

Magnitude of Type I Error When Single-Locus Linkage Analysis Is Maximized over Models: A Simulation Study

Susan E. Hodge,^{1,2,3} Paula C. Abreu,³ and David A. Greenberg⁴

¹Department of Psychiatry, Columbia University, ²Clinical-Genetic Epidemiology Unit, New York State Psychiatric Institute, ³Division of Biostatistics, Columbia University School of Public Health, and ⁴Department of Psychiatry, Mt. Sinai School of Medicine, New York

Summary

It is well known that maximizing the maximum LOD score over multiple parameter values or models (i.e., the method of mod scores, or MMLS), will inflate type I error, compared with assuming only one parameter value/model in the linkage analysis. On the other hand, a mod score often has greater power to detect linkage than does a LOD score (Z) calculated under a wrong genetic model. Therefore, it is of interest to determine the actual magnitude of type I error in realistic genetic situations. Simulated data sets with no linkage were generated under three dominant and three recessive single-locus models, with reduced penetrance ($f = .8, .5$, and $.2$). Data sets were analyzed for linkage by (1) maximizing over penetrance only, (2) maximizing over “dominance model” (i.e., dominant versus recessive), and (3) maximizing over both penetrance and dominance model simultaneously. In (1), the resultant significance levels were approximately doubled, compared with baseline values if one had not maximized over penetrances (i.e., compared with a one-sided χ^2_1). In (2), significance levels were increased somewhat less, and, in (3), they were increased by approximately two to three times (but not more than four times) over those of the one-sided χ^2_1 . This means that, for a given size of test α , an investigator would need to increase the Z used as a test criterion, by ~ 0.30 LOD units for analyses as in (1) or (2) and by 0.60 Z units for analyses as in (3). These guidelines, which are valid up to $\approx Z = 3.0$, are conservative for (1) and are very conservative for (2) and (3). By quantifying the increase in significance level (or, correspondingly, the increase in Z), our findings will enable users to rationally assess the advantages versus the disadvantages of mod scores.

Introduction

The method of “mod scores” (Clerget-Darpoux et al. 1986; Clerget-Darpoux and Bonaïti-Pellié 1992), or “MMLS” (“maximizing the maximum LOD score”; Greenberg 1989), involves calculating the maximum LOD score, Z_{\max} , for more than one genetic model, then maximizing this result to obtain the maximized Z_{\max} ($\max Z_{\max}$) over multiple models. The theoretical underpinnings and justification for this approach have been discussed by several authors (e.g., see Elston 1989; MacLean et al. 1993; Hodge and Elston 1994). The major appeal of this method is its greater potential to detect linkage, when the true mode of inheritance of the disease is unknown. It is well known (e.g., see Clerget-Darpoux et al. 1986; Greenberg and Hodge 1989; Greenberg 1990; Vieland et al. 1992a, 1992b, 1993) that, if one analyzes linkage under a single model, and if that model happens to be far from the correct model, one is considerably more likely to miss a true linkage than if the correct model had been used. The same problem can arise with an affected-sib-pair method, which Knapp et al. (1994) have shown to be statistically equivalent to (i.e., have the same rejection region as) a LOD score (Z) analysis assuming *recessive* inheritance (also see Greenberg et al. 1996). On the other hand, if one analyzes linkage under at least a couple of different genetic models, one’s chance of detecting a true linkage increases.

There is no “free lunch,” and, obviously, as investigators consider more genetic models in this kind of analysis, they will pay a price in increased probability of type I error. If one were not maximizing over models or parameters, significance levels would be given by a one-sided χ^2_1 (the subscript indicates the number of df); we will refer to these as “baseline” significance levels. Clerget-Darpoux et al. (1990), Weeks et al. (1990), and Risch (1991) have examined some aspects of this problem, but there has been no systematic investigation of the significance levels that would result from judicious use of mod scores.

In this study we used computer simulation to study this issue. We generated data sets under six single-locus models with reduced penetrance—three dominant (D) models and three recessive (R) models—with no linkage between the disease and the marker. We then analyzed

Received July 26, 1996; accepted for publication September 26, 1996.

Address for correspondence and reprints: Dr. Susan E. Hodge, New York State Psychiatric Institute, Unit 14, 722 West 168th Street, New York, NY, 10032. E-mail: hodge@child.cpmc.columbia.edu

© 1997 by The American Society of Human Genetics. All rights reserved.
0002-9297/97/6001-0028\$02.00

the data in several ways: by (1) maximizing over penetrance only, (2) maximizing over "dominance model" (i.e., D or R), and (3) maximizing over penetrance and dominance model simultaneously. We compared the resultant significance levels with each other and with the baseline—that is, with the lower bound represented by a one-sided χ^2_1 .

Methods

Assumptions and Terminology

We considered simple single-locus D and R models with reduced penetrance (f) and no sporadics. For all models (both generating models and analysis models) we assumed Hardy-Weinberg equilibrium, random mating, and no epistasis or pleiotropy.

In what follows, the term "dominance model" refers specifically to mode of inheritance—that is, in this study, single-locus D model versus single-locus R models. The more-general term "model" (or "genetic model") encompasses both mode of inheritance and the value of f .

Generating Models

We generated data sets under three dominant (D) models (D80, D50, and D20) and three R models (R80, R50, and R20), where the number after "D" or "R" indicates the percent penetrance, $f = .8, .5, \text{ or } .2$. The disease and marker were always unlinked, since we were interested only in evaluating significance levels (P values) in this study. Gene frequency of the disease allele was always .01.

To begin, each simulation consisted of $N = 1,000$ data sets of 20 nuclear families each. For some analyses, we generated additional, $N = 1,000$ – $2,000$, data sets of 20 families each, yielding a total of $N = 1,000$ – $3,000$ data sets for each analysis. Additionally, some simulations examined data sets of 40 nuclear families each; these simulations contained 500 data sets each. The reasons for these choices are outlined below (see Discussion). The nuclear families consisted of two parents plus a variable number of children. The number of children was determined by a family-size distribution with the following conditional probabilities (conditioned on there being at least two children): $P(2) = .49$, $P(3) = .18$, $P(4) = .13$, $P(5) = .08$, and then diminishing probabilities up to the maximum sibship size, $P(10) = .009$. All matings were fully informative for the marker. Families were selected for linkage analysis on the basis of having at least two affected children. Data sets were simulated by use of our extensively tested simulation program (Greenberg 1989; Durner and Greenberg 1992), which uses a random process for each step in the simulation (picking the mating type, family size, segregation of alleles from parents to offspring, etc.).

Analysis Models

We investigated three different kinds of MMLS or mod-score analyses: (1) that in which, for a fixed dominance model, Z_{\max} was maximized over penetrance; (2) that in which, for a fixed penetrance, Z_{\max} was maximized over the two possible dominance models, D or R; and (3) that in which Z_{\max} was maximized simultaneously over dominance model and penetrance. We used LIPED (Ott 1974) to analyze linkage, with gene frequency set at .01. Z values were calculated over a grid of recombination values, $\theta = 0, .02, .04, \dots, .50$, and, for any given model, Z_{\max} was chosen along that grid.

Before beginning the study proper, we first confirmed that our simulated significance levels were following a one-sided χ^2_1 , as expected. We did this by performing linkage analyses under the correct genetic model, including penetrance. We also did confirmatory analyses under *one* wrong genetic model (both wrong dominance model and wrong penetrance), since significance levels for linkage analysis assuming any one wrong model are still asymptotically one-sided χ^2_1 (Williamson and Amos 1990). All these confirmatory analysis did yield significance levels following a one-sided χ^2_1 . We also performed preliminary analyses under a different ascertainment scheme, requiring (only) at least one affected child per family. Results were indistinguishable from those found when at least two affected children were required. Since the latter scheme is more realistic for linkage studies, these are the results that we present here. Moreover, we had no reason to expect that we would have found major differences if we had used "denser" ascertainment schemes. We now describe the three parts of the study in detail.

1. Maximizing Z_{\max} over penetrance.—In this part of the study, linkage was analyzed under *either* the right dominance model (part 1a) *or* the wrong one (part 1b), and the linkage analysis was repeated with 10 different penetrance values, $f = .1, .2, \dots, .8, .9, .99$. For each assumed penetrance value, first Z was maximized over the grid of θ values, to yield Z_{\max} for that assumed f ; then the resultant Z_{\max} values were maximized over the 10 penetrances. This max Z_{\max} was reported as the final result. (Thus, max Z_{\max} represents the largest of 10 Z_{\max} values.) All simulations using the right dominance model (part 1a) consisted of either $N = 2,000$ data sets or $N = 3,000$ data sets containing 20 nuclear families each; these simulations then were repeated for 500 data sets, containing 40 nuclear families each. The simulations using the wrong dominance model (part 1b) consisted of $N = 1,000$ data sets containing 20 families each.

2. Maximizing Z_{\max} over dominance model.—A single analysis penetrance value, $f = .5$, was chosen arbitrarily, irrespective of the generating model used, since here we were considering the effects of the dominance model only. Linkage was analyzed under both the right and wrong dominance models, both assuming $f = .5$. The

Table 1

Summary of All Analyses, Showing Which Assumptions Went into the Analysis Model and How Many Data Sets, N , Were Simulated for Each Analysis

GENERATING MODEL	ANALYSIS MODEL (N) FOR			
	Part 1a ^a	Part 1b	Part 2	Part 3
D80	D, maximized over f (3,000)	R, maximized over f (1,000)	$f = .5$, maximized over D, R (2,000)	Maximized over D, R, f (1,000)
D50	D, maximized over f (2,000)	R, maximized over f (1,000)	$f = .5$ maximized over D, R (2,000)	Maximized over D, R, f (1,000)
D20	D, maximized over f (3,000)	R, maximized over f (1,000)	$f = .5$, maximized over D, R (2,000)	Maximized over D, R, f (1,000)
R80	R, maximized over f (2,000)	D, maximized over f (1,000)	$f = .5$, maximized over D, R (2,000)	Maximized over D, R, f (1,000)
R50	R, maximized over f (2,000)	D, maximized over f (1,000)	$f = .5$, maximized over D, R (2,000)	Maximized over D, R, f (1,000)
R20	R, maximized over f (2,000)	D, maximized over f (1,000)	$f = .5$, maximized over D, R (2,000)	Maximized over D, R, f (1,000)

NOTE.—Each data set contains 20 nuclear families.

^a All analyses in part 1a also were repeated with 500 data sets of 40 families each.

larger of the two Z_{\max} values was reported as the max Z_{\max} . Each simulation consisted of $N = 2,000$ data sets of 20 nuclear families each.

3. *Maximizing Z_{\max} over penetrance and dominance model.*—Here Z_{\max} was maximized over the same penetrance values as were used in (1), as well as over the two possible dominance models. That is, the largest of 20 Z_{\max} values was reported as the max Z_{\max} value. Each simulation consisted of $N = 1,000$ data sets of 20 families each. Table 1 summarizes all the analyses—what their assumptions were and how many data sets were simulated for each.

Presentation of Results

Observed significance levels, $P(Z)$, were determined as a function of max Z_{\max} , as follows: $P(Z) \equiv (\text{number of data sets yielding max } Z_{\max} \geq Z)/N$, where N represents the number of data sets generated for that simulation. For parts 1 and 2, these significance levels were plotted versus Z and were compared with corresponding plots for a one-sided χ^2_1 , denoted " $\frac{1}{2}\chi^2_1$ " in the figures; for a two-sided χ^2_1 , denoted " χ^2_1 "; for half of a χ^2_2 , denoted " $\frac{1}{2}\chi^2_2$ "; and, for the mean of a χ^2_1 and a χ^2_2 , denoted " $\frac{1}{2}(\chi^2_1 + \chi^2_2)$." For part 3, $P(Z)$ was compared with tail probabilities from $\frac{1}{2}\chi^2_1$, χ^2_1 , and $\frac{1}{2}(\chi^2_1 + \chi^2_2)$ and with twice a χ^2_1 , denoted " $2\chi^2_1$." For the χ^2 graphs, values along the horizontal axis were transformed to a "LOD" scale by dividing by $2\ln 10 = 4.605$.

For certain Z cutoff values of interest, we report tables of our observed significance levels, with their 95% confidence intervals. These confidence intervals are based on the normal approximation to the binomial distribution, except where $N \times \text{observed significance level}$ was < 5 , in which case we calculated an exact binomial confidence

interval. These exact confidence intervals were two sided, except when the observed number equaled 0, in which case we gave a one-sided interval. We also tabulated the approximate increase in Z needed to achieve a given test size α .

Results

Maximizing Z_{\max} over Penetrance (Part 1)

For all simulations in this part of the study, the observed significance levels matched those of a two-sided χ^2_1 quite closely, within sampling variation. All curves were essentially bounded below by $\frac{1}{2}\chi^2_1$ and above by $\frac{1}{2}\chi^2_2$. These observations held whether data were analyzed under the right dominance model (part 1a) or the wrong one (part 1b) and whether data sets consisted of 20 or 40 families each. Figure 1 shows significance levels from two representative simulations (D80 and R80) in part 1a, and figure 2 shows two representative simulations (D20 and R50) from part 1b. Thus, significance levels were approximately doubled when Z_{\max} was maximized over penetrance values, within either the right or the wrong dominance model, compared with the baseline.

Certain Z cutoff values are of particular interest. The values $Z = 0.59, 0.83, 1.17$, and 1.44 correspond to tests of size $\alpha = .05, .025, .01$, and $.005$, respectively, on the basis of a one-sided χ^2 test with 1 df. Although few linkage researchers would use the lower of these Z values, it is nevertheless interesting to compare α with the observed significance levels for these Z values. In addition, the values $Z = 2.0$ and $Z = 3.0$ are routinely used by human geneticists to indicate evidence for linkage. Tables 2 and 3 give observed significance levels,

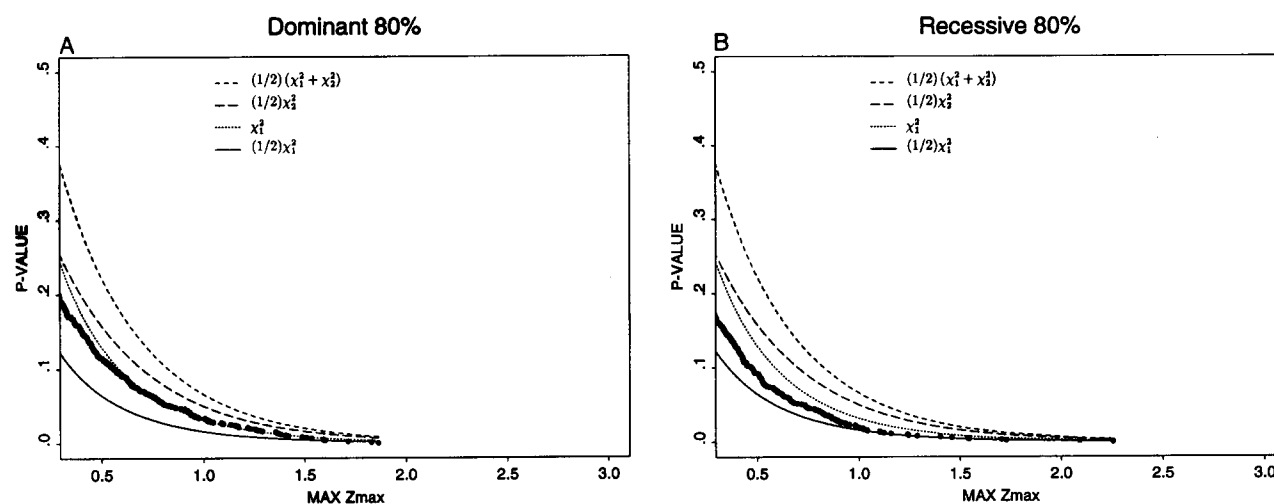


Figure 1 Simulated significance levels for two representative simulations from part 1a, maximizing over penetrance, analyzed under correct dominance model (table 2). Results are compared with tail probabilities from four χ^2 curves, as labeled. A, Generating model D80; $N = 3,000$ data sets. B, Generating model R80; $N = 2,000$ data sets.

along with 95% confidence intervals, for these cutoff values of Z . The results in the tables support the conclusion from the figures—that significance levels are approximately doubled by maximizing Z_{\max} over penetrance, within a single dominance model.

From another viewpoint, how much would an investigator have to add, in LOD-score units, to the Z value used as a test criterion, in order to achieve a desired test size α ? To approach this question, we took the two-sided χ_1^2 as the working approximation, and we estimated χ^2 values corresponding to selected values of α . (For example, for $\alpha = .01$, the one-sided χ^2 is 5.4, which corresponds to $Z \approx 1.17 = 5.4/(2\ln 10)$, whereas the two-sided χ^2 is 6.63 ($\Rightarrow Z \approx 1.44$). Thus, in this exam-

ple, the increase Δ needed for Z is $.27 = 1.44 - 1.17$.) Table 4 shows Δ values ranging from .24 when $\alpha = .05$ to .29 when $\alpha = .0001$. Thus, for Z values of interest (up to $\sim Z = 3.0$), it appears that adding 0.3 to the Z value used as a test criterion is conservative.

Maximizing Z_{\max} over Dominance Models (Part 2)

When Z_{\max} was maximized over D versus R, for a fixed value of penetrance, significance levels were increased over baseline—but, in most cases, less than in part 1. When the generating model was D with high penetrance ($f = .8$), significance levels were approximately doubled, but, for lower penetrances of generating models and for R generating models, significance levels

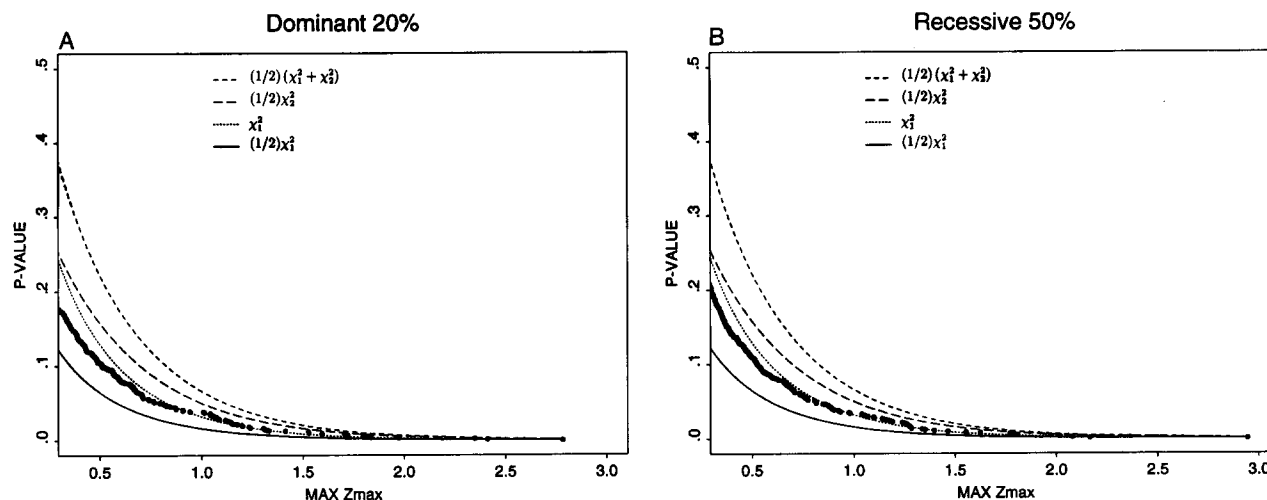


Figure 2 Simulated significance levels for two representative simulations from part 1b, maximizing over penetrance, analyzed under wrong dominance model (table 3). Comparison χ^2 curves are same as in figure 1. A, Generating model D20; $N = 1,000$ data sets. B, Generating model R50; $N = 1,000$ data sets.

Table 2

Observed Significance Levels and Associated 95% Confidence Intervals for Analyses from Part 1a, with Maximization over Penetrance, under Correct Dominance Model (Figure 1)

α^a	Z	P (95% CONFIDENCE INTERVAL) FOR GENERATING MODEL		
		D80 (N = 3,000)	D50 (N = 2,000)	D20 (N = 3,000)
Generated dominant, analyzed dominant:				
.05				
.025	.59	.095 (.084-.105)	.087 (.074-.099)	.102 (.091-.113)
.01	.83	.051 (.043-.059)	.045 (.036-.054)	.056 (.048-.064)
.005	1.17	.023 (.017-.028)	.020 (.014-.026)	.026 (.020-.032)
≈.001	1.44	.009 (.006-.012)	.012 (.007-.017)	.017 (.012-.021)
≈.0001	2.0	0 (0-.0010) ^b	.002 (.0005-.0051) ^b	.003 (.001-.005)
	3.0	0 (0-.0010) ^b	0 (0-.0015) ^b	0 (0-.0010) ^b
		R80 (N = 2,000)	R50 (N = 2,000)	R20 (N = 2,000)
Generated recessive, analyzed recessive:				
.05				
.025	.59	.070 (.059-.081)	.070 (.059-.081)	.066 (.055-.077)
.01	.83	.038 (.029-.046)	.044 (.035-.052)	.039 (.031-.048)
.005	1.17	.011 (.006-.016)	.021 (.015-.027)	.018 (.012-.023)
≈.001	1.44	.007 (.003-.011)	.011 (.006-.015)	.011 (.006-.015)
≈.0001	2.0	.002 (.0005-.0051) ^b	.003 (.001-.005)	.003 (.001-.005)
	3.0	0 (0-.0015) ^b	0 (0-.0015) ^b	0 (0-.0015) ^b

^a Baseline values.

^a Baseline values.

^b Exact binomial confidence interval.

Table 3

Observed Significance Levels and Associated 95% Confidence Intervals for Analyses from Part 1b, with Maximization over Penetrance, under Wrong Dominance Model (Figure 2)

		P (95% CONFIDENCE INTERVAL) FOR GENERATING MODEL		
α^a	Z	D80	D50	D20
Generated dominant, analyzed recessive:				
.05	.59	.082 (.065-.099)	.092 (.074-.110)	.085 (.068-.102)
.025	.83	.046 (.033-.059)	.053 (.039-.067)	.046 (.033-.059)
.01	1.17	.019 (.011-.028)	.022 (.013-.031)	.022 (.013-.031)
.005	1.44	.014 (.007-.021)	.009 (.003-.015)	.012 (.005-.019)
$\approx .001$	2.0	.005 (.001-.009)	.001 (.0000-.0056) ^b	.003 (.0006-.0087) ^b
$\approx .0001$	3.0	0 (0-.0030) ^b	0 (0-.0030) ^b	0 (0-.0030) ^b
		R80	R50	R20
Generated recessive, analyzed dominant:				
.05	.59	.092 (.074-.110)	.085 (.068-.102)	.100 (.081-.119)
.025	.83	.048 (.035-.061)	.048 (.035-.061)	.054 (.040-.068)
.01	1.17	.021 (.012-.030)	.025 (.015-.035)	.024 (.015-.033)
.005	1.44	.014 (.007-.021)	.011 (.005-.017)	.013 (.006-.020)
$\approx .001$	2.0	.002 (.0002-.0072) ^b	.004 (.0011-.0102) ^b	.003 (.0006-.0087) ^b
$\approx .0001$	3.0	.002 (.0002-.0072) ^b	0 (0-.0030) ^b	0 (0-.0030) ^b

NOTE.—N = 1,000 for all analyses.

NOTE.—N = 1,000 for all analyses.

^a Baseline values.

^b Exact binomial confidence interval.

Table 4

Approximate New Values of Z to Be Used as Test Criterion, and Δ in Z , for Given Values of α

α	Z^a	NEW Z (Δ) FOR	
		Parts 1 and 2	Part 3
.05	.59	.83 (.24)	1.09 (.50)
.025	.83	1.09 (.26)	1.35 (.52)
.01	1.17	1.44 (.27)	1.71 (.54)
.005	1.44	1.71 (.27)	1.99 (.55)
$\approx .001$	2.00	2.28 (.28)	2.56 (.56)
$\approx .0001$	3.00	3.29 (.29)	3.58 (.58)

^a Baseline values.

^b For parts 2 and 3 these guidelines are very conservative; see Discussion.

were only ~ 1.5 times higher than baseline. (On theoretical grounds, the *maximum* that asymptotic significance levels could be increased in this part of the study is two times; see Discussion.) Table 5 gives observed significance levels and 95% confidence intervals for all simulations, and figure 3 shows graphs of significance levels for two representative simulations (D80 and R80). To answer the question of how much to increase Z for a given α , an investigator could use the same guidelines as are used for part 1, on the basis of approximately doubled significance levels (table 4), but should recognize that, for part 2, these guidelines are more conservative than they were for part 1.

Maximizing Z_{\max} over Penetrance and Dominance Model (Part 3)

When Z_{\max} was maximized over the two dominance models, D and R, and over penetrance within each model, significance levels were increased approximately two to three times over baseline, although somewhat more when the generating model was D with high penetrance. On theoretical grounds, the maximum to which asymptotic significance levels could rise would be the sum of significance levels from parts 1a and 1b—that is, not more than approximately four times over baseline (see Discussion). Table 6 gives the results, and figure 4 shows graphs of significance levels for D50 and R50. On the basis of this “4 \times ” guideline, an investigator would increase Z (up to $Z = 3.0$) by an increment of 0.50–0.58 (table 4). However, this guideline is very conservative (see Discussion).

Discussion

Summary of Findings

In parts 1 and 2, we maximized Z_{\max} over either penetrance model or dominance model but not over both. Significance levels were approximately doubled in part 1 and were, at most, doubled in part 2, compared with baseline significance levels from a one-sided χ^2_1 . This result corresponds roughly to going from a one- to a two-sided χ^2_1 . A doubling of significance levels corresponds to increasing Z by ~ 0.24 – 0.29 (table 4). Thus, for maximizing over penetrance (part 1), increasing Z by 0.3 would be a conservative course of action in this

Table 5

Observed Significance Levels and Associated 95% Confidence Intervals for Analyses from Part 2, with Maximization over Dominance Model Only (Figure 3)

α^a	Z	P (95% CONFIDENCE INTERVAL) FOR GENERATING MODEL		
		D80	D50	D20
.05	.59	.100 (.087–.113)	.082 (.070–.094)	.074 (.063–.085)
.025	.83	.050 (.040–.060)	.047 (.038–.056)	.041 (.032–.050)
.01	1.17	.017 (.011–.023)	.018 (.012–.024)	.016 (.011–.022)
.005	1.44	.008 (.004–.012)	.012 (.007–.017)	.008 (.004–.012)
$\approx .001$	2.0	.002 (.0005–.0051) ^b	.0015 (.0003–.0044) ^b	.003 (.001–.005)
$\approx .0001$	3.0	0 (0–.0015) ^b	0 (0–.0015) ^b	0 (0–.0015) ^b
α^a	Z	R80	R50	R20
.05	.59	.068 (.057–.079)	.068 (.057–.079)	.065 (.054–.076)
.025	.83	.039 (.031–.048)	.032 (.024–.040)	.035 (.027–.043)
.01	1.17	.016 (.011–.022)	.017 (.011–.023)	.015 (.010–.020)
.005	1.44	.010 (.006–.014)	.008 (.004–.012)	.010 (.006–.014)
$\approx .001$	2.0	.003 (.001–.005)	.004 (.001–.007)	.004 (.001–.007)
$\approx .0001$	3.0	.0005 (.0000–.0028) ^b	.001 (.0001–.0036) ^b	0 (0–.0015) ^b

NOTE.— $N = 2,000$ for all analyses.

^a Baseline values.

^b Exact binomial confidence interval.

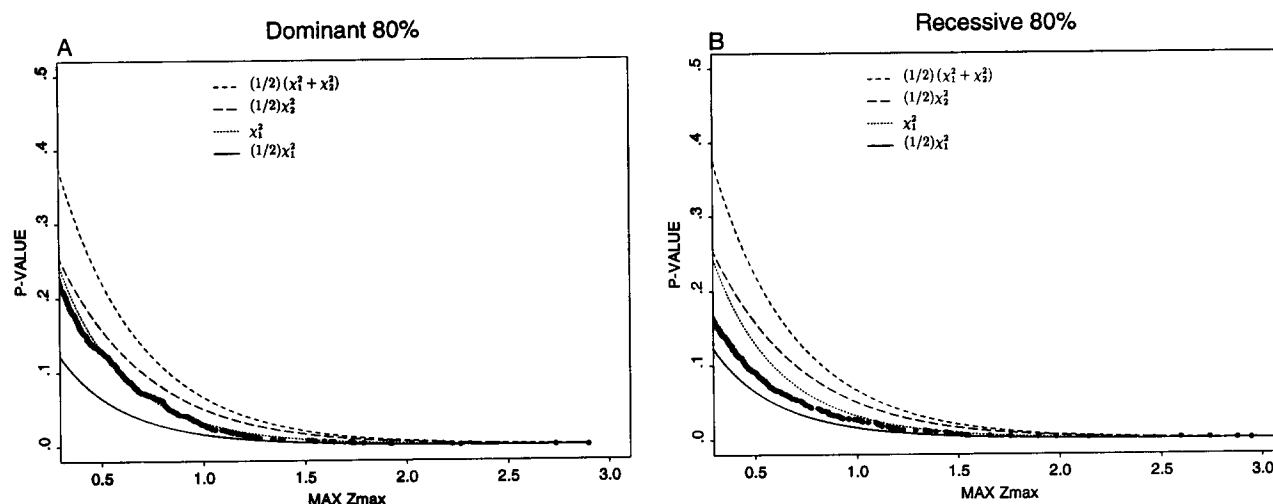


Figure 3 Simulated significance levels for two representative simulations from part 2, maximizing over dominance model (table 5). Comparison χ^2 curves are same as in figure 1. A, Generating model D80; $N = 2,000$ data sets. B, Generating model R80; $N = 2,000$ data sets.

range of Z values (up to $\approx Z = 3.0$) and would be even more conservative for analyses maximizing only over dominance model (part 2); also see below.

In part 3, Z_{\max} was maximized over dominance model and over penetrance, so significance levels were higher than in either part 1 or part 2, rising to approximately two to three times baseline values—or even slightly higher, when the true generating model was D with high penetrance. For this part of the study it is more difficult

to give simple guidelines concerning how much to raise the cutoff value of Z for a given test size α . However, very conservative upper limits of 0.50–0.58 LOD units are shown in table 4, based on the $4\times$ upper limit on significance levels that has been described above in Results.

We did not perform true continuous maximization of Z values; rather, we used a grid of θ values for finding Z_{\max} and a grid of f values for maximizing Z_{\max} in parts

Table 6

Observed Significance Levels and Associated 95% Confidence Intervals for Analyses from Part 3, with Maximization over Dominance Model and over Penetrance (Figure 4)

α^a	Z	P (95% CONFIDENCE INTERVAL) FOR GENERATING MODEL		
		D80	D50	D20
		R80	R50	R20
.05	.59	.170 (.147–.193)	.153 (.131–.175)	.131 (.110–.152)
.025	.83	.092 (.074–.110)	.090 (.072–.108)	.078 (.061–.095)
.01	1.17	.039 (.027–.051)	.033 (.022–.044)	.036 (.025–.048)
.005	1.44	.025 (.015–.035)	.020 (.011–.029)	.015 (.008–.023)
$\approx .001$	2.0	.006 (.001–.011)	.006 (.001–.011)	.003 (.0006–.0087) ^b
$\approx .0001$	3.0	0 (0–.0030) ^b	0 (0–.0030) ^b	0 (0–.0030) ^b
		R80	R50	R20
.05	.59	.104 (.085–.123)	.107 (.088–.126)	.086 (.069–.103)
.025	.83	.062 (.047–.077)	.067 (.052–.082)	.055 (.041–.069)
.01	1.17	.023 (.014–.032)	.030 (.019–.041)	.027 (.017–.037)
.005	1.44	.015 (.008–.023)	.014 (.007–.021)	.020 (.011–.029)
$\approx .001$	2.0	.004 (.0011–.0102) ^b	.002 (.0002–.0072) ^b	.006 (.001–.011)
$\approx .0001$	3.0	.002 (.0002–.0072) ^b	.001 (.0000–.0056) ^b	.002 (.0002–.0072) ^b

NOTE.— $N = 1,000$ for all analyses.

^a Baseline values.

^b Exact binomial confidence interval.

1 and 3. However, this approximation should not affect our findings noticeably. The θ grid was quite fine (increments of .02). The f grid was coarser (increments of .1); however, on the basis of both previous work (Greenberg 1989, 1990) and our observations in this study, Z_{\max} as a function of assumed penetrance is relatively flat. Moreover, as mentioned earlier, our confirmatory studies agreed with the predicted baseline significance levels (results not shown but available on request).

Of the total of 38,000 20-family data sets in the whole study, only 1 yielded a Z value >4.0 , and only 10 yielded Z values >3.0 . (The highest Z observed over all maximizations was 4.89; the next highest was 3.76.) Therefore, we do not attempt to give significance levels for Z values >3.0 ; our sample sizes are too small.

It is well known, on the basis of common experience (or see Clerget-Darpoux et al. 1986), that, in a linkage analysis, moderate changes in gene frequency have little impact on Z values. (This is in contrast to the potential impact of gene frequency on estimates of θ ; e.g., see Clerget-Darpoux et al. 1986.) Since we were interested in Z values here, not in estimating θ , we used just one gene frequency in this study.

How generalizable are our simulation results? It is true that we did not consider any generating models other than D and R with high and low penetrances. However, we do not believe that *significance levels* (as opposed to power) would change much if the true model were, say, intermediate or two-locus but were analyzed as D or as R (e.g., see Greenberg 1990). Also, our previous work has shown that two-locus models can be adequately approximated by an appropriate single-locus linkage analysis (Vieland et al. 1992a, 1992b, 1993). But also of significance is that our results in the present study were very similar when we analyzed data under the *wrong* model (table 3). Thus, although it is possible that the guidelines that we have derived here ultimately may prove inaccurate for other genetic models, until further work is done these guidelines should serve as workable rules.

Theoretical Issues

The situation being considered here represents a difficult one in statistics—since it is a situation in which (at least) one parameter which plays a role in the alternative hypothesis is not defined under the null hypothesis. In our application, penetrance is that parameter. This complication arises because, under the null hypothesis of “no linkage,” there is no information about penetrance (Hodge and Elston 1994).

Intuitively, for our investigations in part 1, we might expect the df to fall, in some sense, “between” the df for a likelihood ratio (LR) test with one unspecified parameter and the df for an LR test with two unspecified parameters; that is, if the alternative hypothesis on θ were two-sided, we would expect the distribution of the LR statistic to lie

between χ^2_1 and χ^2_2 . In our case, since the alternative is one-sided ($\theta < 1/2$), we would expect a distribution between $1/2\chi^2_1$ and $1/2(\chi^2_1 + \chi^2_2)$. However, even this intuition may not hold as a general statistical principle, although in fact none of our analyses in part 1 exceeded $1/2(\chi^2_1 + \chi^2_2)$ —or even came close; see figures 1 and 2. We now examine these theoretical issues further, first from the viewpoint of correlated versus independent statistical analyses, then by examining other approaches.

Correlated versus independent analyses.—As noted, the one-sided χ^2_1 represents the baseline to which we compare our findings. A priori, the significance levels in parts 1 and 2 must be at least as great as baseline, since, in both part 1 and part 2, the max Z_{\max} cannot be less than the Z_{\max} obtained from analysis under a single genetic model. By similar reasoning, significance levels in part 3 must be at least as high as the greater of those in parts 1 and 2, since, in part 3, Z_{\max} was maximized over both penetrance and dominance model but, in parts 1 and 2, was maximized over only one or the other.

However, it is not a priori obvious how much higher the significance levels would be for any of the three parts of this study. By maximizing Z_{\max} over dominance model and/or over penetrance, are we adding the equivalent of 1 df? the equivalent of $1/2$ df? This is the question that our study attempted to answer.

One way to think about these issues is to ask to what extent different analyses performed on the same data set are *correlated*. Consider the case in which an investigator performs two analyses on the data set, as in part 2. Let p represent the baseline significance level associated with the one-sided χ^2_1 . If the two analyses were independent, the new asymptotic significance level would be $2p - p^2$; if the two analyses were negatively correlated, the *maximum possible* would be $2p$. A priori, we expect the two analyses in part 2 to be positively correlated, which is why the observed significance levels tend to be $<2p$ (or $2p - p^2$, which is indistinguishable from $2p$). If anything, it is surprising, for D models with high penetrance, that the significance levels are as close to $2p$ as we have observed. This seems to imply that, for these generating models and with our ascertainment scheme (families with at least two affected children), many families/data sets yield positive evidence for linkage when they are analyzed under one dominance model but yield negative or no evidence for linkage when they are analyzed under the other. This situation would lead to uncorrelated (or even negatively correlated) linkage analyses in part 2. This also would explain why our observed significance levels are higher for these same generating models in the analyses in part 3.

As mentioned, we can put an upper bound of $2p$ on the (asymptotic) significance levels in part 2, where p represents the baseline significance level. However, in part 1, where we maximize over 10 f values for each model, this kind of reasoning is not helpful. These 10

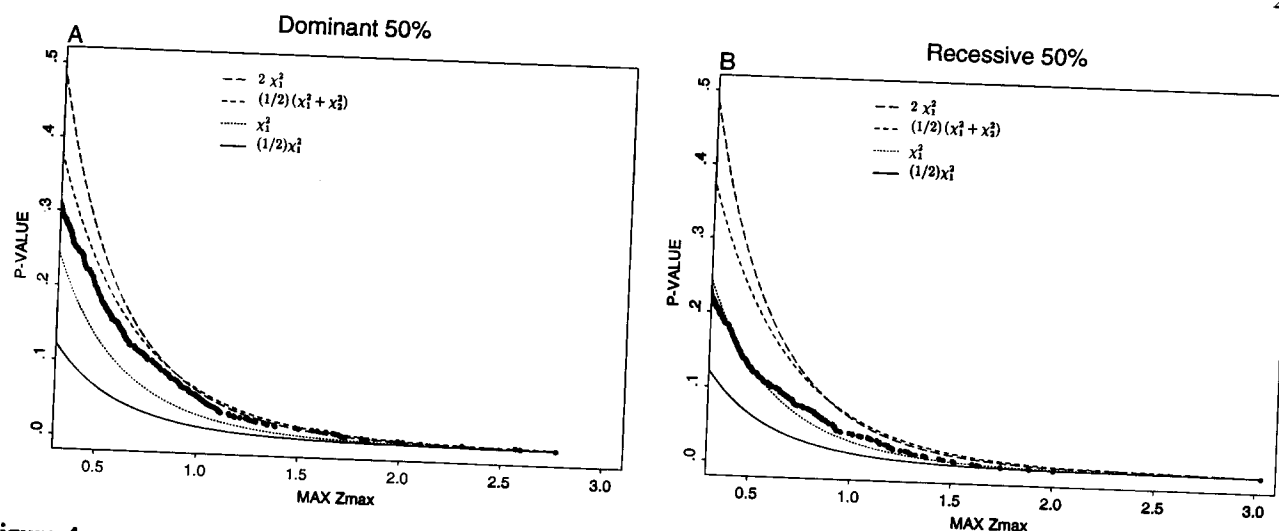


Figure 4 Simulated significance levels for two representative simulations from part 3, maximizing over penetrance and dominance model (table 6). Note that the comparison χ^2 curves are different from those in the other figures. A, Generating model D50; $N = 1,000$ data sets. B, Generating model R50; $N = 1,000$ data sets.

values of f represent an approximation to maximizing over a continuous range of f values, and clearly these analyses are highly positively correlated with each other.

Part 3 represents a "combination" of the other two parts, in the sense that we are maximizing over both dominance model and penetrance. However, the analyses in part 3 also can be viewed as simply taking the maximum of the two $\max Z_{\max}$'s from parts 1a and 1b. Hence, the significance level for part 3 cannot exceed the sum of those from parts 1a and 1b, by the same reasoning as has been used for part 2 above. Since significance levels in parts 1a and 1b were, at most, approximately doubled, we conclude that those in part 3 will be increased, at most, $4\times$. In fact, taking $4\times$ as a guideline is quite conservative, as can be seen in table 6 and figure 4 (see the tail probabilities for the $2\chi^2$ curve). The guidelines for how much to raise the cutoff value of Z (table 4) also are based on this reasoning and hence also are very conservative.

Other theoretical approaches.—Davies (1977, eq. [3.7]) has considered this general statistical problem and has derived an asymptotic upper bound on the significance level of such a test. This result is a complex formula that must be tediously evaluated for any particular application. D. Rabinowitz (personal communication) has calculated this upper bound for a simple genetic model (D with reduced penetrance), in triads consisting of one phase-unknown parent and two sibs.

These theoretical results of Davies and Rabinowitz are asymptotic, and it is reasonable to ask how large a data set we would need in order to use asymptotic properties of the LR statistic. In part 1 of this study we examined both 20-family and 40-family data sets. The resultant distributions of significance levels were virtually identical to each other, implying that a data set of

20 nuclear families of varying sizes is reasonably close to asymptoticity for this kind of analysis. Hence we considered only 20-family data sets subsequently. (The 40-family results are not shown in the tables or figures but are available on request.)

In part 1a of this study, we generated larger numbers of data sets ($N = 2,000$ – $3,000$) in order to produce confidence intervals narrower than what we would get from $N = 1,000$ data sets. However, once we saw the pattern that the results were following, $N = 1,000$ seemed sufficient for our analyses in part 1b. The runs are time-consuming, and we did not wish to perform more than were necessary to answer the question being posed. For part 2, we simulated $N = 2,000$ data sets for each generating model, to confirm the observed pattern whereby significance levels were higher for data sets generated under D80 than for those generated under the other models. For part 3, $N = 1,000$ data sets seemed sufficient.

Practical Guidelines for Investigators

When analyzing complex genetic diseases for linkage, the investigator often does not know the mode of inheritance of the disease being studied. In this situation, several approaches are available. One can choose a single genetic model (perhaps based on evidence from family studies or on a segregation analysis) and perform linkage analysis based on that model. One can turn to affected-sib-pair, affecteds-only, and other "nonparametric" methods. Or one can perform linkage analysis under the assumption of two or more genetic models, which allows the option of more intensive data exploration, as we have recommended elsewhere (Hodge et al. 1993). Each approach (and there are others, as well, not discussed here) has its advantages and disadvantages. Our focus

here is on the third course of action, calculating Z values under multiple models. Previous research has shown that, when the mode of inheritance is unknown, analyzing data under different mode-of-inheritance assumptions has more power to detect linkage than does analyzing data under a single genetic model (e.g., see Greenberg and Hodge 1989; Greenberg 1990; Vieland et al. 1992b; Greenberg and Berger 1994). Taking this approach has the advantage of using more of the genetic information available in a data set, thereby not only increasing the chances of detecting linkage but also yielding other information about the trait or disease, such as mode of inheritance and penetrance. The disadvantage of this approach—that is, increasing the type I error—is what we have attempted to quantify in this study.

Several investigators have considered some aspects of the increase in type I error when linkage is analyzed under several models. Weeks et al. (1990) presented a computer-simulation method for evaluating the inflation in Z after Z_{\max} has been maximized over *diagnostic schemes* and/or over penetrances. They presented rough guidelines for “deflating” a Z of 3.0 by 0.3–1.0 units but commented that “the bulk of the lod score inflation is apparently due to maximization over diagnostic schemes, rather than over penetrances” (Weeks et al. 1990, p. 242). Unfortunately, other investigators have not preserved this distinction and have concluded from the Weeks et al. work that they must deflate a Z of 3.0 by a whole LOD-score unit when maximizing over penetrance. As we have shown here, however, even the 0.3 figure is already conservative if the user maximizes only over penetrance or only over dominance model, and a figure of 0.6 is *very* conservative if one maximizes over both penetrance and dominance model. (It is also important to note that the Weeks et al. work was not a general simulation study but was based on a particular published data set [Sherrington et al. 1988] involving schizophrenia.) Clerget-Darpoux et al. (1990) focused on what happens when Z_{\max} is maximized simultaneously over penetrance, diagnostic classification, and multiple laboratories. They considered a single three-generation pedigree, with a fixed distribution of affected individuals within this pedigree, and they examined only analyses under dominant modes of inheritance. They found that type I errors were greatly increased when all these multiple analyses were performed. Risch (1991) derived a conservative correction (i.e., deflate Z_{\max} by $\log_{10} t$, where t is number of models) based on treating all analyses as independent; however, as discussed above, this is excessively conservative.

To some extent, the goal of these cited studies is to provide a post hoc warning to readers, a warning with which we strongly agree. As Clerget-Darpoux et al. (1990, p. 245) say, “the significance of a LOD-score value of 3 is very difficult to assess . . . multiple tests

may have been performed, voluntarily or otherwise.” In other words, when one reads a linkage study reporting a “significant” Z , in which the investigators have maximized Z values over multiple factors, the reader has no reliable way to evaluate the finding. However, our goal in this study is different. We are attempting to provide investigators with guidelines to use *before* they undertake their study. We do not advocate maximizing Z values indiscriminately over all possible factors; rather, we propose maximizing over a well-defined parameter, such as penetrance, and/or maximizing over dominance model. Our results here enable the investigator to accurately evaluate the increase in significance level that will result from such a course of action.

In summary, then, for Z values ≤ 3.0 and with the potential limitations of our simulation studies kept in mind:

1. For maximizing only over penetrance, significance levels are approximately doubled. To achieve a given probability of type I error, increase Z by ~ 0.3 LOD units (conservative).
2. For maximizing only over two dominance models (D vs. R), significance levels are increased by one and one-half to two times. Use the 0.3-LOD-unit guideline above, but note that here it is very conservative.
3. For maximizing simultaneously over penetrance and two dominance models, significance levels are increased by 2–3 times in most cases—and certainly not $>4\times$. Increasing Z by 0.6 LOD units is very conservative.

A prudent investigator may decide that it is worth maximizing Z_{\max} over penetrance and/or over dominance model, because these increases in significance levels (P values) are relatively modest, in return for the higher probability of detecting linkage.

Acknowledgments

We thank Dr. Daniel Rabinowitz for several helpful discussions. This work was supported in part by the National Institutes of Health (grants MH-48858, MH-28274, MH-36197, and DK-31813, all to S.E.H.; and grants DK-31775 and NS-27941, both to D.A.G.) and by the Junta Nacional de Investigação Científica e Tecnológica of Portugal (support to P.C.A.).

References

- Clerget-Darpoux F, Babron MC, Bonaïti-Pellié C (1990) Assessing the effect of multiple linkage tests in complex diseases. *Genet Epidemiol* 7:245–253
- Clerget-Darpoux F, Bonaïti-Pellié C (1992) Strategies on marker information for the study of human diseases. *Ann Hum Genet* 56:145–153
- Clerget-Darpoux F, Bonaïti-Pellié C, Hochez J (1986) Effects

- of misspecifying genetic parameters in lod score analysis. *Biometrics* 42:393-399
- Davies RB (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64:246-254
- Durner M, Greenberg DA (1992) Effect of heterogeneity and assumed mode of inheritance on lod scores. *Am J Med Genet* 42:271-275
- Elston RC (1989) Man bites dog? the validity of maximizing lod scores to determine mode of inheritance. *Am J Med Genet* 34:487-488
- Greenberg DA (1989) Inferring mode of inheritance by comparison of lod scores. *Am J Med Genet* 34:480-486
- (1990) Linkage analysis assuming a single-locus mode of inheritance for traits determined by two loci: inferring mode of inheritance and estimating penetrance. *Genet Epidemiol* 7:467-479
- Greenberg DA, Berger B (1994) Using lod-score differences to determine mode of inheritance: a simple, robust method even in the presence of heterogeneity and reduced penetrance. *Am J Hum Genet* 55:834-840
- Greenberg DA, Hodge SE (1989) Linkage analysis under "random" and "genetic" reduced penetrance. *Genet Epidemiol* 6:259-264
- Greenberg DA, Hodge SE, Vieland VJ, Spence MA (1996) Affecteds-only linkage methods are not a panacea. *Am J Hum Genet* 58:892-895
- Hodge SE, Durner M, Vieland VJ, Greenberg DA (1993) Better data analysis through data exploration. *Am J Hum Genet* 53:775-776
- Hodge SE, Elston RC (1994) Lods, wrods, and mods: the interpretation of lod scores calculated under different models. *Genet Epidemiol* 11:329-342
- Knapp M, Seuchter SA, Baur MP (1994) Linkage analysis in nuclear families. II. Relationship between affected sib-pair tests and lod score analysis. *Hum Hered* 44:44-51
- MacLean CJ, Bishop DT, Sherman SL, Diehl SR (1993) Distribution of lod scores under uncertain mode of inheritance. *Am J Hum Genet* 52:354-361
- Ott J (1974) Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. *Am J Hum Genet* 26:588-597
- Risch N (1991) A note on multiple testing procedures in linkage analysis. *Am J Hum Genet* 48:1058-1064
- Sherrington R, Brynjolfsson J, Petursson H, Porter M, Dudleston K, Barraclough B, Wasmuth J, et al (1988) Localization of a susceptibility locus for schizophrenia on chromosome 5. *Nature* 336:164-167
- Vieland VJ, Greenberg DA, Hodge SE (1993) Adequacy of single-locus approximations for linkage analyses of oligogenic traits: extension to multigenerational pedigree structures. *Hum Hered* 43:329-336
- Vieland V, Greenberg DA, Hodge SE, Ott J (1992a) Linkage analysis of two-locus diseases under single-locus and two-locus analysis models. In: MacCluer JW, Chakravarti A, Cox D, Bishop DT, Bale SJ, Skolnick MH (eds) *Genetic Analysis Workshop 7: Issues in Gene Mapping and Detection of Major Genes*. Cytogenet Cell Genet 59:145-146
- Vieland VJ, Hodge SE, Greenberg DA (1992b) Adequacy of single-locus approximations for linkage analyses of oligogenic traits. *Genet Epidemiol* 9:45-59
- Weeks DE, Lehner T, Squires-Wheeler E, Kaufman C, Ott J (1990) Measuring the inflation of the lod score due to its maximization over model parameter values in human linkage analysis. *Genet Epidemiol* 7:237-243
- Williamson JA, Amos CI (1990) On the asymptotic behavior of the estimate of the recombination fraction under the null hypothesis of no linkage when the model is misspecified. *Genet Epidemiol* 7:309-318