Am J Hum Genet 36:858–872, 1984

Maximum Likelihood Estimation of Genetic Parameters of HLA-Linked Diseases Using Data from Families of Various Sizes

W. J. EWENS[1,2] AND CHRISTINE P. CLARKE[2]

SUMMARY

This paper is concerned with estimating parameters associated with *HLA*-linked diseases. We consider a single disease locus closely linked to *HLA*, allowing a disease and a normal allele. The parameters to be estimated are the penetrances of the genotypes at the disease locus, the population frequency of the disease allele, and the distance of the disease locus from *HLA*. The presently used method of estimation uses *HLA*-sharing information from affected sib-pairs. The method proposed here generalizes the previous approach, using data from all sibs (affected or unaffected) in a family of any size. It allows immediate generalizations to the use of information on parental affectedness status and population prevalence.

INTRODUCTION

The basis of *HLA*-linked diseases has received considerable attention in recent years. The simplest model so far analyzed is that of one susceptibility locus linked to *HLA*, admitting one normal and one susceptibility allele. For data consisting of families of two sibs, both of whom are affected by the disease, the most frequently used theoretical approach for this model is the shared-haplotype method of Thomson and Bodmer [1], described in the next section.

A problem currently attracting much interest is how data from families of various sizes, with each family having possibly more than two affected sibs, can be analyzed. One broad strategy is to continue to focus on the shared-haplotype method, and to approximate a number of families of various sizes by an "equivalent" number of families of size two by taking pairs of affected sibs in families

Received November 3, 1983; revised February 14, 1984.

[1] Department of Biology, University of Pennsylvania, Philadelphia, PA 19104.
[2] Department of Mathematics, Monash University, Clayton, Victoria 3168, Australia.

having three of more affected (Louis et al. [2], Motro and Thomson [3]), or to reduce the data in some other way (Green et al. [4]). We argue here against this strategy, for two reasons. First, there are problems inherent in the shared-haplotype approach itself as well as inefficiencies and possible biases in reducing families with more than two affected sibs to artificial families having exactly two affected sibs. Second, a reduction of this nature is unnecessary. A relatively simple maximum likelihood estimation procedure is available specifically designed for families with various numbers of affected sibs. This procedure is described under THE MAXIMUM LIKELIHOOD APPROACH. Since maximum likelihood methods have optimality properties, do not involve the biases possibly involved in an artificial data reduction, and have other advantages described later, we advocate their use for the form of data in question.

We make two broad comments that apply to any analysis of family data used for *HLA*-linked diseases. First, since a family must usually satisfy some condition to enter a sample (e.g., have at least two affected sibs), any estimation procedure must incorporate ascertainment sampling theory as well as segregation analysis theory. Second, as noted by Louis et al. [2] and Spielman et al. [5], it is desirable to conduct a "sensitivity analysis" of any parameter estimation method. This investigates the sensitivity of any parameter estimate to assumptions about the numerical values of other parameter estimates as well as to parts of the data. We expand on both these comments later.

Our main aim in this paper is to set out the maximum likelihood estimation theory for the "one-susceptibility-locus, one-susceptibility-allele" model when the data consist of families of various sizes and have various numbers of affected sibs. We realize that the DR3/4 excess and other observations imply that the genetic basis of insulin-dependent diabetes mellitus (IDDM) is quite possibly more complex than that described by this model. Nevertheless, we apply our methods to the Cudworth data on IDDM recently published by Green et al. [4] (and recently analyzed using this model by Green et al. [4] and by Motro and Thomson [3]), so as to exhibit properties of the maximum likelihood estimation procedure using real data and to compare our estimates with those deriving from a "shared-haplotype" approach.

THE SHARED-HAPLOTYPE METHOD

We begin by describing the classical shared-*HLA*-haplotype method for the "one-susceptibility-locus-linked-to-*HLA*, one-susceptibility-allele" model, since, although, as mentioned in the INTRODUCTION, problems can arise with applying this method for families of various sizes, it serves as a natural introduction to the maximum likelihood approach.

Denote the susceptibility locus by D, with alleles D and d. We define the population frequency of D by p and assume the penetrances (i.e., probabilities of contracting the disease) of the three genotypes to be

Genotype:	DD	Dd	dd	
Penetrance:	x	λx	0	(1)

Of the three parameters, x, p, and λ, we are particularly interested in λ, and in considering the hypothesis "$\lambda = 0$," that is, D recessive to d. In terms of x, p, and λ, the population prevalence of the disease is given by the expression

$$p^2x + 2\lambda xp(1 - p) . \tag{2}$$

The susceptibility locus is assumed to be linked to *HLA* with recombination fraction R. To compare our data analysis for IDDM, described later, with those of Green et al. [4] and Motro and Thomson [3], we initially assume $R = 0$, since these authors made this assumption in their analyses. It is then easily shown (see Thomson and Bodmer [1], Suarez et al. [6]) that in families of two sibs, both of whom are affected, the respective probabilities X, Y, and Z that the two sibs share 2, 1, or 0 *HLA* haplotypes are

$$X = x^2[(\tfrac{1}{4} - \tfrac{1}{2}\lambda^2)p^2 + \tfrac{1}{2}\lambda^2 p)]/T , \tag{3a}$$

$$Y = x^2[(\tfrac{1}{2} - \lambda)p^3 + \lambda(1 - \tfrac{1}{2}\lambda)p^2 + \tfrac{1}{2}\lambda^2 p]/T , \tag{3b}$$

$$Z = x^2[(\tfrac{1}{2} - \lambda)^2p^4 + 2\lambda(\tfrac{1}{2} - \lambda)p^3 + \lambda^2p^2]/T , \tag{3c}$$

where

$$T = x^2[(\tfrac{1}{2} - \lambda)^2p^4 + 2(\tfrac{1}{4} - \lambda^2)p^3 + (\tfrac{1}{4} + \lambda)p^2 + \lambda^2 p] . \tag{3d}$$

Note that x^2 cancels in the equations for X, Y, and Z so that x cannot be estimated using only shared-haplotype frequency data from two affected sibs in a family of two. We have left the x^2 term in equations (3) for comparison with other equations developed later. Only two of the equations (3a)–(3c) are independent, the third being implied by the other two; for reasons of symmetry, we exhibit all three equations.

Suppose, in a sample of n families, all having exactly two sibs both of whom are affected, there are n_i sib-pairs sharing i *HLA* haplotypes ($i = 0, 1, 2$; $n_0 + n_1 + n_2 = n$). The essence of the shared-haplotype method is to equate X, Y, and Z to the observed frequencies and, hence, to estimate λ and p. As Louis et al. [2] and Spielman et al. [5] pointed out, an undesirable feature of this procedure is the sensitivity of the estimate of p to the observed fraction $Z = n_0/n$ of sib-pairs sharing no *HLA* haplotypes. This sensitivity of parameter estimation to the value of Z is not specific to the shared-haplotype method: a parallel sensitivity arises, as we show later, for maximum likelihood estimation. One possible problem does, however, apply specifically for the shared-haplotype method. Suppose, for example, that the families in the sample are chosen by an ascertainment scheme in which each family can be of any size but must have exactly two affected sibs. We show later that the shared-haplotype ratios above are not correct for such families, so that in view of the sensitivity of the estimate of p to the observed value of Z, possibly inaccurate estimates of p would arise by using (3a)–(3c) as the theoretical values of X, Y, and Z for data arising from these families. Although

amended *XYZ* ratios can be calculated for these families, it remains true that in the past all shared-haplotype estimation procedures have used the ratios (3) no matter what the family size.

A further difficulty with the shared-haplotype approach is that one is not able, using only X, Y, and Z, to use it to estimate the parameter x. Additional problems arise if we allow nonzero values of R. If R is nonzero, the equations for X, Y, and Z involve R (as well as λ and p), and if the data consist only of families of size two, as assumed above, the estimating equations become underidentified (i.e., one is trying to estimate three parameters from only two data points). No unique estimation of λ, p, and R is now possible unless extrinsic assumptions are made about p or λ. We will note later that problems of sensitivity arise in assuming particular values for some parameters: in particular, setting λ or R equal to 0 leads to quite different estimates of x and p than when these are allowed to be free parameters. Thus, assuming λ or $R = 0$ simply to be able to solve the *XYZ* equations typified by equations (4) below can lead to biased parameter estimates.

An additional difficulty with the shared-haplotype approach is the following. The random variables n_0 and n_1, coming from a trinomial distribution, would have, even in moderate-size samples, approximately normal distributions. However, the transformation from n_0 to n_1 to $\hat{p}$ and $\hat{\lambda}$ is nowhere near linear, so that $\hat{p}$ and $\hat{\lambda}$ cannot be expected, except in very large samples, to have approximately normal distributions. (We observe later that the joint distribution of $\hat{p}$ and $\hat{\lambda}$, in samples of reasonable size, is quite nonnormal.) Little reliance, then, can be placed on the value of standard error estimates of $\hat{p}$ and $\hat{\lambda}$, and, in particular, the "two standard deviation rule" may not be invoked. We amplify these remarks later.

A final difficulty with the shared-haplotype methodology is that by its very nature it focuses on affected sibs. However, data are now becoming available in which the *HLA* haplotypes of nonaffected sibs, as well as the affectedness status of parents, are available. It is very difficult to see how the shared-haplotype approach in its present form can handle the information provided by these data. (Although we agree that the information available in unaffected sibs could be quite small, it is shown in Spielman and Ewens [7] that use of information on affectedness statuses of parents will decrease the variances of maximum likelihood estimates of parameters, for typically occurring parameter values, to about one-sixth or less of the values when this information is not used. This significant decrease does not seem available under the shared-haplotype approach.)

For all the reasons mentioned above, we do not advocate data reduction followed by application of the shared-haplotype method for families of sizes two or more with two or more affected sibs. We turn now to the approach we do advocate, namely, maximum likelihood.

THE MAXIMUM LIKELIHOOD APPROACH

In a recent discussion of estimation procedures in ascertainment sampling, one of us (Ewens [8]) obtained the maximum likelihood equations for the estimation of genetic (and, if appropriate, ascertainment) parameters in various ascertainment

sampling schemes, in particular, in those where no constraint is made on the family size. We now write down these equations, with a revised notation adapted to conform with shared-*HLA*-haplotype data, with computer programming requirements, and also, as far as possible, with that used by Green et al. [4]. In ascertainment schemes, a family must satisfy some requirement to be a potential member of the sample, and in the case of Cudworth's data on IDDM, it is that at least two sibs in the family be affected by IDDM. We thus focus here on the theory appropriate for this requirement: the corresponding theory in other cases will parallel closely that given here. We let $P(m)$ be the probability that a family having m sibs has at least two affected: this probability is a function of p, x, λ, R, and ascertainment parameters. For families of size m, we further define a family to be of "type" mki if the family contains k affected sibs whose *HLA*-sharing characteristics are of type i, defined in more detail below ($m = 2, 3, 4 \ldots; k = 2, 3, \ldots, m; i = 1, 2, 3 \ldots$). Suppose in the data that there are $n(m)$ families of size m and that of these $n(m, k, i)$ are of size m and "type" mki. Let $P(m, k, i)$ be the probability that a family of size m and type mki enter the sample; as with $P(m)$, this is a function of p, x, λ, R, and ascertainment parameters. The maximum likelihood approach reduces to maximization, as a function of p, x, λ, R, and ascertainment parameters, of the function

$$\sum_m \sum_k \sum_i n(m, k, i) \log P(m, k, i) - \sum_m n(m) \log P(m) \tag{4a}$$

$$= \sum_m \sum_k \sum_i n(m, k, i) \log [P(m, k, i)/P(m)] \; . \tag{4b}$$

Green et al. [4] produce estimation equations that derive from conditioning on the observed numbers of m and k for each family, a procedure with which we do not agree, since maximum likelihood theory (see [8]) shows that correct conditioning is as given in equations (4) above.

At this point, we must take up the question of the ascertainment model to be used. Various models are possible, and the choice of the most appropriate model, for any given data set, is unfortunately seldom clear. (For a thorough discussion of this much debated point, see Stene [9].) One widely adopted approach (see, for example, Motro and Thomson [3]) is to consider several ascertainment schemes and compare the estimates arising from each. For the moment, we assume the "complete ascertainment" model, under which it is assumed that there is a fixed (unknown) probability α that a family of any size, having two or more affected sibs, enters the sample. Under this assumption, the parameter α drops out of the expressions (4) and thus cannot be estimated, and for the purposes of estimating p, x, λ, and R, we can put $\alpha = 1$. We do this from now on, and will consider in a later section the effects of choosing other ascertainment models.

To use the functions (4), it is necessary to compute the $P(m, k, i)$ and the $P(m)$ values. The latter are comparatively easy to calculate since they do not involve HLA status. We have five parental mating types that can produce affected offspring, and we enumerate these as follows (PMT stands for parental mating type):

Mating type:	$DD \times DD$	$DD \times Dd$	$Dd \times Dd$	$DD \times dd$	$Dd \times dd$
Enumeration: j =	1	2	3	4	5
Frequency: PMT(j) =	p^4	$4p^3(1-p)$	$4p^2(1-p)^2$	$2p^2(1-p)^2$	$4p(1-p)^3$.

(5)

Given parental mating type j, $j = 1, \ldots, 5$, the number of nonaffected offspring has a binomial distribution with paremeters shown (PNA stands for probability of not being affected):

$$\text{PNA}(1) = 1 - x,\ \text{PNA}(2) = 1 - \tfrac{1}{2}(1 + \lambda)x,\ \text{PNA}(3) = 1 - \tfrac{1}{4}x - \tfrac{1}{2}\lambda x\ ,$$

$$\text{PNA}(4) = 1 - \lambda x,\ \text{PNA}(5) = 1 - \tfrac{1}{2}\lambda x\ . \tag{6}$$

The probability that, in a family of size m, there are at least two affected is then

$$P(m) = 1 - (1-p)^4 - \sum_{j=1}^{5} \text{PMT}(j)[\text{PNA}(j)]^m - m \sum_{j=1}^{5} \text{PMT}(j)[\text{PNA}(j)]^{m-1}[1 - \text{PNA}(j)]\ . \tag{7}$$

The form (7) for $P(m)$ is the most suitable for computer calculation.

The probabilities $P(m, k, i)$ are evaluated by an extension of the argument that led to equation (7). The probability that a family of size m has k affected, (and, hence, $m - k$ nonaffected), and that the HLA configuration of the k affected sibs is of type i, is of the form

$$P(m, k, i) = \binom{m}{k} \sum_{j=1}^{5} \text{PMT}(j)[\text{PNA}(j)]^{m-k} w(j, k, i)\ , \tag{8}$$

where the weights $w(j, k, i)$ are the probabilities that, in a family with parental mating type j, there are k affected sibs having HLA configuration of type i. To use these probabilities, it is necessary to adopt a convention for listing the possible HLA configurations of affected sibs, and we do this, for $k = 2, 3, 4$, and 5, in table 1.

The weights $w(j, k, i)$ are given in table 2 for $k = 2, 3, 4$, and 5 (covering effectively all values arising in practice), and are calculated by a routine (although tedious) examination of all possible recombination events leading to the required HLA configuration.

One immediate use that can be made of equations (7) and (8) is to calculate the theoretical *XYZ* ratios for families of size larger than two (recall that the equations (3) apply only to families of size two). We have done this for various x, p, and λ combinations and exhibit some results in table 3. This table reveals two important features. First, while the ratios are no longer independent of x [as

TABLE 1

NOTATION FOR HLA PATTERNS OF AFFECTED SIBS

No. affected sibs (k)	Pattern no. (k, i)	HLA pattern
2	2, 1	*ac/ac*
	2, 2	*ac/ad*
	2, 3	*ac/bd*
3	3, 1	*ac/ac/ac*
	3, 2	*ac/ac/ad*
	3, 3	*ac/ac/bd*
	3, 4	*ac/ad/bd*
4	4, 1	*ac/ac/ac/ac*
	4, 2	*ac/ac/ac/ad*
	4, 3	*ac/ac/ac/bd*
	4, 4	*ac/ac/ad/bd*
	4, 5	*ac/ac/ad/ad*
	4, 6	*ac/ac/bd/bd*
	4, 7	*ac/ac/ad/bc*
	4, 8	*ac/ad/bc/bd*
5	5, 1	*ac/ac/ac/ac/ac*
	5, 2	*ac/ac/ac/ac/ad*
	5, 3	*ac/ac/ac/ad/ad*
	5, 4	*ac/ac/ac/ac/bd*
	5, 5	*ac/ac/ac/ad/bc*
	5, 6	*ac/ac/ac/ad/bd*
	5, 7	*ac/ac/bc/ad/ad*
	5, 8	*ac/ac/ac/bd/bd*
	5, 9	*ac/ac/bc/ad/bd*
	5, 10	*ac/ad/ad/bc/bc*

they were in equations (3)], the dependence on x is very weak, and this implies that no really satisfactory estimate of x can be found using shared-haplotype information only. Second, the value of Z depends quite significantly on family size (at least for large x, p, and λ), and in view of the sensitivity of parameter estimation under the shared-haplotype approach to the value of Z, it appears unwise to use the ratios (3) for estimation purposes for families that have exactly two affected sibs but may be of various sizes.

NUMERICAL RESULTS

Our main aim here is to exhibit the maximum likelihood equations and to compare their solutions with those arising from other approaches. Thus, although we have doubts that the model we use (one susceptibility locus, one susceptibility allele) is appropriate for IDDM, we present in this section some numerical results found by applying the maximum likelihood methods to Cudworth's data for IDDM (see table 4). These data have also been analyzed recently by Motro and Thomson [3] and by Green et al. [4]. Both these analyses assume $R = 0$, and thus we also make this assumption for the moment. We also temporarily assume a complete ascertainment model, since this perhaps comes closest to the ascertainment model implicitly assumed by Green et al.

Application of the maximum likelihood procedures that we have described leads to

TABLE 2

VALUES OF THE WEIGHT FUNCTION $w(j, k, i)$ FOR $k = 2, 3, 4$, and 5 (R NONZERO)

Define: $A = 1 - R + R\lambda$
$B = R + \lambda - R\lambda$
$C = (1 - R)^2 + 2R(1 - R)\lambda$
$D = R(1 - R) + \{R^2 + (1 - R)^2\}\lambda$
$E = R^2 + 2R(1 - R)\lambda$

$k = 2$. $w(j, 2, i)$ is $x^2/16$ times the value listed, for each i, j pair

	$j =$	1	2	3	4	5
$i =$	1	4	$2(A^2 + B^2)$	$C^2 + 2D^2 + E^2$	$4\lambda^2$	$2\lambda^2[R^2 + (1 - R)^2]$
	2	8	$2(A + B)^2$	$4D(C + E)$	$8\lambda^2$	$2\lambda^2$
	3	4	$4AB$	$2(D^2 + CE)$	$4\lambda^2$	$4\lambda^2R(1 - R)$

$k = 3$. $w(j, 3, i)$ is $x^3/64$ times the value listed, for each i, j pair

	$j =$	1	2	3	4	5
$i =$	1	4	$2(A^3 + B^3)$	$C^3 + 2D^3 + E^3$	$4\lambda^3$	$2\lambda^3[R^3 + (1 - R)^3]$
	2	24	$6(A^3 + A^2B + B^2A + B^3)$	$6(C^2D + D^2C + E^2D + D^2E)$	$24\lambda^3$	$6\lambda^3[R^2 + (1 - R)^2]$
	3	12	$6(A^2B + B^2A)$	$6D^3 + 3EC(C + E)$	$12\lambda^3$	$6\lambda^3R(1 - R)$
	4	24	$12(A^2B + B^2A)$	$12CDE + 6D^2(E + C)$	$24\lambda^3$	$12\lambda^3R(1 - R)$

(Table continues on p. 866)

TABLE 2 (*Continued*)

$k = 4$. $w(j, 4, i)$ is $x^4/256$ times the value listed, for each i, j pair

$j =$	1	2	3	4	5
$i = 1$	4	$2(A^4 + B^4)$	$C^4 + 2D^4 + E^4$	$4\lambda^4$	$2\lambda^4[R^4 + (1 - R)^4]$
$i = 2$	32	$8(A^4 + A^3B + B^3A + B^4)$	$8(C^3D + D^3C + E^3D + D^3E)$	$32\lambda^4$	$8\lambda^4[R^3 + (1 - R)^3]$
$i = 3$	16	$8(A^3B + B^3A)$	$4(C^3E + E^3C + 2D^4)$	$16\lambda^4$	$8\lambda^4R(1 - R)[R^2 + (1 - R)^2]$
$i = 4$	96	$24(A^3B + 2A^2B^2 + B^3A)$	$24CDE(C + E) + 24D^3(C + E)$	$96\lambda^4$	$24\lambda^4R(1 - R)$
$i = 5$	24	$6(A^4 + 2A^2B^2 + B^4)$	$12D^2(C^2 + E^2)$	$24\lambda^4$	$6\lambda^4[R^2 + (1 - R)^2]^2$
$i = 6$	12	$12A^2B^2$	$6(C^2E^2 + D^4)$	$12\lambda^4$	$12\lambda^4R^2(1 - R)^2$
$i = 7$	48	$24(A^3B + B^3A)$	$12D^2(C + E)^2$	$48\lambda^4$	$24\lambda^4R(1 - R)[R^2 + (1 - R)^2]$
$i = 8$	24	$24A^2B^2$	$24CD^2E$	$24\lambda^4$	$24\lambda^4R^2(1 - R)^2$

$k = 5$. $w(j, 5, i)$ is $x^5/1{,}024$ times the value listed, for each i, j pair

$j =$	1	2	3	4	5
$i = 1$	4	$2(A^5 + B^5)$	$C^5 + 2D^5 + E^5$	$4\lambda^5$	$2\lambda^5[R^5 + (1 - R)^5]$
$i = 2$	40	$10(A^5 + A^4B + AB^4 + B^5)$	$10(C^4D + D^4C + E^4D + D^4E)$	$40\lambda^5$	$10\lambda^5[R^4 + (1 - R)^4]$
$i = 3$	80	$20(A^5 + A^3B^2 + A^2B^3 + B^5)$	$20(C^3D^2 + D^3C^2 + E^3D^2 + D^3E^2)$	$80\lambda^5$	$20\lambda^5[R^5 + R^3(1 - R)^2 + R^2(1 - R)^3 + (1 - R)^5]$
$i = 4$	20	$10(A^4B + B^4A)$	$5(C^4E + E^4C + 2D^5)$	$20\lambda^5$	$10\lambda^5R(1 - R)[R^3 + (1 - R)^3]$
$i = 5$	80	$40(A^4B + B^4A)$	$20(C^3D^2 + 2D^3EC + E^3D^2)$	$80\lambda^5$	$40\lambda^5R(1 - R)[R^3 + (1 - R)^3]$
$i = 6$	160	$40(A^4B + A^3B^2 + A^2B^3 + AB^4)$	$40(C^3DE + D^4C + D^4E + E^3CD)$	$160\lambda^5$	$40\lambda^5R(1 - R)[R^2 + (1 - R)^2]$
$i = 7$	240	$60(A^4B + A^3B^2 + A^2B^3 + AB^4)$	$60(C^2D^3 + E^2D^3 + D^2C^2E + D^2E^2C)$	$240\lambda^5$	$60\lambda^5R(1 - R)[R^2 + (1 - R)^2]$
$i = 8$	40	$20(A^3B^2 + A^2B^3)$	$10(C^3E^2 + 2D^5 + C^2E^3)$	$40\lambda^5$	$20\lambda^5R^2(1 - R)^2$
$i = 9$	240	$120(A^3B^2 + A^2B^3)$	$60(C^2D^2E + 2CD^3E + CD^2E^2)$	$240\lambda^5$	$120\lambda^5R^2(1 - R)^2$
$i = 10$	120	$60(A^3B^2 + A^2B^3)$	$30(DC^4 + 2C^2DE^2 + ED^4)$	$120\lambda^5$	$60\lambda^5R^2(1 - R)^2$

TABLE 3

THEORETICAL *X*, *Y*, AND *Z* RATIOS IN FAMILIES OF SIZES 2, 3, AND 4, EACH FAMILY HAVING EXACTLY TWO AFFECTED SIBS, FOR VARIOUS x, p, AND λ VALUES

p	λ	Family size	$x = .1$			$x = .8$		
			X	*Y*	*Z*	*X*	*Y*	*Z*
.05	.1	2	.6006	.3743	.0251	.6006	.3743	.0251
		3	.6002	.3750	.0248	.5966	.3809	.0225
		4	.5998	.3757	.0246	.5891	.3900	.0209
.05	.4	2	.4587	.4988	.0425	.4587	.4988	.0425
		3	.4594	.4989	.0417	.4665	.4990	.0345
		4	.4602	.4989	.0409	.4728	.4992	.0279
.20	.1	2	.5374	.3981	.0645	.5374	.3981	.0645
		3	.5401	.3967	.0632	.5689	.3813	.0498
		4	.5427	.3953	.0620	.5853	.3713	.0434
.20	.4	2	.3835	.4978	.1187	.3835	.4978	.1187
		3	.3849	.4978	.1172	.4004	.4979	.1017
		4	.3863	.4979	.1158	.4147	.4980	.0872

$$\hat{p} = .160(\pm .089), \hat{\lambda} = .103(\pm .041), \hat{x} = .470(\pm .167) , \tag{9}$$

where values in parentheses are $\pm$ 1 standard error. We mentioned in the previous section that, under the shared-haplotype approach, the estimators p and λ would not, except in very large samples, have approximately normal distributions. The same remark is true of maximum likelihood estimates, for much the same reason as that given for *XYZ*-derived estimators. We therefore do not believe, for example, that it is approximately 95% likely that λ lies in $.103 \pm .082$ and that x lies in $.470 \pm .334$. Nor do we believe that the standard normal theory testing procedures using the chi-square statistic $2 \log [L_1(\max)/L_2(\max)]$ are valid for testing hypotheses about parameters in this model.

This view is supported by the fact that although $\hat{\lambda}$ [given in equations (9)] is 2.5 standard deviations above zero, formally a significant excess, the chi-square statistic for the hypothesis $\lambda = 0$ takes the value 2.00, which is not large enough to reject the hypothesis. This seeming contradiction arises from, and confirms, the quite nonnormal joint distribution of $\hat{\lambda}$ and $\hat{p}$.

If we assume that $\lambda = 0$, the maximum likelihood estimators of p and x are

$$\hat{p} = .346, \hat{x} = .309 . \tag{10}$$

We note the rather high value of $\hat{p}$ and the consequent high estimated population prevalence of .370 [see equation (2)]. We will return to the assumption $\lambda = 0$ later.

We have just mentioned the population prevalence. If this can be assumed known extrinsically, a trivial programming amendment allows us to estimate parameters subject to the prevalence value's being given. The estimates (9) lead to an estimated prevalence of .025, six times larger than the often quoted value of about .004 for IDDM (Spielman et al. [5]). Under the constraint that the

prevalence is .004, the maximum likelihood estimates of p, x, and λ are $\hat{p} = .039, \hat{\lambda} = .093$, and $\hat{x} = .470$.

We now turn briefly to the sensitivity of the estimates to small changes in the data. In Cudworth's data (table 4), there is an increase as the family size increases (rather than the theoretical decrease under the model presently being considered) in the proportion of families having two affected with the HLA configuration *ac/bd*. This is very largely caused by the four families of size three having two affected sibs with this HLA configuration. If we were to ignore these four families, our estimates of p, x, and λ become

$$\hat{p} = .083, \hat{\lambda} = .105, \hat{x} = .630 , \tag{11}$$

and the estimate of p is now about half its previous value. This illustrates the sensitivity of the estimate of p to data from four families only (out of 133). This sensitivity will apply to any estimation procedure, in particular to any generalized shared-haplotype procedure that attempts to use data from families of size two or more, and makes us quite wary of many published parameter estimates for IDDM and other diseases.

TABLE 4

FAMILY TYPES REPRESENTED IN THE DATA OF CUDWORTH, TOGETHER WITH NO. FAMILIES OF EACH TYPE OBSERVED

Family size	No. affected	Type no.	No. observed
2	2	2, 2, 1	26
		2, 2, 2	19
		2, 2, 3	1
3	2	3, 2, 1	21
		3, 2, 2	14
		3, 2, 3	4
3	3	3, 3, 1	2
4	2	4, 2, 1	14
		4, 2, 2	9
		4, 2, 3	1
4	3	4, 3, 1	1
		4, 3, 2	2
		4, 3, 3	1
4	4	4, 4, 1	1
5	2	5, 2, 1	4
		5, 2, 2	2
		5, 2, 3	1
5	3	5, 3, 2	2
		5, 3, 4	1
6	2	6, 2, 1	4
		6, 2, 2	1
11	2	11, 2, 2	1
11	3	11, 3, 1	1
Total			133

We now compare our estimating procedure, and our estimates, with those of Green et al. [4] and Motro and Thomson [3]. Green et al. center their estimation procedure around a statistic ΣN, the sum over all families of the sum over each parent in each family of the numbers of the most frequently occurring *HLA* haplotypes from each parent. We believe, first, that use of this statistic loses part of the information afforded by the data. Second, and more important, in computing the mean and variance of ΣN, Green et al. do not apply the condition that at least two sibs in each family be affected. In ascertainment sampling schemes such as that leading to Cudworth's data, where a family must satisfy this condition to enter the sample, it is necessary that this condition be applied: for example, the probabilities $P(m, k, i)/P(m)$ in equations (4) are probabilities conditional on exactly this requirement. We are therefore not confident that their parameter estimates are accurate and do not pursue them further here.

The analysis of Motro and Thomson [3] is centered around the *XYZ* ratios and applies these ratios for families of size two or more, with two or more affected sibs. Motro and Thomson use the classical values (3) in their analysis, applying these pairwise to affected sibs when more than two affected sibs arise in the family. Although there is no bias in this procedure, we believe it does not lead to estimates as efficient as those given by maximum likelihood.

Using three different ascertainment models, Motro and Thomson find unexpectedly high estimates of p, not only for IDDM but also for other diseases. This does not occur because they use a different estimation procedure than we do, but, rather, is due to their making the assumption $\lambda = 0$. For essentially the same data set as we have used, and assuming $\lambda = 0$, they estimate p at .330 for the ascertainment model considered in this section. (Our estimate, for $\lambda = 0$, is quite close to this, namely, .346. The closeness of the two estimates arises because most families in the data set contain only two sibs: if all families contained only two sibs, the estimates would be identical.) The approach of Motro and Thomson is to test the hypothesis $\lambda = 0$ (which, with their data, they accept) and then to estimate p assuming $\lambda = 0$. In view of the difficulty of formulating a reliable test of the hypothesis $\lambda = 0$ (arising from the quite nonnormal joint distribution of $\hat{p}$ and $\hat{\lambda}$ as described earlier), we have some reservations about this procedure. Our preference is to estimate p and λ jointly, since, as we have seen, the estimate of p is very sensitive to the choice of λ, and even if the hypothesis $\lambda = 0$ can be accepted using a valid test (a procedure not yet available), a large difference can be found in the estimate of p by forcing $\lambda = 0$ and by allowing λ to be estimated simultaneously with p.

NONZERO RECOMBINATION

In all of the above analyses, we have put R (the recombination fraction between the susceptibility locus and *HLA*) equal to zero, so as to compare our estimates with those of Green et al. [4] and Motro and Thomson [3], who make this assumption. Our estimation procedure, however, allows nonzero values of R. When R is a free parameter, the likelihood of the observations is a function of four parameters, x, p, λ, and R, and may be maximized with respect to all four

parameters by standard numerical methods. The estimates of these parameters, using all the data of table 4, are

$$\hat{p} = .00087\ (\pm .0169), \quad \hat{\lambda} = .0040\ (\pm .0492),$$

$$\hat{x} = .773\ (\pm .280), \quad \hat{R} = .133\ (\pm .054)\ . \tag{12}$$

Note the great difference in the estimates of p, λ, and x from those in equation (10), where the restriction $R = 0$ is imposed. This suggests that great caution should be exercised in assuming zero recombination between a susceptibility locus and *HLA*, and that estimates of genetic parameters obtained when this restriction is made should be viewed with great caution.

DIFFERENT ASCERTAINMENT MODELS

The ascertainment model assumed above is that of "complete ascertainment": the ascertainment probability α of any family with at least two affected sibs is independent of the number of affected sibs. The actual choice of ascertainment model is, of course, a matter of considerable difficulty, and we prefer to adopt the approach, used by Motro and Thomson, of considering several ascertainment models. A broad class of ascertainment schemes is covered by the assumption that the probability of ascertainment of a family with k affected sibs is of the form

$$\alpha[1 - (1 - \pi)^k]\ , \tag{13}$$

where α and π are unknown parameters. The choice $\pi = 1$ corresponds to complete ascertainment while the limiting case $\pi \to 0$ is that of single ascertainment. As with our previous analysis, the parameter α drops out of the estimation equations and thus cannot be estimated. We may thus assume $\alpha = 1$, for simplicity, in the estimation process.

We have generalized our theory to allow for an ascertainment scheme described by equation (13). We do not give the mathematical details here. In the generalized model, we may either restrict π to particular values (for example, a value close to zero, corresponding to the single-ascertainment model) or let π be a further free parameter, and maximize the likelihood with respect to it and the remaining parameters. Initially, we considered both approaches, but found that under the second approach the standard error of the estimate of π was extremely large (often the asymptotic maximum likelihood formula gave a standard error numerically larger than 1) and, as a consequence of the correlation between the parameter estimates, this gave unrealistically large values of the standard errors of the remaining parameter estimates. We found a much more satisfactory approach was to fix π at particular chosen values (and, in particular, on at two cases $\pi \simeq 0$, $\pi = 1$, corresponding to single ascertainment and complete selection, respectively), and to estimate the remaining parameters for each fixed π value.

The maximum likelihood estimates of the remaining parameters when we set π equal to the (small) value .001 are:

$$\hat{p} = .00090 \pm .01234, \quad \hat{\lambda} = .0050 \pm .0365,$$

$$\hat{x} = .573 + .243, \qquad \hat{R} = .129 \pm .050 \tag{14}$$

We note, by comparing equations (12) and (14), that the estimates of p and R are very little affected by the choice of π, that is, they are very close in the single-ascertainment and the complete-ascertainment models. (They are, in fact, close for all choices of π between 0 and 1.) The estimate of λ changes rather more, from .0040 to .0050 as π decreases from 1 to 0, while, as expected, the value of x changes the most (from .773 to .570). Thus, there is a far smaller change in parameter estimates as one changes the ascertainment model than there is by assuming special values for parameters within a fixed-ascertainment model, a conclusion that is of some interest in view of the considerable attention paid recently to choices of ascertainment models.

To emphasize this point, we calculated the single ascertainment (i.e., $\pi = .001$) estimates of p and x under the assumptions $R = 0$, $\lambda = 0$. Our estimates were

$$\hat{p} = .332 \pm .045, \hat{x} = .229 \pm .059 . \tag{15}$$

Comparison of equations (10) and (15) shows very little difference in the estimated values of p under complete- and single-ascertainment models, and a moderate change to the estimate of x. Motro and Thomson reached, so far as estimation of p is concerned, an essentially identical conclusion. On the other hand, comparison of equations (14) and (15) shows that large changes to the estimates of the remaining parameters can arise by imposing the assumptions $R = 0$ and $\lambda = 0$, even within a given ascertainment scheme. This is why we have claimed that it is dangerous, without some good reason, to make these assumptions.

REMARKS

We emphasize that our main aim in this paper is not to produce parameter estimates for IDDM, but to outline the theory of maximum likelihood estimation for the one-susceptibility-locus, one-susceptibility-allele model when data come from families of different sizes. We do not feel sufficiently confident that this is the correct model for IDDM for the parameter estimates given above to be accurate. We have used IDDM data for three reasons: to illustrate the theory, to compare the parameter estimates with those found by Motro and Thomson [3] and Green et al. [4] (who do assume this model), and to demonstrate the sensitivity of parameter estimation (in the maximum likelihood approach, at least), to assuming specific values for some parameters and to small subsets of the data.

Regarding the latter point, we believe that all parameters should be estimated freely and not under assumptions concerning particular parameter values. Not enough is known about the true values of these parameters to make the fixing of parameter values a safe approach in the light of the sensitivity of estimation referred to above, nor are formal normal-theory tests of hypotheses about parameter

values reliable in the light of the marked nonnormal distributions of parameter estimates.

In view of the sensitivity of parameter estimation to small parts of the data, it is clear that not only a much larger data set than Cudworth's will be necessary before reliable estimates are found, but also data sets involving more information from each family. In this connection, we remark that the theory presented here can be extended readily to cover data where the disease status of parents is used as part of the data. It will be shown (Spielman and Ewens [7]) that use of this information reduces variances of parameter estimates to about one-sixth of their previous values. We are, thus, wary about the value of any parameter estimates published in the literature that are arrived at without using this or similar information.

ACKNOWLEDGMENTS

We thank Barbara Morrison in Melbourne and Jane Wu in Philadelphia for their computing assistance. We also thank Drs. Richard Spielman, Sue Hodge, and Glenys Thomson for very many invaluable comments, and Dr. Kathy Gogolin for drawing our attention to the problems taken up in this paper.

REFERENCES

1. THOMSON G, BODMER W: The genetic analysis of HLA and disease, in *HLA and Disease*, edited by DAUSSET J, SVEJGAARD A, Copenhagen, Munksgaard, 1977, pp 56–63
2. LOUIS EJ, THOMSON G, PAYAMI H: The affected sib method. II. The intermediate model. *Ann Hum Genet* 47:225–243, 1983
3. MOTRO U, THOMSON G: The affected sib method. I: Statistical features of the affected sib pair method. Submitted for publication
4. GREEN JR, HENG CHIN LOW, WOODROW JC: Inference on inheritance of disease using repetitions of HLA haplotypes in affected siblings. *Ann Hum Genet* 47:73–82, 1983
5. SPIELMAN RS, BAKER L, ZMIJEWSKI CM: Gene dosage and susceptibility to insulin-dependent diabetes. *Ann Hum Genet* 44:135–150, 1980
6. SUAREZ BK, RICE J, REICH T: The generalized sib pair IBD distribution: its use in the detection of linkage. *Ann Hum Genet* 42:87–94, 1978
7. SPIELMAN RS, EWENS WJ: Operating characteristics of maximum likelihood estimates of parameters of HLA-related diseases. Submitted for publication
8. EWENS WJ: Aspects of parameter estimation in ascertainment sampling schemes. *Am J Hum Genet* 34:853–865, 1982
9. STENE J: Choice of ascertainment model. I and II. *Ann Hum Genet* 42:219–229, 493–505, 1979