Original Paper

Human Heredity

Hum Hered 1996;46:265–273

Phenocopies versus Genetic Heterogeneity: Can We Use Phenocopy Frequencies in Linkage Analysis to Compensate for Heterogeneity?

Martina Durner[a,c]
David A. Greenberg[a,b]
Susan E. Hodge[d]

[a] Departments of Psychiatry and
[b] Biomathematics, Mount Sinai Medical Center, New York, N.Y., USA,
[c] Clinic of Epileptology, University of Bonn, Germany,
[d] Departments of Psychiatry and Biostatistics, Columbia University and New York State Psychiatric Institute, New York, N.Y., USA

Abstract

In this study we explore whether a *phenocopy frequency* (defined as a 'penetrance' for nondisease genotypes) can approximate or model genetic heterogeneity in a single-locus analysis. We simulated two types of heterogeneity situations: 'sporadic models', where there are two forms of a disease, one genetic and linked to a marker and the other purely random, and 'genetic heterogeneity models', where the disease is caused by either of two different loci, one linked to the marker and the other unlinked. We analyzed simulated data sets for linkage, assuming a single-locus analysis with varying phenocopy frequency, in analogy with earlier work on epistatic two-locus models. We found that in the presence of purely random sporadics, there was a difference between assuming *any* nonzero phenocopy frequency and a zero frequency, but that the actual value of the assumed phenocopy frequency had little effect on the maximum lod score. In contrast, when both forms of disease are genetic, and are generated under similar genetic parameters, assuming a positive phenocopy frequency will not, in general, compensate for the presence of the unlinked form. However, when the modes of inheritance of the two forms differ, the assumption of a nonzero phenocopy frequency does have an effect, either to increase or decrease the maximum lod score, depending on the modes of inheritance of the two disease forms. We conclude with practical recommendations for investigators, based on these results.

Key Words
Sporadics
Heterogeneity
Maximized maximum lod score
Mod scores
Phenocopies

KARGER
E-Mail karger@karger.ch
Fax + 41 61 306 12 34
http://www.karger.ch

© 1996 S. Karger AG, Basel
0001–5652/96/0465–0265$10.00/0

Dr. Martina Durner
Department of Psychiatry, Box 1229
Mount Sinai Medical Center
1 Gustave Levy Place
New York, NY 10029 (USA)

Received: January 31, 1995
Revision received: December 12, 1995
Accepted: January 11, 1996

Introduction

While linkage analysis provides a powerful tool to detect the chromosomal location of a genetic disease and has proven successful in detecting genes for single-locus mendelian disorders, it has been less successful when used to look for disease loci for the so-called common complex diseases, such as epilepsy, schizophrenia, etc. These are conditions in which there is no clear mode of inheritance and where confounding factors, such as reduced penetrance, heterogeneity and phenocopies, are thought to exist.

The work we report here is part of an ongoing effort to determine optimal approaches to analyzing linkage data from the common complex diseases. Following up our earlier work on the problem of heterogeneity [1–3] we wanted to test the notion that one could compensate for heterogeneity in a linkage analysis by varying the assumed phenocopy frequency.

In this work, we consider two (out of many possible) heterogeneity models – a 'sporadic' model and a 'genetic heterogeneity' model (to be defined more precisely under Methods). These two kinds of heterogeneity lead to entirely different ways of interpreting a phenocopy frequency assumed in a linkage analysis, as we will show.

Our motivation for this work arose from our earlier work on epistatic models [4–7]. There, we had shown that when a disease is controlled by two or more mendelian loci acting epistatically, this disease can be modeled in a linkage analysis as a single-locus mendelian disease with reduced penetrance. That is, 'reduced penetrance' can be used to approximate or mimic the action of a second, epistatic locus. In the current work we hoped to draw an analogy between the epistatic and heterogeneity situations. Just as assuming reduced penetrance can compensate for the action of the second locus in the epistatic case, we wondered whether a 'phenocopy frequency' (to be defined below) might be able to mimic the action of a second, unlinked locus in a genetic heterogeneity model (also see the Discussion).

Before approaching the above question, as a control, we also explored the effect of varying the phenocopy frequency in a linkage analysis when the *true* model is genetic plus sporadic cases. We wanted to determine how accurately one could estimate the phenocopy frequency using the maximized maximum lod score or mod score (MMLS) method [1, 8–11], i.e., maximizing the maximum lod score over different values of the phenocopy frequency.

When we assumed a 'sporadic model' to analyze data generated from a 'sporadic model', we found that maximum lod scores were extremely insensitive to phenocopy frequencies used to analyze the simulated data; the only thing that mattered was whether one allowed for the existence of phenocopies at all. We found that we could not estimate a phenocopy frequency accurately using MMLS with the data set sizes that we used (40 nuclear families). Secondly, when we used a 'sporadic model' to analyze data generated under *genetic heterogeneity,* we found that in general this approach did not compensate very well for genetic heterogeneity in a linkage analysis. We found that there are also circumstances in which it is better *not* to assume a positive phenocopy frequency – even in the presence of genetic heterogeneity. We will conclude with practical recommendations for the investigator.

Methods

We define the *phenocopy frequency,* g = P[affected | nondisease genotype]. That is, g represents a 'penetrance function' for the genotype aa (if the disease

is dominant) or for the genotypes Aa and AA (if the disease is recessive). This is not to be confused with the phenocopy rate, usually defined as P[nondisease genotypes | affected], i.e., the proportion of sporadic cases *among all* affected individuals in the population [12].

We describe first the generating models under which we create the data. then the models used for analysis.

Generating Models

Family data were generated by computer simulation, with two clinically indistinguishable forms of disease in the population. For all cases, *one* form was caused by a locus tightly linked to the marker, but the second form differed for the 'sporadic' and 'genetic heterogeneity' models (see below). We used our simulation program [3, 13] to generate data under both 'sporadic' and 'heterogeneity' models, as follows:

Sporadic Models. Under these models, the second form of disease was sporadic, or purely random. Individuals having this second form are called 'sporadic cases'. The probability of having this form is strictly random, i.e. not correlated within families. The phenocopy frequency g defined above gives the probability that an individual who is not at genetic risk will have this form of the disease.

In the simulations, the linked (i.e., first) form of the disease is autosomal dominant (for D) or autosomal recessive (for R) with full penetrance, and a disease allele frequency of 0.005 (D) or 0.1 (R). These gene frequencies were chosen to yield a population prevalence of about 1%. The disease is tightly linked to the marker (recombination fraction $\theta = 0.01$). The second disease form is modeled by fixing g at the values 0.002, 0.006, and 0.010 (sporadic models 1–6; table 1). Thus, the proportion of sporadic cases among all affected ranged from 0.17 to 0.5. We use the notation D+P and R+P to indicate dominant + phenocopies and recessive + phenocopies, respectively. Table 1 summarizes these generating parameters.

For each of the six 'sporadic' models, 125 data sets consisting of 40 families each were generated, by the following protocol: Population samples of nuclear families (consisting of two parents and at least two offspring) were generated, and families were selected if they had at least one affected member. Sibship sizes were determined according to a negative binomial distribution with mean = 2.8 and standard deviation = 2.3. [14] For these 'sporadic' models, some families exhibited the genetic form, and others, the sporadic form. Only rarely did a 'genetic' family also include a 'sporadic' case.

Table 1. Description and model numbers of 10 'sporadic' generating models and 12 'genetic heterogeneity' generating models used in this study

A. Sporadic models

D+P models with $f = 1, q = 0.005, \theta = 0$		R+P models with $f = 1, q = 0.1, \theta = 0$	
model	g	model	g
1	0.002	4	0.002
2	0.006	5	0.006
3	0.010	6	0.010

B. Genetic heterogeneity models

form 1 (linked): $f_1 = 1, \theta = 0$;
form 2 (unlinked): $f_2 = 1, \theta = 0.5$

D+D models with $q_1 = 0.005$		R+D models with $q_1 = 0.1$	
model	q_2	model	q_2
1	0.005	7	0.005
2	0.0025	8	0.0025
3	0.0005	9	0.0005
D+R models with $q_1 = 0.005$		**R+R models with $q_1 = 0.1$**	
model	q_2	model	q_2
4	0.100	10	0.100
5	0.071	11	0.071
6	0.032	12	0.032

f = Penetrance; q = gene frequency; θ = recombination fraction. In the sporadic models, g = phenocopy frequency. In the heterogeneity models, the subscript 1 refers to the linked locus, and 2, to the unlinked locus. For the 'genetic heterogeneity' models, gene frequencies were chosen to ensure that the unlinked form would occur in frequencies which were equal to, one half of, and 10% of the population frequency of the linked form. D+P models: autosomal dominant; R+P models: autosomal recessive.

Genetic Heterogeneity Models. Under these models, the second form of the disease is, like the first form, genetic, but its locus is not linked to the marker. Both D and R models of inheritance were simulated at each disease locus. We use the notation D+D, D+R, etc., where the first letter describes the mode of inheritance of the linked form of disease, the second describes the unlinked form, and the + sign indicates heterogeneity (as opposed to epistasis), as in [2] (also see the Discussion).

In the simulations, there was full penetrance at both loci. We generated families under three sets of parameters each of the D+D, D+R, R+D and R+R models. For the linked disease locus, the gene frequency was always 0.005 (D) or 0.1 (R), as with the 'sporadic' models. For the unlinked form, the gene frequency was chosen so that the unlinked form would occur in frequencies which were equal to, one half of, and 10% of the population frequency of the linked form. (That is, 0.005, 0.0025, and 0.0005 for D and 0.1, 0.0707 and 0.0316 for R.) See table 1 for a summary.

Data sets were generated by the same protocol as for the 'sporadic models', except that the families were put into 250 data sets of 20 families each, because these data sets were expected to be more informative (since both kinds of families contained more linkage information). The average data set contained approximately 100 individuals.

Analysis Models

All data sets, whether generated under a 'sporadic model' or a 'heterogeneity model', were analyzed assuming a 'sporadic model'. Lod scores were calculated by LIPED [15].

The data were analyzed under the correct mode of inheritance for the linked form (AD or AR), with the correct gene frequencies at that locus. The analysis phenocopy frequency, g, was varied over the values 0.0, 0.002, 0.004, ..., 0.010 in the sporadic and heterogeneity data sets. The analysis penetrance was set to ≈ 1.0. In all data sets, Z_{max} (the maximum lod score) was calculated for each value of the assumed g. We calculated the means and standard deviation of Z_{max} for each assumed g, and plotted mean Z_{max} versus g.

Results

Our findings were twofold. First, we determined that, in the cases where the unlinked form of the disease is sporadic, the value of the assumed phenocopy frequency g, makes little difference to the final lod score, as long as that assumed g is not zero. Second, if the unlinked disease has a genetic basis, we found that assuming a phenocopy frequency usually does *not* compensate for this other form in the linkage analysis. Below, we examine the results for the sporadic and genetic heterogeneity models in detail.

Sporadic Models

Figure 1 shows graphs of mean Z_{max} as a function of assumed g for the six 'sporadic' models. Note that the Z_{max} values remain similar, no matter what value of g was used in the analysis, provided only that the assumed g was positive. These results suggest that the actual phenocopy frequency cannot be easily estimated from a linkage analysis, in contrast with how accurately *penetrance* can be estimated from a linkage analysis [4]. When we assumed g equal zero, in contrast, Z_{max} was lower than for any positive g.

Because Z_{max} appeared to rise so sharply as g rose from 0 to 0.002, we also examined phenocopy frequencies smaller than 0.002, i.e., from 10^{-9} to 0.001 (results not shown). The initial increase in Z_{max} remains steep, with the curve then flattening at phenocopy values greater than 10^{-5}. These results further support the idea that Z_{max} is almost independent of the assumed g, as long as this assumed g is positive.

Genetic Heterogeneity Models

Under the 'genetic heterogeneity models', the behaviors of the Z_{max} versus g curves were less clear-cut. For some models, Z_{max} *decreased* monotonically as the analysis g increased. Even when the curve did peak at some nonzero g value, this g did nor appear to represent any useful or meaningful 'estimate' of a phenocopy frequency. Beyond that, the results differed depending on the heterogeneity model.

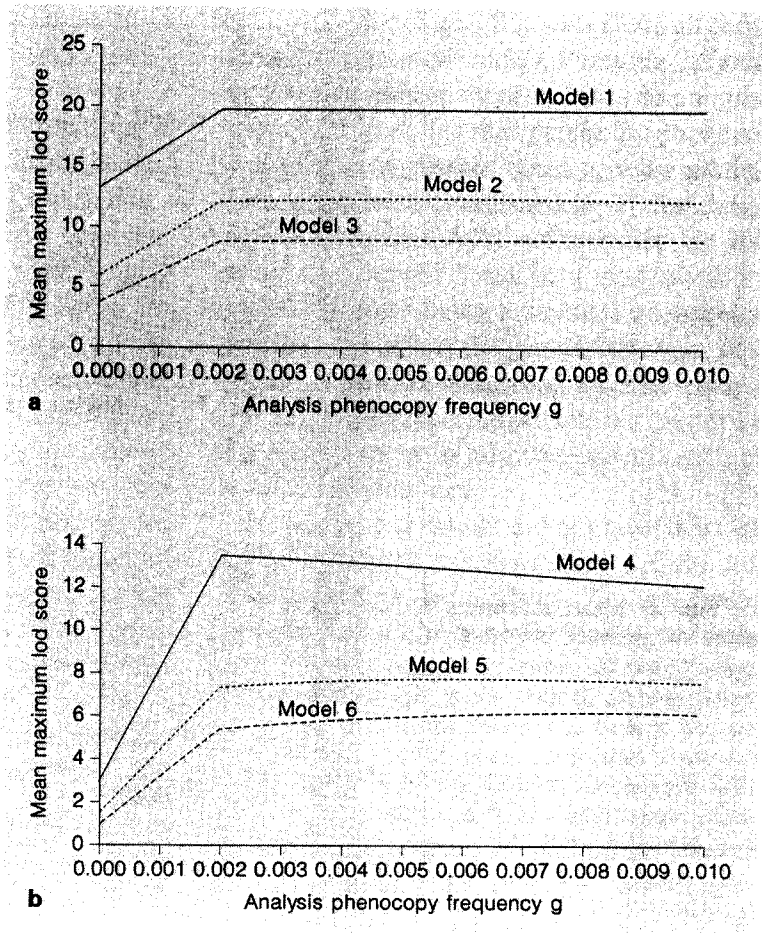

Fig. 1. Mean maximum lod score vs. assumed phenocopy frequency g for the 'sporadic' models. Models are as described in table 1. **a** Models where the linked locus is dominant (D+P, models 1–3). **b** Models where the linked locus is recessive (R+P, models 4–6).

Figure 2a shows graphs for the models where the linked locus was D. Under the D+D model, the actual value of the assumed g – whether zero or nonzero – made little difference in the Z_{max}. Unlike the models with true sporadics (D+P, fig. 1), Z_{max} changed only in the second or first decimal place as the analysis g increased from zero to 0.02. Under the D+R model, in contrast, Z_{max} did increase sharply and discontinuously as the assumed g rose from zero, although less so than when the model was D+P. Tha is, again allowing for *any* phenocopies increased Z_{max} relative to allowing for zero phenocopies. After this initial increase, Z_{max} decreased monotonically with increasing g. In contrast, when the first locus was R, assuming a positive g *lowers* Z_{max} compared to assuming a g of zero (fig. 2b). In all these models, Z_{max} decreased monotonically with increasing g.

Discussion

Our most striking findings were (a) when the true model was 'sporadic', there was a marked difference between the analyses assuming a zero phenocopy frequency and those assuming *any* of a broad range of phenocopy frequencies, and (b) when the true model was

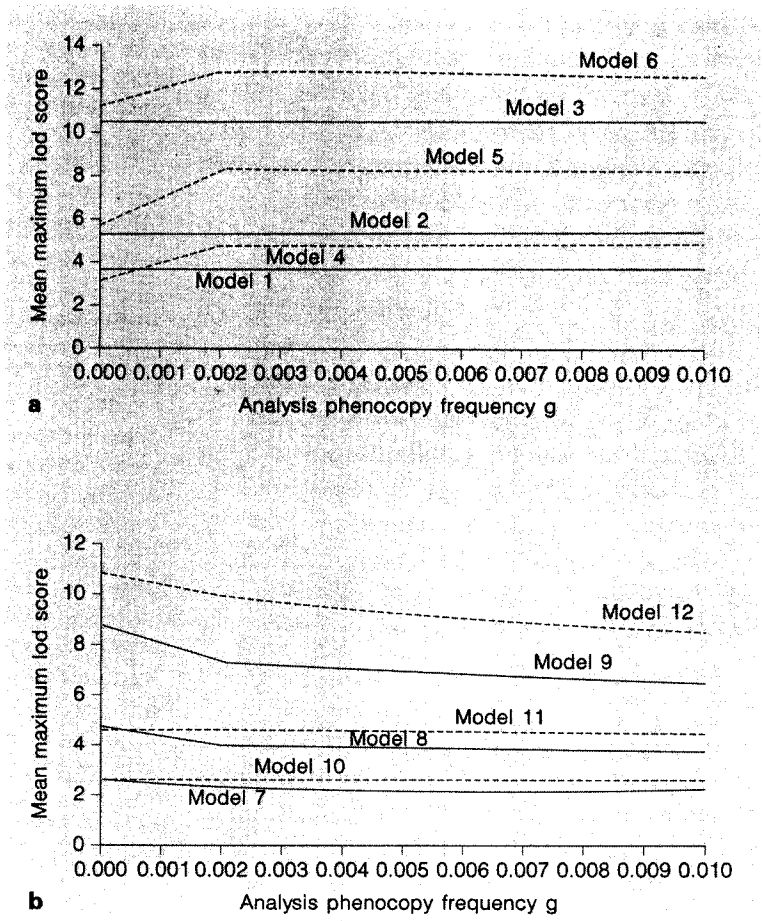

Fig. 2. Mean maximum lod score vs. assumed phenocopy frequency g for the 'genetic heterogeneity' models. Models are as described in table 1. **a** Models where the linked locus is dominant. Solid lines indicate D+D (models 1–3); dashed lines indcate D+R (models 4–6). **b** Models where the linked locus is recessive. Solid lines indicate R+D (models 7–9); dashed lines indicate R+R (models 10–12).

'genetic heterogeneity', a phenocopy frequency does not appear, in general, to compensate for heterogeneity. We now discuss the implications of these findings in more detail.

Sporadic Models

It is well known that under the correct genetic mechanism, one can estimate unknown genetic parameters by maximizing maximum lod scores over those parameters (MMLS or mod scores; [8–11]). Thus, our goal was not to determine *whether* one could estimate the parameter g, but rather *how well* one could estimate g in reasonable-sized data sets. Our simulations yielded a relatively flat Z_{max} versus g curve, from which g could not be estimated with any precision, even with the larger data sets of 40 families each.

The other interesting finding was the 'discontinuity' between using an assumed g of 0 and using any positive value. We surmise that when the possibility of sporadics is not allowed for (i.e., when one assumes a zero g value in the analysis), all families are considered genetic. This produces evidence against linkage. Assuming a phenocopy frequency then allows for the possibility of another form of disease. This allowance is sufficient to make the sporadic families be incorporated as sporadic, as opposed to genetic. Yet, as noted

above, there is still too little information in the data – at least, in data sets of these sizes – to estimate g.

When one looks at the family structures in a dataset generated under, for example, the D+P model, the mixture does not look like a fully penetrant autosomal dominant disease. We wondered whether analyzing the data assuming reduced penetrance could have the same effect as analyzing assuming a nonzero g. We analyzed the 'sporadic' data sets under a zero phenocopy frequency but assuming several penetrance values (results not shown). We found that the lod score increased only slightly and that the increase did not exceed one lod score unit at any value of the assumed penetrance. In neither the dominant nor recessive case did the lod score increase as dramatically as we observed when g was assumed to be positive. Therefore, simply assuming a reduced penetrance will not compensate for sporadics.

Genetic Heterogeneity Models

The main focus of this study was on the 'genetic heterogeneity' models, because we had been struck by the analogy between epistatic two-locus (2L) models and genetic heterogeneity models. For example, one can use 'DD' to refer to an epistatic disease in which an individual must have one copy of the disease gene at *each* of the two disease loci in order to be affected, i.e., locus A *and* locus B. Then 'D+D' can denote a two-locus heterogeneity model in which an individual must have one copy of the disease gene at *either* of the two disease loci to be affected, i.e., locus A *or* locus B [2, 16]. Thus, where the epistatic models are 'multiplicative', their heterogeneity counterparts are 'additive'. Moreover, if disease status is determined by, say, a DD model, then the unaffected state can be viewed as being caused by an R+R model; that is, DD is *complementary* to R+R. Similarly, DR – R+D, RD – D+R, and RR – D+D are all complementary pairs of models. Pursuing the analog, the role played by penetrance in an epistatic model matches that played by phenocopy rate g in a genetic heterogeneity model. Since it has been established [6, 7] that penetrance can successfully compensate for the action of a second locus in a multiplicative model – despite being a random probability that does not reflect the genetic action of that second locus – we wondered whether phenocopy rate could also compensate for the action of a second locus in additive models such as those considered here.

However, our results did not bear this out. As figure 2 shows, when we increased the value of g used in the analysis, Z_{max} increased for only one set of models (D+R), and then only slightly; under all other models, Z_{max} stayed steady or actually decreased with increasing g.

We speculate that this is because, in contrast to the epistatic models, where the linked disease locus is linked in *all* families, here in the case of genetic heterogeneity not all families are actually linked. Therefore, when one assumes $g > 0$, families with the disease caused by the unlinked locus give less negative evidence for linkage, but the lod score of the linked families decreases. The final lod score depends on how much information from the linked families is lost compared to what extent the negative contribution of the unlinked families is 'canceled' out. When the modes of inheritance and penetrance of the linked and unlinked form of the disease are the same, the unlinked families are similar in structure to the linked form, so the loss and gain in information neutralize each other. Thus, despite our assumption of a range of positive g values, the lod score for the D+D model did not change as a function of g. However, for the D+R case, the lod score rose when a positive g was assumed, just as it did

when true sporadics existed. It appears that the more 'sporadic-like' the unlinked form is, compared to the linked form, the more effective it is to assume a nonzero g.

Moreover, the structure of recessive families, especially those ascertained through only one affected member, resembles that of sporadic families more than that of dominant families. It appears that the more the second disease 'resembles' a sporadic, the greater the effect of assuming a nonzero g on the analysis. In the models where the linked form was recessive, the lod score either dropped (R+D model, fig. 2c) when we assumed a positive g or remained approximately the same (R + R model, fig. 2d). The drop in lod score when we assumed a positive g can be attributed to the loss of linkage information from many of the linked families because they are more 'sporadic-like'.

Another concern is whether assuming a nonzero g could lead to spurious indications of linkage, i.e., to inflation of type 1 error. To investigate this possibility, we performed additional simulations, with (a) only data sets with linked families, but at a higher recombination fraction; (b) only 'sporadic' data sets, and (c) only data sets with tightly linked families. We observed *no* inflation of type 1 error rate in these simulations, so we conclude there is probably little if any such inflation in real data sets.

In another set of simulations, we deliberately selected families with both forms of disease present. Obviously, one would not deliberately sample families in this way, but we did this in an effort to understand the behavior of the lod scores more thoroughly. We were also motivated by our previous work suggesting that exclusive sampling of 'high-density' pedigrees might increase *intra*familial heterogeneity [2]. We found that in these families specially selected for having both forms of the genetic disease present, assuming a phenocopy frequency *does* compensate for the second, unlinked form of disease. Not only is the yield in linkage information (i.e., Z_{max}) higher, but also the estimate of the recombination fraction at this maximized Z_{max} is much closer to the generating value.

Therefore, it seems to be possible to use a phenocopy frequency in the analysis to reduce the influence of the second unlinked form of the disease in families with intrafamilial heterogeneity, at least in the specially selected families we used. It appears to work best when the linked form of the disease is dominant. The reason why this approach appears to yield better results with *intra*familial heterogeneity than with *inter*familial heterogeneity may be because the linked form within a family still provides positive linkage information, whereas the *individuals* affected with the second, unlinked form will give reduced negative information when a phenocopy frequency is assumed. In the interfamilial heterogeneity situation, in contrast, whole *families* will be forced to give little linkage information. This finding gives more insight into the behavior of lod score analysis in the presence of sporadics, and can also be useful for some study designs which may, perhaps inadvertently, increase intrafamilial heterogeneity [2].

Finally, one might ask why an investigator would even consider using this simple analysis, rather than an admixture heterogeneity linkage analysis [17, 18]. We feel that this approach presents another tool for the analysis of complex diseases, particularly for an initial genome scan, where an investigator may prefer a simpler, quicker method. Also, studies have been published in which the investigators assumed a positive g without considering a zero g, but we have demonstrated here that this is not necessarily a good course of action.

Recommendations

We cannot generalize overmuch, since simulation results are limited to the models considered. On the other hand, this study has covered a broad spread of simple genetic heterogeneity models, so we feel it is safe to make the following recommendation:

In the analysis of complex common disease, where phenocopies and heterogeneity are a possibility, we recommend testing the effect of assuming a small phenocopy frequency (g = 0.002–0.005) in the analysis, as well as zero g. If Z_{max} increases, the presence of another form of disease is possible, and the information in favor of linkage increases. In this case keep the phenocopy frequency in the analysis. If a positive lod score under the assumption of g = 0 decreases when g is positive, this suggests that the positive g is canceling out positive linkage information because a 'more penetrant' unlinked form is present in the data set. In this case, go back to a phenocopy frequency of zero. We think it important, however, not to simply assume a phenocopy frequency, because it is possible to inaccurately *decrease* the evidence for linkage in some situations. The lod score may increase substantially when there are sporadics in the data set. It will increase moderately when there is intrafamilial heterogeneity and the increase will be greater when the linked form is dominant and the unlinked form recessive.

References

1 Durner M, Greenberg DA: Effect of heterogeneity and assumed mode of inheritance on lod scores. Am J Med Genet 1992;42:271–275.

2 Durner M, Greenberg DA, Hodge SE: Inter- and intrafamilial heterogeneity: Effective sampling strategies and comparison of analysis methods. Am J Hum Genet 1992; 51:859–870.

3 Greenberg DA, Berger B: 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 1994;55:834–840.

4 Greenberg DA, Hodge SE: Linkage analysis under 'random' and 'genetic' reduced penetrance. Genet Epidemiol 1989;6:259–264.

5 Greenberg DA: Linkage analysis assuming a single-locus mode of inheritance for traits determined by two loci: Inferring mode of inheritance and estimating penetrance. Genet Epidemiol 1990;7:467–479.

6 Vieland VJ, Greenberg DA, Hodge SE: Adequacy of single-locus approximations for linkage analyses of oligogenic traits. Genet Epidemiol 1992;9:45–59.

7 Vieland VJ, Greenberg DA, Hodge SE: Adequacy of single-locus approximations for linkage analyses of oligogenic traits: Extension to multigenerational pedigree structures. Hum Hered 1993;43:329–336.

8 Clerget-Darpoux F, Bonaïti-Pellié C, Hochez J: Effects of misspecifying genetic parameters in lod score analysis. Biometrics 1986;42:393–399.

9 Greenberg DA: Inferring mode of inheritance by comparison of lod scores. Am J Med Genet 1989;34: 480–486.

10 Hodge SE, Elston RC: Lods, wrods, and mods: The interpretation of lod scores calculated under different models. Am J Hum Genet 1994;11: 329–342.

11 Elston RC: Man bites dog? The validity of maximizing lod scores to determine mode of inheritance. Am J Med Genet 1989;34:487–488.

12 Ott J: Analysis of human genetic linkage. Baltimore, Johns Hopkins University Press, 1991, p 148.

13 Greenberg DA: Simulation studies of segregation analysis: Application to two-locus models. Am J Hum Genet 1984;36:167–176.

14 Cavalli-Sforza LL, Bodmer WF: The genetics of human populations. San Francisco, Freeman, 1971, pp 310–313.

15 Ott J: Estimation of the recombination fraction in human pedigrees: Efficient computation of the likelihood for human linkage studies. Am J Hum Genet 1974;26:588–597.

16 Schork NJ, Boehnke M, Terwilliger JD, Ott J: Two-trait-locus linkage analysis: A powerful strategy for mapping complex genetic traits. Am J Hum Genet 1993;53:1127–1136.

17 Hodge SE, Anderson CE, Neiswanger K, Sparkes RS, Rimoin DL: The search for heterogeneity in insulin-dependent diabetes mellitus (IDDM): Linkage studies, two-locus models, and genetic heterogeneity. Am J Hum Genet 1983;35:1139–1155.

18 Ott J: Linkage analysis and family classification under heterogeneity. Ann Hum Genet 1983;47:311–320.