-1)\left(\sum_{i=1}^{n} \frac{m_i}{b_i}\right)^2} \sum_{i=1}^{n} \left(\frac{s_i}{b_i} - \hat{\mu}\frac{m_i}{b_i}\right)^2 \\ &= \frac{n}{(n-1)\left(\sum_{i=1}^{n} \frac{m_i}{b_i}\right)^2} \left[\sum \left(\frac{s_i}{b_i}\right)^2 - 2\hat{\mu} \sum \frac{s_i m_i}{b_i^2} + \hat{\mu}^2 \sum \left(\frac{m_i}{b_i}\right)^2\right] \end{aligned} \tag{6}$$

(COCHRAN 1977).

Remarks

1. The weights given to the u_i, v_i, w_i, m_i and s_i values compensate for the over- or underrepresentation in our sample of families with different numbers of affected children. For incomplete truncate selection $b_i = 1$, $i = 1, \ldots, n$, and all families contribute with equal weight all their sib pairs to the analysis, as discussed above. If families are ascertained with probabilities directly proportional to the number of affected sibs in the family (*incomplete single selection*, CAVALLSI-FORZA and BODMER 1971, p. 853), then $b_j/b_k = j/k$, and a family with, say, four affected sibs is twice as likely to be sampled than is a family with only two affected children. To overcome this bias, we give the data from the former family only half the weight of that given to the data from the latter family.

2. (Missing observations): If, for a cluster of size m, data is available for only m' ($m' < m$) of its elements, the sum of observations for the whole cluster can be estimated by the sum over the m' elements, multiplied by m/m'. Thus, if, for a family with m affected sib pairs, the number of shared haplotypes can be obtained only for m' of the sib pairs, then s (the sum of the number of shared haplotypes) can be estimated by $(m/m')s'$, where s' is the sum over the m' affected sib pairs for which data is available, etc.

If the appropriate weighting for a given ascertainment scheme as given above is not followed, then the $\hat{X}$, $\hat{Y}$ and $\hat{Z}$ values obtained *do not* have the expectations given in (1a)–(1c). They then have expectations that are complicated functions of family size and number of unaffected children. The strength of the approach presented here is that, if the appropriate weighting is given, then the simplicity of the X, Y and Z values, in that they are functions only of the disease allele frequency p_D and the ratio λ of the penetrance of disease heterozygotes to homozygotes is maintained.

ESTIMATING THE FREQUENCY OF THE DISEASE ALLELE IN THE GENERAL POPULATION

When the appropriately weighted estimates $\hat{X}$, $\hat{Y}$ and $\hat{Z}$ have been obtained for data from a patient population, these can be used to obtain estimates of the disease-predisposing allele frequency, p_D, for specified modes of inheritance, that is, specified λ values, where λ [see (1a)–(1c)] denotes the relative penetrance of the heterozygotes *Dd* to homozygotes *DD*.

Obviously, one would initially estimate p_D values and test the observed data against expectations under additive and recessive models, as these are the two simplest alternative models to first consider in disease modeling. (The expectations for additive and dominant models are very similar, and since the theory for additive models is much simpler, consideration will be restricted to this model.) Estimates of p_D could also be obtained and tested against expectations for specified values of λ, for example, to test against previously estimated values of λ.

It is possible to obtain joint estimates using the appropriately weighted sib-pair haplotype sharing data, for both λ and p_D (LOUIS, THOMSON and PAYAMI 1983). No test of goodness of fit of these estimates can be carried out, since only two of the X, Y and Z values are independent and these have been used to estimate the two parameters.

Maximum likelihood estimates (MLEs) of p_D for recessive and additive models can be obtained under the assumption of random sampling from the population of sib pairs. However, since we must take account of the fact that sampling is of families and, hence, that our sib pairs are not all statistically independent, we must use a moments method (MOOD and GRAYBILL 1963) and appropriate cluster-sampling techniques (COCHRAN 1977).

For a specified value of λ, if $\hat{\mu}$, an estimate for μ, the mean number of shared haplotypes per affected sib pair, is substituted in (2), this yields a cubic equation in p_D (for $\lambda \neq 0$ or ½).

$$\hat{\mu} = \frac{2(1 - 2\lambda)\hat{p}_D^2 + 2(1 - \lambda)(1 + 3\lambda)\hat{p}_D + 6\lambda^2}{(1 - 2\lambda)^2\hat{p}_D^3 + 2(1 - 2\lambda)(1 + 2\lambda)\hat{p}_D^2 + (1 + 4\lambda)\hat{p}_D + 4\lambda^2} \tag{7}$$

$\hat{\mu}$ is estimated by the sample mean number of shared haplotypes per affected sib pair using the moments method [see (5)].

If we use first order approximations from a Taylor expansion, the approximate variance of $\hat{p}_D$, which is an upper bound given the restriction $0 \leq p_D \leq 1$, is given by

$$\mathrm{Var}(\hat{p}_D) = \mathrm{Var}(\hat{\mu})/[\mu'(\hat{p}_D)]^2 \tag{8}$$

where $\mu'(\hat{p}_D)$ is the first derivative with respect to $\hat{p}_D$ of the expression $\mu(\hat{p}_D)$ given in (2). Equation (8) can be written as

$$\mathrm{Var}(\hat{p}_D) = B^4\ \mathrm{Var}(\hat{\mu})/C^2 \tag{9}$$

where B is given after (3), and

$$\begin{aligned} C = {} & 2(1 - 2\lambda)^3\hat{p}_D^4 + 4(1 - 2\lambda)^2(1 - \lambda)(1 + 3\lambda)\hat{p}_D^3 \\ & + 2(1 - 2\lambda)(1 + 4\lambda + 11\lambda^2 - 30\lambda^3)\hat{p}_D^2 + 8\lambda^2(1 - 2\lambda)(1 + 6\lambda)\hat{p}_D \\ & + 2\lambda^2(1 + 2\lambda)(6\lambda - 1) \end{aligned}$$

For a recessive mode of inheritance ($\lambda = 0$), the solution of (7) is

$$\hat{p}_D = \begin{cases} \dfrac{2 - \hat{\mu}}{\hat{\mu}} & \text{if } \hat{\mu} \geq 1 \\ 1 & \text{if } \hat{\mu} < 1 \end{cases} \tag{10}$$

and an upper bound for the variance of $\hat{p}_D$ is $4\mathrm{Var}(\hat{\mu})/\mu^4$. (Under the assumption of random sampling of sib pairs the moments estimate in the recessive case is the same as the MLE.)

For the additive model ($\lambda = ½$), the solution of (7) is

$$\hat{p}_D = \begin{cases} 0 & \text{if } \hat{\mu} > 1.5 \\ \dfrac{3 - 2\hat{\mu}}{6\hat{\mu} - 5} & \text{if } 1 \leq \hat{\mu} \leq 1.5 \\ 1 & \text{if } \hat{\mu} < 1 \end{cases} \tag{11}$$

and an upper bound for $\mathrm{Var}(\hat{p}_D)$ is $(1 + 3\hat{p}_D)^4\mathrm{Var}(\hat{\mu})/4$ [from (9) using (3)]. (The MLE, for the additive model, under the assumption of random mating, is only the same as the moments estimate when $n_1 = n/2$, where n_1 is the number of sib pairs in the sample of size n sharing 1 haplotype. The MLE is $\hat{p}_D = n_0/(2n_2 - n_0)$ if $n_2 \geq n_0$, and $\hat{p}_D = 1$ if $n_2 < n_0$, where n_2 and n_0 are the numbers of sib pairs in the sample sharing 2 and 0 haplotypes, respectively.)

EXAMPLE

We consider data from the study of WALKER and CUDWORTH (1980), concerning families with at least two IDDM-affected children. (The data on families with more than two affected sibs was kindly supplied by A. G. CUDWORTH.) In total, the data set comprised $n = 135$ families, 119 of which had two affected children, 14 had three affected and two families had four affected children. We summarize these data in Table 1.

We consider three different ascertainment schemes: I, incomplete truncate selection [all families with at least two affected children have the same probability of being sampled; hence, see discussion before (4), $b_i = 1$, $i = 1, \ldots, n$); II, incomplete single selection (families are ascertained with proba-

TABLE 1

Sib pair data on IDDM

m_i	u_i	v_i	w_i	$s_i = 2u_i + v_i$	Frequency
1	1	0	0	2	69
1	0	1	0	1	43
1	0	0	1	0	7
					119
3	3	0	0	6	4
3	1	2	0	4	6
3	0	2	1	2	1
3	0	3 (1)[a]	0	3 (1)[a]	3
					14
6	6	0	0	12	1
6	2 (1)[a]	4 (2)[a]	0	8 (4)[a]	1
					2

The data are drawn from WALKER and CUDWORTH (1980) and unpublished results of A. G. CUDWORTH. The parameters m_i, u_i, v_i, and w_i, represent the number of affected sib pairs in the family sharing two, one and no haplotypes, respectively.

[a] In three families with three affected and in one family with four affected children, one of the sibs was either a recombinant or untyped. The numbers in parentheses are the available data, and the numbers to their left are the estimated values.

bilities directly proportional to the number of affected sibs in the family, thus $b_j/b_k = j/k$) and III, incomplete sib-pair selection [families are ascertained proportional to the number of affected sib pairs in the family, thus $b_j/b_k = \binom{j}{2}/\binom{k}{2}$].

Under the assumption that the mode of inheritance of the disease-predisposing allele is recessive, the following estimates of the disease allele frequency are obtained (10): I, $\hat{p}_D = 0.3308$ (SD = 0.0452); II, $\hat{p}_D = 0.3285$ (SD = 0.0436); III, $\hat{p}_D = 0.3235$ (SD = 0.0445).

In this case very similar estimates for the frequency of the disease allele (and their standard deviations), under the assumption of a recessive mode of inheritance, have been obtained for each of the three modes of ascertainment. (For all three cases the estimate $\hat{p}_D = 0.0$ is obtained for the additive case.)

Note that incomplete sib-pair selection can also be analyzed by taking one sib pair from each family. Such a procedure involves loss of information. However, this has been the standard procedure to take account of the lack of independence of sib pairs from families with more than two affected sibs.

TESTING THE MODE OF INHERITANCE OF AN HLA-LINKED DISEASE

The fact that in actual data collection the sampling units are families and not affected sib pairs leads to difficulties in the use of the χ^2 test of goodness of fit to compare observed and expected values under hypotheses about the mode of inheritance of the disease. If we include more than one sib pair from each sib trio, quartet, etc., there is a problem in that these pairs are not independent. On the other hand, since the distribution of the number of shared haplotypes is different for families having different numbers of affected children, we cannot consider using only one affected sib pair from each family. (The only case for which we could would be where ascertainment is by incomplete sib-pair selection.) Alternate statistical tests that take account of clustering and ascertainment bias in our data collecting will be discussed elsewhere (E. J. LOUIS, unpublished).

DISCUSSION

The distribution of the number of haplotypes shared by sibs affected with the same disease can be expressed in an explicit form (SUAREZ 1978); this distribution depends on the mode of inheritance of the disease, as well as on the frequency of the disease-predisposing allele. The distribution of the number of shared haplotypes among affected sib pairs in a sample can be used to estimate p_D values and test the observed data against expectation for specified λ values, which will usually be restricted to recessive and additive (dominant) models. (These estimates apply to single-allele single-locus models with the assumption of zero recombination between the predisposing gene and the HLA region and no selective disadvantage of the trait.) The sample data can also be used to give estimates of the degree of dominance λ of the disease allele and its frequency for an intermediate model. In practice this requires numerical iterations by computer (LOUIS, THOMSON and PAYAMI 1983).

This paper presents procedures for obtaining estimates for the frequency of the disease-predisposing allele, for recessive, additive and dominant models, and for intermediate models for a specified value of λ. When using affected sib-pair data for statistical inference, an important fact must be taken into consideration, namely, that in most cases affected sib pairs are not selected by simple random sampling. Usually, we select families, and, since sib pairs belonging to the same family are not statistically independent, this has to be taken into account in our data treatment. Moreover, families with different numbers of affected sibs may be ascertained with different probabilities. Since the distribution of the number of shared haplotypes for sib pairs belonging to families with only two affected children is different from sib pairs belonging to families with, say, three affected children [as has been shown empirically by WEITKAMP (1981), and theoretically by RUBINSTEIN, GINSBERG-FELLNER and FALK (1981)], appropriate compensation should be given to overcome the over- or underrepresentation of families with different sizes of affected siblings.

Detailed treatment has been given to the IDDM data of WALKER and CUDWORTH (1980). The null hypothesis of a recessive mode of inheritance was not rejected in any of the three cases assuming three different patterns of family ascertainment. The three ascertainment patterns are I, equal probability of selection for each family, independent of affected sibship size (incomplete truncate selection); II, probability of selection proportional to the number of affected sibs (incomplete single selection) and III, probability of selection proportional to the number of affected sib pairs in the family (incomplete sib-pair selection). If the mode of inheritance is recessive, the three assumed ascertainment patterns yield very similar estimates for the frequency of the disease-predisposing allele, namely, 32–33%. An additive mode of inheritance was rejected for all three ascertainment patterns.

IDDM data collected by WEITKAMP (1981) have also been analyzed. For these data a recessive mode of inheritance for the HLA-linked disease-predisposing allele is again not rejected for the three ascertainment patterns. However, in this case, the three ascertainment patterns give a wider range for the estimate of p_D, namely, 0.4136 (incomplete truncate selection), 0.3846 (incomplete single selection) and 0.3521 (incomplete sib-pair selection). For these data an additive mode of inheritance is not rejected when the ascertainment of families is assumed to be incomplete truncate selection or incomplete single selection.

Since one can usually not be certain of the extent of bias inherent in the ascertainment of families with multiple affected sibs, we suggest that sib-pair data be analyzed under all three ascertainment patterns outlined above. Provided the number of families with three or more affected sibs is sufficiently small compared to the number of families with two affected sibs, the analysis of the data under these different ascertainment patterns will not usually lead to vastly different estimates of the disease allele frequency, nor to different conclusions regarding the mode of inheritance of the HLA-linked disease-predisposing allele. This is not the case if we consider the data analyzed by WEITKAMP (1981). However, preliminary examination of an extension of Weitkamp's data set on haplotype sharing in families with three or more affected sibs indicates that, with the larger data set, the deviations in haplotype sharing in the families with three or more affected sibs, compared to that in families with two affected sibs, are not as large as the deviations found in the original data (PAYAMI *et al.* 1985).

Analysis of Caucasian haplotype sharing data for autoimmune thyroid disease gives very different results, both in terms of estimates of disease allele frequencies and tests of mode of inheritance hypotheses, for the three ascertainment schemes (H. PAYAMI, personal communication). In this case there is evidence that incomplete sib-pair selection is the appropriate ascertainment scheme.

We stress that one cannot predict *a priori* the effects of the three ascertainment procedures on the allele frequency estimates and the test of mode of inheritance. Unless there is strong evidence in favor of a particular ascertainment scheme, we suggest analyzing the data under all three ascertainment schemes, as detailed above.

An interesting feature of the affected sib-pair haplotype-sharing data from a number of diseases is that the estimates of the "disease" allele frequency are often quite high (THOMSON 1983a). For example, under a recessive hypothesis, the IDDM data set yields estimates for the frequency of the HLA-linked disease-predisposing allele that are larger than 0.3. Under an additive hypothesis, the estimated frequencies of the multiple sclerosis-predisposing allele are larger than 0.14 (data not included here).

Such disease allele frequencies are often too high to be compatible with the population prevalence of the disease, given estimated penetrance values (see for example SPIELMAN, BAKER and ZMIJEWSKI 1980; LOUIS, THOMSON and PAYAMI 1983). Consideration of an intermediate mode of inheritance, instead of a recessive model, for IDDM leads, of course, to a lower estimate of p_D. This does not necessarily imply a large decrease in the population prevalence of the disease, since heterozygous individuals will now contribute to the patient pool. The different (λ, p_D) estimates for IDDM given by LOUIS, THOMSON and PAYAMI (1983) do not lead to large differences in the predicted relative values of MZ twin concordance rates and recurrence risks in sibs and parents or children, and these are also similar to the values for the recessive case. None of the estimated parameter sets yield population rates that are compatible with the observed values. These results imply the necessity of investigating two-locus disease-predisposing models as well as more realistic single-locus models.

The strength of using affected sib-pair haplotype-sharing data to estimate disease allele frequencies and test modes of inheritance for single-allele single-locus models is that these estimates are unaffected by the presence of additional non-HLA-linked loci that predispose to disease, for which there is increasing evidence for IDDM, provided the penetrance values in the multilocus system have a multiplicative structure and there is no linkage disequilibrium between the loci (THOMSON 1981; HODGE and SPENCE 1981; LOUIS, THOMSON and PAYAMI 1983). The effect of breaking the assumptions of non-zero recombination and selective disadvantage of the trait are well understood (RISCH 1982; PAYAMI, THOMSON and LOUIS 1984). The disease allele frequency is overestimated when non-zero recombination is ignored and usually underestimated when the selective disadvantage of the trait is ignored. However, the haplotype-sharing distributions for additive and recessive models still fall on the classical additive and recessive curves, giving considerable robustness to our tests.

In the classical sib-pair analysis information on family size and the haplotype-sharing distribution of unaffected sibs are ignored. Family size information is not always available. Also, more complex disease susceptibility models which are being analyzed using a sib-pair approach, for example, multiple-allele single-locus models for IDDM and other HLA-associated disease (LOUIS, PAYAMI and THOMSON 1984), and more particularly models investigating the genetic interrelationship of the HLA-associated diseases (PAYAMI and THOMSON 1984), may not be amenable to an analysis that takes account of these factors. The data obviously should be analyzed taking account of family size information when possible (GREEN, HENG and WOODROW 1983; EWENS and CLARKE 1984), as well as the haplotype sharing of unaffected sibs and the affectional status of parents. These approaches should be seen as complementary. Discrepancies in results obtained from the different methods could indicate which assumptions of the models are incorrect.

The sib-pair analysis and the appropriate weighting schemes to account for the clustering of the data within families can also be extended to the analysis of sib-pair and sib-trio data from families with three or more affected sibs (PAYAMI *et al.* 1985). Comparison of the results from such analyses with the sib-pair results, and the results of analyses including information on the family size, as well as analysis by the AGFAP antigen genotype frequencies among patients (AGFAP) (THOMSON and BODMER 1977a,b; THOMSON 1981, 1983b; GREENBERG, HODGE and ROTTER 1982) method using population data, can give information about the validity of the assumptions of the models.

This research was supported by National Institutes of Health grant HD 12731. We thank ESTHER HUDES for her careful reading of the manuscript and helpful suggestions.

LITERATURE CITED

BOBROW, M., J. G. BODMER, W. F. BODMER, H. MCDEVITT, J. LORBER and P. SWIFT, 1975 The search for a human equivalent of the mouse T-locus-negative results from a study of HL-A types in spina bifida. Tissue Antigens **5:** 234–237.

CAVALLI-SFORZA, L. L. and W. F. BODMER, 1971 *The Genetics of Human Populations.* Freeman, San Francisco.

COCHRAN, W. G., 1977 *Sampling Techniques,* Ed. 3. John Wiley, New York.

CUDWORTH, A. G. and J. C. WOODROW, 1975 Evidence for HLA-linked genes in "juvenile" diabetes mellitus. Br. Med. J. **3:** 133–135.

DAY, N. E. and M. J. SIMONS, 1976 Disease susceptibility genes: their identification by multiple case family studies. Tissue Antigens **8:** 109–119.

EWENS, W. J. and C. P. CLARKE, 1984 Maximum likelihood estimation of genetic parameters of HLA-linked diseases using data from families of various sizes. Am. J. Hum. Genet. **36:** 858–872.

GREEN, J. R., C. L. HENG and J. C. WOODROW, 1983 Inference on inheritance of disease using repetitions of HLA haplotypes in affected siblings. Ann. Hum. Genet. **47:** 73–82.

GREENBERG, D. A., S. E. HODGE and J. I. ROTTER, 1982 Evidence for recessive and against dominant inheritance at the HLA "linked" locus in coeliac disease. Am. J. Hum. Genet. **34:** 263–277.

GREENBERG, D. A. and J. I. ROTTER, 1981 Two locus models for gluten sensitive enteropathy: population genetic considerations. Am. J. Med. Genet. **8:** 205–214.

HODGE, S. E. and M. A. Spence, 1981 Some epistatic two locus models of disease. II. The confounding of linkage and association. Am. J. Hum. Genet. **33:** 396–406.

KHAN, M. A., I. KUSHNER and L. R. WEITKAMP, 1983 Genetics of HLA associated diseases: rheumatoid arthritis. Tissue Antigens **22:** 182–185.

KIDD, K. K., D. BERNOCO, A. O. CARBONARA, V. DANEO, U. STEIGER and R. CEPELLINI, 1977 Genetic analysis of HLA-associated diseases: the "illness susceptible" gene frequency and sex ratio in ankylosing spondylitis. pp. 72–80. In: *HLA and Disease,* Edited by J. DAUSSET and A. SVEJGAARD. MUNKSGAARD, COPENHAGEN.

LOUIS, E. J., H. PAYAMI and G. THOMSON, 1984 Affected sib methods. pp. 662–663. In: *Histocompatibility Testing 1984,* Edited by E. ALBERT, M. BAUR and W. MAYR. Springer Verlag, New York.

LOUIS, E. J., G. THOMSON and H. PAYAMI, 1983 The affected sib method. II. The intermediate model. Ann. Hum. Genet. **47:** 225–243.

MOOD, A. M. and F. A. GRAYBILL, 1963 *Introduction to the Theory of Statistics.* McGraw Hill, New York.

PAYAMI, H. and G. THOMSON, 1984 Genetic interrelationship amongst HLA associated diseases using the affected sib method. pp. 663–664. In: *Histocompatibility Testing 1984,* Edited by E. ALBERT, M. BAUR and W. MAYR. Springer Verlag, New York.

PAYAMI, H., G. THOMSON and E. LOUIS, 1984 The affected sib method. III. Selection and recombination. Am. J. Hum. Genet. **36:** 352–362.

PAYAMI, H., G. THOMSON, U. MOTRO, E. J. LOUIS and E. HUDES, 1985 The affected sib method IV: sib trios. Ann. Hum. Genet. In press.

RISCH, N., 1982 Affected sib pair marker allele sharing: the effect of reduced fertility, variable family size, and a second unlinked locus (Abstr.). Am. J. Hum. Genet. **34:** 191A.

RUBINSTEIN, P., F. GINSBERG-FELLNER and C. FALK, 1981 Genetics of type I diabetes mellitus: a single, recessive predisposition gene mapping between HLA-B and GLO. Am. J. Hum. Genet. **33:** 865–882.

SPIELMAN, R. S., L. BAKER and C. M. ZMIJEWSKI, 1980 Gene dosage and susceptibility to insulin dependent diabetes. Ann. Hum. Genet. **44:** 135–150.

STEWART, G. J., J. G. MCLEOD, A. BASTEN and H. V. BASHIR, 1981 HLA family studies and multiple sclerosis: a common gene, dominantly expressed. Hum. Immunol. **3:** 13–29.

SUAREZ, B. K., 1978 The affected sib pair IBD distribution for HLA-linked disease susceptibility genes. Tissue Antigens **12:** 87–93.

SUAREZ, B., S. HODGE and T. REICH, 1979 Is juvenile diabetes determined by a single gene closely linked to HLA? *Diabetes* **28:** 527–532.

SUAREZ, B. K., J. RICE and T. REICH, 1978 The generalized sib pair IBD distribution: its use in the detection of linkage. Ann. Hum. Genet. **42:** 87–94.

SVEJEAARD, A., P. PLATZ and L. P. RYDER, 1980 Insulin-dependent diabetes mellitus. pp. 638–656. In: *Histocompatibility Testing 1980,* Edited by P. TERASKI. U.C.L.A. Tissue Typing Laboratory, Los Angeles.

THOMSON, G., 1981 A review of theoretical aspects of HLA and disease associations. Theor. Pop. Biol. **20:** 168–208.

THOMSON, G., 1983a Theoretical aspects of HLA disease associations. pp. 77–88. In: *Human Genetics, Part B: Medical Aspects,* Edited by B. BONNE-TAMIR. Alan R. Liss Inc., New York.

THOMSON, G., 1983b Investigation of the mode of inheritance of the HLA associated diseases by the antigen genotype frequencies amongst diseased method. Tissue Antigens **21:** 81–104.

THOMSON, G. and W. F. BODMER, 1977a The genetics of HLA and disease associations. pp. 545–564. In: *Measuring Selection in Natural Populations,* Edited by F. B. CHRISTIANSEN, T. FENCHEL and O. BARNDORFF-NIELSON. Springer-Verlag, New York.

THOMSON, G. and W. F. BODMER, 1977b The genetic analysis of HLA and disease associations. pp. 84–93. In: *HLA and Disease,* Edited by J. DAUSSET and A. SVEJGAARD. Munksgaard, Copenhagen.

WALKER, A and A. G. CUDWORTH, 1980 Type 1 (insulin-dependent) diabetic multiplex families. Diabetes **29:** 1036–1039.

WEITKAMP, L. R., 1981 HLA and disease: predictions for HLA haplotype sharing in families. Am. J. Hum. Genet. **33:** 776–784.

Communicating editor: W. J. EWENS

APPENDIX

Assuming random sampling of sib pairs and a single-allele single-locus recessive mode of inheritance, with zero recombination between the disease susceptibility locus and the HLA region and no selective disadvantage of the trait, the probabilities of affected sib pairs sharing two, one and zero HLA haplotypes identical by descent, denoted X, Y and Z, respectively, are derived from Table 2.

Thus,

TABLE 2

Calculation of haplotype sharing

Mating type	Frequency in population	Probability of 2 affected sibs	Probability of sharing indicated no. of haplotypes		
			2	1	0
$DD \times DD$	$\mathrm{p}_{\mathrm{D}}^4$	x^2	¼	½	¼
$DD \times Dd$	$4p_D^3p_D$	$x^2/4$	½	½	
$Dd \times Dd$	$4\mathrm{p}_D^2p_d^2$	$x^2/16$	1		

$$X = \frac{\frac{1}{4}\,x^2p_D^4 + \frac{1}{2}\,\frac{x^2}{4}\,4p_D^3p_d + \frac{x^2}{16}\,4p_D^2p_d^2}{x^2p_D^4 + \frac{x^2}{4}\,4p_D^3p_d + \frac{x^2}{16}\,4p_D^2p_d^2} \tag{A1}$$

which reduces to

$$X = \frac{1}{(1 + p_D)^2} \tag{A2}$$

We now consider family size in our calculations. For families of total size two, all of the expressions in Table 2 are appropriate. For families of size three, we must consider the probabilities of the family having two affected sibs *vs.* three affected. For matings of type $DD \times DD$, these two types of family will be represented in the general population in proportions $3x^2(1 - x)$ and x^3, respectively. These expressions appropriately involve the terms $(x/2)$ and $(x/4)$ for matings of type $DD \times Dd$ and $Dd \times Dd$, respectively. All other expressions in Table 2 are appropriate. For families of size four, and mating type $DD \times DD$, families with two, three and four affected sibs will be represented in the general population in proportions $6x^2(1 - x)^2$, $4x^3(1 - x)$ and x^4, respectively. In this case

$$\begin{aligned} X = \Bigg\{ & \frac{1}{4}\,p_D^4[\eta_2b_2x^2 + \eta_3(b_23x^2(1 - x) + \alpha_3b_3x^3) + \eta_4(b_26x^2(1 - x)^2 \\ & + \alpha_3b_34x^3(1 - x) + \alpha_4b_4x^4) + \cdots] \\ & + \frac{1}{2}\,4p_D^3p_d\left[\eta_2b_2\left(\frac{x}{2}\right)^2 + \eta_3\left(b_23\left(\frac{x}{2}\right)^2\left(1 - \frac{x}{2}\right) + \alpha_3b_3\left(\frac{x}{2}\right)^3\right) + \cdots\right] \\ & + 4p_D^2p_d^2\left[\eta_2b_2\left(\frac{x}{4}\right)^2 + \cdots\right]\Bigg\} \Bigg/ \\ \Bigg\{ & p_D^4\Big[\eta_2b_2x^2 + \eta_3(b_23x^2(1 - x) + \alpha_3b_3x^3) + \eta_4(b_26x^2(1 - x)^2 \\ & + \alpha_3b_34x^3(1 - x) + \alpha_4b_4x^4 + \cdots\cdots\Big] \\ & + 4p_D^3p_d\left[\eta_2b_2\left(\frac{x}{2}\right)^2 + \cdots\right] + 4p_D^2p_d^2\left[\eta_2b_2\left(\frac{x}{4}\right)^2 + \cdots\right]\Bigg\} \end{aligned} \tag{A3}$$

The $\eta_i(i = 2, 3, \ldots)$ represent the proportion of families of size i in the general population, the ratio of the probabilities of selection in our sample of a family with j affected *vs.* a family with k affected children is $b_j/b_k(j, k = 2, 3, \ldots)$ and $\alpha_i(i = 3, 4, \ldots)$ represent the weight we will give families with i affected sibs *vs.* families with two affected sibs.

The only weighting scheme that will allow reduction as before from (A1) to (A2) is

$$\alpha_i = \binom{i}{2}b_2/b_i$$

Since $\binom{i}{2}$ is the number of sib pairs in a family of size i, the estimates of X, Y and Z are as given in (4).