BIOMETRICS 42, 393–399
June 1986

Effects of Misspecifying Genetic Parameters in Lod Score Analysis

Françoise Clerget-Darpoux, Catherine Bonaïti-Pellié, and Joëlle Hochez
Unité de Recherches de Génétique Epidémiologique, I.N.S.E.R.M.-U. 155,
Château de Longchamp, Bois de Boulogne, 75016 Paris, France

SUMMARY

The lod score method is widely used to test linkage and to estimate the recombination fraction between a disease locus and a marker locus. The parameters (gene frequency, penetrance, and degree of dominance) are assumed to be known at each locus. This condition may not be fulfilled at the disease locus.

In this paper, we evaluate the errors due to the use of wrong parameters. The power of the linkage test is sensitive to the degree of dominance, and slightly to the penetrance, but not to the gene frequency. In contrast, the estimation of the recombination fraction may be strongly affected by an error on any genetic parameter.

1. Introduction

Lod score analysis has been proposed by Morton (1955) to study genetic linkage between a trait locus and a marker locus. The method was intended to be applied to traits with known mode of inheritance and allele frequencies. It permitted the localisation of certain genes of diseases with simple Mendelian inheritance. Later it was extended to the study of diseases with incomplete penetrance, and appropriate programmes were proposed by Ott (1974). In such a case, the lod score may be considered as a function of several parameters, the recombination fraction θ, the disease gene frequency, and the penetrance vector. The purpose of the lod score analysis is to test $\theta < \frac{1}{2}$ against $\theta = \frac{1}{2}$ and, if significant, to estimate θ assuming the other parameters known.

At the present time, many linkage studies are focused on diseases with uncertain mode of inheritance and even more uncertain parameters.

Misspecifying the disease genetic model necessarily leads to a bias in the recombination fraction. Some authors have pointed out this bias in some specific situations (Ott, 1977; Clerget-Darpoux and Bonaïti-Pellié, 1980; Spielman, Baker, and Zmijewsky, 1980; Hodge and Spence, 1981; Suarez and Van Eerdewegh, 1981; Clerget-Darpoux, 1982; Hodge et al., 1983).

The purpose of this study is to quantify under the single-locus model the effects of using wrong genetic parameters, on the linkage test and on the recombination fraction estimate between a disease locus and a genetic marker.

2. Model

We consider a single locus with an allele, g, responsible for susceptibility to a disease, and a normal allele, G. Let the frequency of the disease allele g be q_0, the penetrance of the homozygote gg be f_0, and the penetrance of the heterozygote Gg be $\lambda_0 f_0$. For the sake of

Key words: Biased recombination fraction; Genetic parameters; Linkage; Lod score.

simplicity, the penetrance of the homozygote *GG* is assumed null, so that the penetrance vector is $(f_0, \lambda_0 f_0, 0)$. Thus, the transmission of the disease is completely specified by the three parameters q_0, f_0, λ_0. Henceforth we shall refer to f_0 as the penetrance and λ_0 as the degree of dominance. We denote the recombination fraction between the disease locus and the marker by θ_0.

3. Method

Let us consider the affected status and the marker genotype of each member of a family. The likelihood of such a family may be considered as a function of θ and depends on q_0, f_0, and λ_0 as defined above. This likelihood is denoted $L_{q_0,f_0,\lambda_0}(\theta)$ and the lod score is defined by

$$Z_{q_0,f_0,\lambda_0} = \log_{10}[L_{q_0,f_0,\lambda_0}(\theta)/L_{q_0,f_0,\lambda_0}(\theta = \tfrac{1}{2})].$$

If the test $\theta < \frac{1}{2}$ against $\theta = \frac{1}{2}$ is significant, the estimate of θ is the value which maximizes the lod score function.

The values q_0, λ_0, and f_0 may be unknown, and in this case we denote by $Z_{q,f,\lambda}(\theta)$ the lod score functions where q, f, and λ are parameters. Our purpose is to assess the bias in the recombination fraction and the lod score due to an error in one of the three parameters q, λ, and f.

The method is based on the same principles as those developed by Clerget-Darpoux and Bonaïti-Pellié (1980) and Clerget-Darpoux (1982). By considering all the possible nuclear families with a given sibship size s, we derive the expected distribution (F_i, p_i) of such families F_i according to the model defined by λ_0, q_0, f_0, θ_0, and to the ascertainment method.

Let $Z^i_{q,f,\lambda}(\theta)$ be the lod score functions of family F_i. The expectations of lod scores for a random nuclear family are

$$\mathrm{E}Z(\theta) = \mathrm{E}[Z_{q,f,\lambda}(\theta)] = \sum p_i Z^i_{q,f,\lambda}(\theta).$$

For $(q, f, \lambda) = (q_0, f_0, \lambda_0)$, $\mathrm{E}Z(\theta) = \mathrm{E}Z_0(\theta)$.

The respective maxima of the functions $\mathrm{E}Z$ are called $\mathrm{E}Z_{\max}$, and the corresponding values of θ, $\hat{\theta}$. Note that $\mathrm{E}Z_{0,\max}$ is obtained for $\theta = \theta_0$ (Edwards, 1972, Chap. 7). The bias in the recombination fraction $\hat{\theta} - \theta_0$ is denoted $\Delta\theta$.

In the following sections, we study $\Delta\theta$ successively as a function of the differences $\Delta q = q - q_0$, $\Delta f = f - f_0$, and $\Delta\lambda = \lambda - \lambda_0$. For instance, to study the effect of Δq, $\mathrm{E}Z(\theta)$ was computed for different sets of parameters (q, f_0, λ_0). In all cases, we considered only families with at least two affected children. The marker was taken highly polymorphic with codominant alleles so that the probability of homozygocity was negligible.

4. Effects of Using Wrong Value of *q*

Underestimation of the gene frequency ($\Delta q < 0$) leads to a negligible change of $\mathrm{E}Z_{\max}$ and a large overestimation of θ ($\Delta\theta > 0$). This is illustrated by Figure 1, which shows $\mathrm{E}Z(\theta)$ for different q-values when $q_0 = 0.20$, $f_0 = 0.05$, $\lambda_0 = 1$, $\theta_0 = 0$. The values $\mathrm{E}Z_{\max}$ remain almost the same, which means that the power of the detection of linkage is unchanged. In contrast, the error $\Delta\theta$ may be quite high: for $q = 0.01$ ($\Delta q = -0.19$), the bias in θ can be as large as 0.15. In other words, a frequent disease gene strictly linked to the marker would seem to be located at a large genetic distance if the lod score analysis was performed using a low gene frequency.

These results are general, whatever the values of s, q_0, f_0, λ_0, and θ_0. That is to say, $\mathrm{E}Z_{\max}$ is almost constant. For a given Δq, $\Delta\theta$ depends neither on sibship size, nor on the q_0 value.

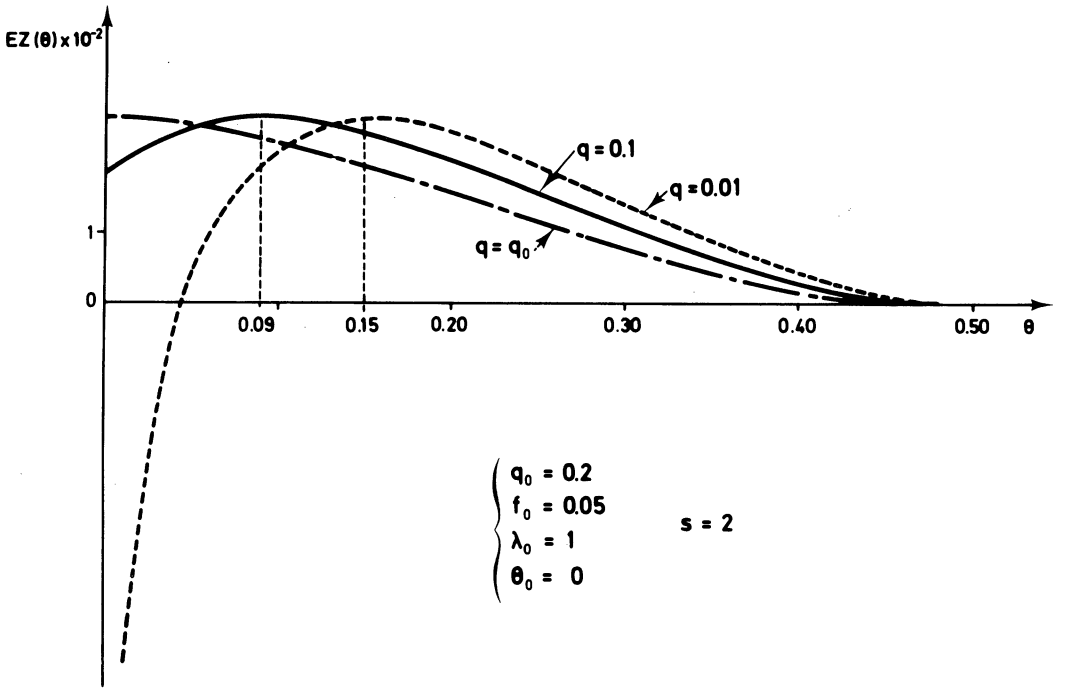

Figure 1. Effects of misspecifying gene frequency on the estimate of the recombination fraction (see text for notations).

Table 1
Values of $\Delta\theta$ *for* $\Delta q = -0.20$ *and for different values of* f_0, λ_0, *and* θ_0 ($s = 2$)

	$\lambda_0 = 1, \theta_0 = 0$		
f_0	0.05	0.50	0.90
$\Delta\theta$	0.16	0.15	0.11
	$f_0 = 0.05, \theta_0 = 0$		
λ_0	0	0.2	0.5	1
$\Delta\theta$	0.09	0.10	0.15	0.16
	$f_0 = 0.05, \lambda_0 = 1$		
θ_0	0	0.10	0.20
$\Delta\theta$	0.16	0.13	0.10

However, $\Delta\theta$ does depend on the values of f_0, λ_0, and θ_0 as shown in Table 1, $\Delta\theta$ being greater when f_0 and θ_0 are small and when λ_0 is not close to zero.

5. Effects of Using Wrong Value of f

Misspecifying the penetrance leads to a slight underestimation of EZ_{max} and to a biased estimate of θ. These biases are small for $s = 2$ but increase with sibship size. Figure 2 shows, for $s = 4$, $EZ(\theta)$ for different f values when $q_0 = 0.20$, $f_0 = 0.05$, $\lambda_0 = 0.5$, and $\theta_0 = 0$. When Δf increases, EZ_{max} and, consequently, the power of detection of linkage both slightly decrease. Overestimating f leads to overestimating θ: for instance, when $f = 0.50$ ($\Delta f = 0.45$), θ is estimated as 0.08.

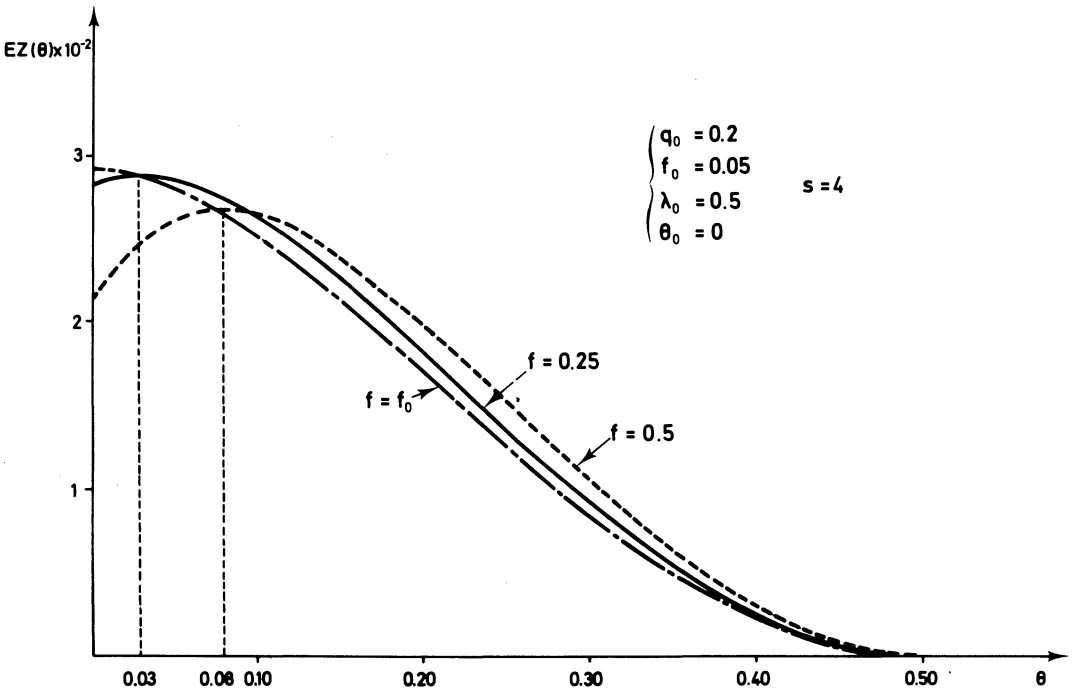

Figure 2. Effects of misspecifying penetrance on the estimate of the recombination fraction (see text for notations).

Table 2
Values of $\Delta\theta$ *for* $\Delta f = 0.40$ *and for different values of* q_0, f_0, λ_0, *and* θ_0 ($s = 4$)

	$f_0 = 0.05, \lambda_0 = 0.5, \theta_0 = 0$		
q_0	0.01	0.20	0.35
$\Delta\theta$	0.01	0.07	0.10
	$\lambda_0 = 0.5, q_0 = 0.20, \theta_0 = 0$		
f_0	0.05	0.20	0.50
$\Delta\theta$	0.07	0.07	0.09
	$q_0 = 0.20, f_0 = 0.05, \theta_0 = 0$		
λ_0	0	0.5	1
$\Delta\theta$	0.06	0.07	0.11
	$q_0 = 0.20, f_0 = 0.05, \lambda_0 = 0.4$		
θ_0	0	0.10	0.20
$\Delta\theta$	0.07	0.06	0.04

Although the results are general, the magnitude of the bias $\Delta\theta$ depends on the values of q_0, f_0, λ_0, and θ_0, as shown in Table 2. For a given Δf, $\Delta\theta$ is greater when q_0 and f_0 are high, θ_0 is small, and λ_0 is not close to zero.

6. Effects of Using Wrong Value of λ

Misspecifying the degree of dominance leads to an underestimation of EZ_{max} and to a biased estimate of θ.

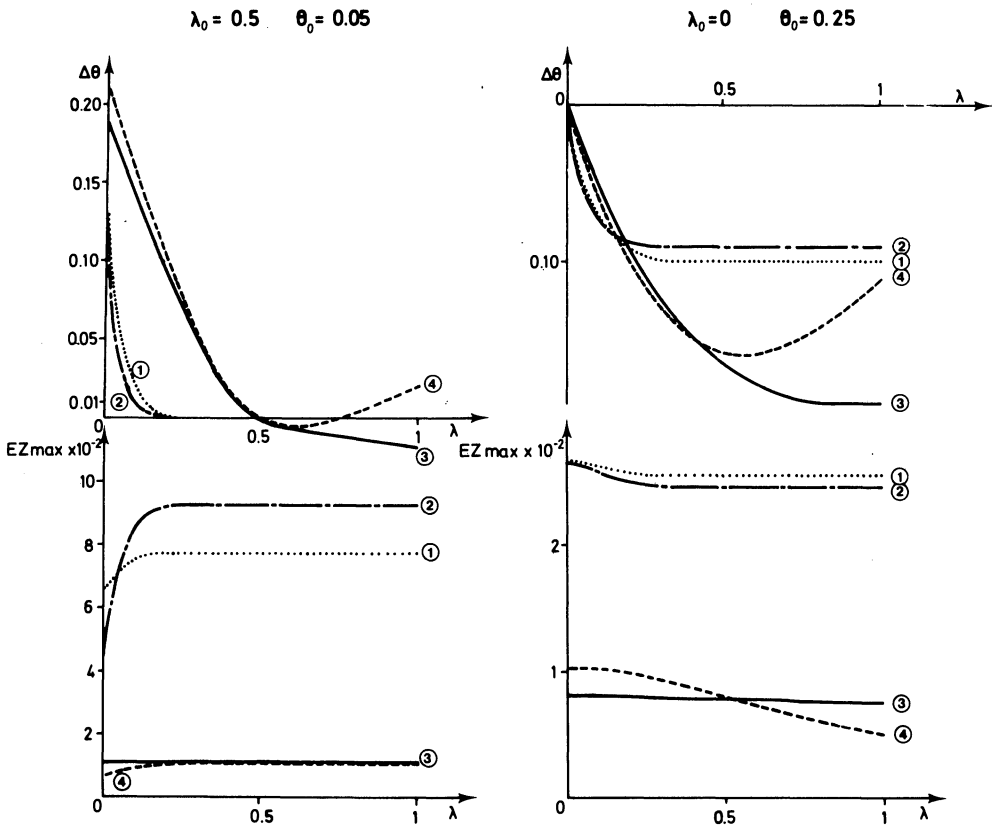

Figure 3. Effects of misspecifying degree of dominance on the estimate of the recombination fraction and on the maximum expectation of lod score for different models ($\lambda_0 = 0.5$, $\theta_0 = 0$); ($\lambda_0 = 0$, $\theta_0 = 0.25$): (1) $q_0 = 0.01$, $f_0 = 0.05$; (2) $q_0 = 0.01$, $f_0 = 0.50$; (3) $q_0 = 0.30$, $f_0 = 0.05$; (4) $q_0 = 0.30$, $f_0 = (0.50)$.

EZ_{max} and $\Delta\theta$ were computed for different values of λ, and for various values of the parameters q_0, f_0, λ_0, and θ_0. The shapes of the curves are quite different according to the values of these parameters, as illustrated in Figure 3. In some cases, the underestimation of EZ_{max} may not be negligible. A general pattern is visible: while analyzing under dominant or additive modes leads to very similar results, the conclusions may be very different when mistaking recessive and intermediate models.

7. Discussion

In this study, considering a single-locus model for a disease, we have shown that misspecifying the genetic parameters at this locus may lead to a bias in the recombination fraction estimate θ between this locus and a marker locus. This had already been mentioned by other authors in specific situations. For example, Ott (1977) pointed out the effects of misclassification errors (such as reduced penetrance). Spielman et al. (1980) showed that θ varied with the λ value. Suarez and Van Eerdewegh (1981) obtained different θ estimates when analyzing simulated data with different sets of parameters. However, none of these studies provided quantitative information on this bias nor permitted the respective effect of each parameter to be measured.

The advantage of our approach is that the conclusions are not limited by the small-sample properties. The bias in the recombination fraction has been computed for a random

nuclear family, using the lod score expectation $EZ(\theta)$. Maximizing this function provides an unbiased estimate of θ.

In this paper, we show that (i) The bias in θ may be quite large; and (ii) The effect of misspecifying one parameter depends not only on its true value but also on the values of the two other parameters.

Note that these errors may be combined because of the prevalence constraint:

$$P = q^2 f + 2q(1 - q)\lambda f.$$

In contrast, the underestimation of the maximum lod score is negligible in most cases (provided that the function $EZ(\theta)$ reaches its maximum within the interval $[0, \frac{1}{2}]$).

Similarly, other kinds of errors may also lead to erroneous conclusions. For example, an epistatic interaction between the disease locus and a marker (in particular, a two-locus model) could mimic loose linkage (Clerget-Darpoux and Bonaïti-Pellié, 1980; Hodge and Spence, 1981).

Ignoring gametic disequilibrium may also affect linkage analysis: it reduces the power of the test but has little effect on the θ estimate (Clerget-Darpoux, 1982).

In conclusion, the lod score method may be unsuitable to estimate the recombination fraction in those diseases where the mode of inheritance is uncertain. It would be more appropriate to maximize the lod score function for the four parameters q, f, λ, and θ. The function thus defined no longer depends only on θ but also on the parameters of the disease model; in work as yet unpublished, Clerget-Darpoux and Bonaïti-Pellié refer to it as the "mod score." Maximizing the mod score function is equivalent to maximizing the conditional probability of marker genotype and disease status given the disease status, as proposed by Risch (1984). However, considering the very small decrease of EZ_{max} induced by errors on parameters, discrimination between different sets of parameters will not be possible. In particular, an infinite number of sets (q, θ) will provide similar maximum mod scores. However, when there is a population association between the disease and the marker, under the assumption of a single-locus model with no epistatic interaction, one may set the constraint $\theta = 0$, and maximize the mod score function in order to estimate the other parameters.

Acknowledgements

We are grateful to the referees for their useful comments and to C. Vaillant for preparation of the manuscript.

Résumé

La méthode des lod scores est largement utilisée pour tester la liaison génétique et estimer le taux de recombinaison entre un locus maladie et un locus marqueur. Cette méthode présuppose que les paramètres (fréquence génique, pénétrance et degré de dominance) sont connus à chacun des locus. En fait, cette condition n'est pas toujours remplie au locus maladie.

Dans cet article, nous évaluons les erreurs induites par l'utilisation de faux paramètres. La puissance du test de la liaison génétique est sensible à une erreur sur le degré de dominance, et légèrement à une erreur sur la pénétrance, mais n'est pas modifiée par une erreur sur la fréquence génique. En revanche, une erreur sur l'un des paramètres génétiques peut fortement biaiser l'estimation du taux de recombinaison.

References

Clerget-Darpoux, F. (1982). Bias of the estimated recombination fraction and lod score due to an association between a disease gene and a marker gene. *Annals of Human Genetics* **46,** 363–372.

Clerget-Darpoux, F. and Bonaïti-Pellié, C. (1980). Epistasis effect: An alternative to the hypothesis of linkage disequilibrium in HLA associated diseases. *Annals of Human Genetics* **44,** 195–204.

Edwards, A. W. F. (1972). *Likelihood.* Cambridge: Cambridge University Press.

Hodge, S. E., Anderson, C. E., Neiswanger, K., Sparkes, R. S., and Rimoin, D. L. (1983). The search for heterogeneity in insulin-dependent diabetes mellitus (IDDM): Linkage studies, two-locus models, and genetic heterogeneity. *American Journal of Human Genetics* **35,** 1139–1155.

Hodge, S. E. and Spence, M. A. (1981). Some epistatic two-locus models of disease. II. The confounding of linkage and association. *American Journal of Human Genetics* **33,** 396–406.

Morton, N. E. (1955). Sequential tests for the detection of linkage. *American Journal of Human Genetics* **7,** 277–318.

Ott, J. (1974). Estimation of the recombination fraction in human pedigrees: Efficient computation of the likelihood for human linkage studies. *American Journal of Human Genetics* **26,** 588–597.

Ott, J. (1977). Linkage analysis with misclassification at one locus. *Clinical Genetics* **12,** 119–124.

Risch, N. (1984). Segregation analysis incorporating linkage markers. I. Single-locus models with an application to type I diabetes. *American Journal of Human Genetics* **36,** 363–386.

Spielman, R. S., Baker, L., and Zmijewski, C. M. (1980). Gene dosage and susceptibility to insulin dependent diabetes. *Annals of Human Genetics* **44,** 135–150.

Suarez, B. K. and Van Eerdewegh, P. (1981). Type I (insulin dependent) diabetes mellitus. Is there strong evidence for a non-HLA linked gene? *Diabetologia* **20,** 524–529.

Received September 1984; *revised June* 1985 *and February* 1986.