Original Paper

Human Heredity

Hum Hered 2002;53:216–226
DOI: 10.1159/000066195

Received: April 3, 2002
Revision received: June 17, 2002
Accepted: June 18, 2002

Evaluating Genetic Heterogeneity in Complex Disorders

Deb K. Pal[a, b] David A. Greenberg[a, b]

[a]Clinical and Genetic Epidemiology Unit, Department of Psychiatry, and [b]Division of Statistical Genetics, Biostatistics Department, Joseph Mailman School of Public Health, Columbia University, New York, N.Y., USA

Key Words
Genetic heterogeneity · Complex diseases · Admixture test · Linkage analysis · Ascertainment

Abstract
Objectives: The Admixture test is routinely used in linkage analysis to take account of genetic heterogeneity, and yields an estimate of the proportion of families (α) segregating the linked disease gene. In complex disorders, the assumptions of the Admixture test are violated. We therefore explore how the estimate of α relates to the true proportion of linked families with a complex disorder in a population or dataset. ***Methods:*** We simulated a two-locus heterogeneity model and varied genetic parameters, ascertainment scheme and phenocopy frequency. ***Results:*** In this model, α is almost always overestimated, by as little as 5% to as much as 60%. The bias is largely attributable to (1) intrafamilial heterogeneity arising from ascertainment of families with many affected members or from analysis of dense pedigrees; (2) low informativeness, which occurs in the presence of reduced penetrance; and (3) differences in the evidence for linkage in linked and unlinked families. This bias is also affected by the analysis phenocopy frequency, but only if the linked locus is dominant and the unlinked locus is recessive. ***Conclusions:*** We conclude that, in complex diseases, the Admixture test has greater value in detecting linkage than in estimating the proportion of linked families in a dataset.

Introduction

Genetic heterogeneity exists when a single disease phenotype can be caused by independent genetic loci. Often there are no phenotypic features to help distinguish the different genetic forms. Genetic heterogeneity is a major complicating factor in linkage analysis. If we perform linkage analysis on a genetically heterogeneous group of affected families with a marker near one disease locus, some families will show evidence of linkage, whereas other families will show evidence against linkage. Heterogeneity can completely mask evidence in favor of linkage even when a proportion of families in the dataset are truly linked to the marker locus of interest.

When there is no predetermined criterion to separate genetic subforms, the presence of genetic heterogeneity can be taken into account in linkage analysis using the Admixture Test [1] in which one estimates an admixture parameter 'α'. In this model a proportion (α) of families are linked to the marker of interest ($\theta < 0.5$) whereas the remaining (1-α) families are unlinked ($\theta = 0.5$). The admixture test has been used to test for linkage and heterogeneity [2], using the likelihood function: $L(\alpha,\theta) = \alpha\, L(\theta) + (1-\alpha)\, L(0.5)$, and has also been incorporated into programs for two-point (HOMOG) [3], and multipoint (GENEHUNTER) [4] linkage analysis. Use of the heterogeneity LOD score (HLOD) allows one to search for linkage in the presence of heterogeneity. The estimate of α from the HLOD is also commonly cited as an indication of the proportion of families in a dataset that are linked to the marker of interest.

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

0001–5652/02/0534–0216$18.50/0

Accessible online at:
www.karger.com/hhe

Dr. Deb K. Pal, PhD MRCP
Clinical and Genetic Epidemiology Unit, Department of Psychiatry
1051 Riverside Drive, Unit 24
New York, NY 10032 (USA)
Tel. +1 212 543 5796, Fax +1 212 342 0484, E-Mail dkp28@columbia.edu

The use of the HLOD to estimate the percentage of linked families is undermined by the fact that estimates of α appear to be accurate only when the data conform exactly to the assumptions of the test. Recall that the parameter α is defined as the fraction of linked families in the *population*, not in the data set. The assumptions of the admixture test are: (1) that there is no intrafamilial heterogeneity – a reasonable assumption in rare diseases but a problematic one in common ones [5]; (2) that the different diseases comprising the clinical phenotype share the same genetic model [6]. In studies of common disease, these assumptions are unrealistic. Also, in situations where there is low power to detect linkage, the estimate of α is biased, especially when only a few families are linked [7].

Whittemore and Halpern [8] recently investigated some theoretical aspects of estimating α. They discussed the assumptions on which the estimation of α is based, the effect of unequal penetrance of the two disease forms, phenocopy frequency, and different gene frequencies, but the definition of their heterogeneity parameter p is not exactly the same as α. Although they proposed a way to correct the estimate of α for some of the confounding parameters they investigated, particularly phenocopies, they concluded that α should never be estimated, even if α is treated only as a nuisance parameter for the sake of taking account of heterogeneity in a linkage analysis.

Whilst their findings are insightful and interesting, these conclusions may be too broad for two reasons. First, it is not reasonable to assert that α should never be estimated, even as a nuisance parameter when assessing evidence for linkage. It is well known that in two-point analysis, in which the lod score is maximised (Z_{max}) over the recombination fraction θ, the presence of unlinked families has the effect of increasing the estimate of θ. In small, phase-unknown families, θ and α are highly correlated. Thus in such families, calculating Z_{max} with respect to θ can, depending on the circumstances, yield similar evidence for linkage as when α is taken into account [9]. Even if using α did not increase the evidence for linkage, simply knowing that heterogeneity exists and what fraction of families are linked can guide how to proceed in genetic studies of disease. Even more critically, for multipoint analysis, the HLOD can mean the difference between detecting evidence of linkage and not detecting it [10].

Second, our previous work on phenocopies did not support the prediction that misspecifying the phenocopy frequency would always lead to a biased estimate of α when using the linkage analysis computer programs at the levels of phenocopies we investigated [11]. We had found that the overall lod score in a two-locus heterogeneity model was unresponsive to the assumed phenocopy frequency when both loci were dominant (D + D) or both recessive (R + R). However, assuming any phenocopy frequency resulted in an increase in lod score for D + R, and decrease in lod score for R + D models. It is possible therefore, that the analysis phenocopy frequency might affect the estimate of α only when the two loci differ in their mode of inheritance. Furthermore, we do not know the effect on linkage detection of trying to reduce or eliminate phenocopies by requiring more than one affected offspring.

Moreover, despite the inherent weaknesses of the Admixture test, there are compelling reasons to investigate its behaviour in complex disorders. First, it is the most widely used for detecting heterogeneity in complex disorders and there are a number of investigations on the HLOD's power to detect linkage [12–17]. On the other hand, relatively little has been written on the relationship between the estimated and true proportion of linked families *in a dataset*. As noted above, the definition of α is the proportion of linked families in the population. In fact, Vieland and Logue [6] showed that when the trait models at linked and unlinked loci differ, the estimate of α derived from the HLOD actually reflects neither the percentage of linked families in the population nor in a dataset. Second, by knowing how the HLOD behaves when its assumptions are violated, we can better understand the results we obtain in everyday analysis situations with complex disorders. We can also use this understanding to devise better tools to assess heterogeneity. Third, once we have identified heterogeneity, we want to use that knowledge to determine the source of the heterogeneity, and find phenotypic characteristics that can separate linked and unlinked disease forms. This will become especially important as large collaborative linkage datasets grow in number, and reliable estimates of α are sought.

The current work centers on the accuracy of the heterogeneity parameter $\hat{\alpha}$, as an estimator of the proportion of linked families in the population, but more importantly in the data set, and the sources of bias. We will explore factors influencing this estimate, and what the estimate signifies for a complex genetic disease. In this study we used computer simulation to study how the estimate of heterogeneity $\hat{\alpha}$ relates to the true proportion of families α in the dataset that are linked to the marker. To explore this relationship between $\hat{\alpha}$ and α we used a two-locus heterogeneity model and varied (a) the genetic parameters (viz mode of inheritance and penetrance of the linked and unlinked loci), (b) the ascertainment scheme, and (c) the generating and analysis phenocopy frequency.

Table 1. Two-locus heterogeneity models, showing penetrance matrices of loci A and B when both have $f = 0.9$

Linked locus	Unlinked locus		
	BB	Bb	bb
D + D model			
AA	0.9	0.9	0.9
Aa	0.9	0.9	0.9
aa	0.9	0.9	0.0
R + R model			
AA	0.0	0.0	0.9
Aa	0.0	0.0	0.9
aa	0.9	0.9	0.9
D + R model			
AA	0.9	0.9	0.9
Aa	0.9	0.9	0.9
aa	0.0	0.0	0.9
R + D model			
AA	0.9	0.9	0.0
Aa	0.9	0.9	0.0
aa	0.9	0.9	0.9

Methods

One of the critical points for this work is the definition of 'linked' and 'unlinked' families. The Admixture test assumes a mixture of families, in a proportion (α) of which the disease locus is linked to the marker and a proportion (1-α) of which it is not. The admixture test also assumes the same genetic parameters at both loci. By varying these parameters, we will be violating the assumptions of the test. In reading what follows, bear in mind that we define α in its narrowest meaning, that is, the proportion of families in which the disease is caused *only* by the disease locus linked to the marker. We do not include in our definition of α families in which both genetic forms are segregating (see below). We define the estimate of heterogeneity from the HLOD as $\hat{\alpha}$, the value obtained by maximizing the HLOD over ø and α.

We generated the data using our extensively tested simulation program [18]. We generated nuclear families and three-generation pedigrees from a population in which two disease genes exist, one linked and one unlinked to a marker. Each of the two disease loci was set to be either dominant (D) or recessive (R). When the disease was dominantly inherited, the gene frequency was set at 0.043, and when recessive, the gene frequency was 0.29, unless otherwise specified. The recombination fraction between the linked locus and the marker was set at $\theta = 0.0$. All matings were fully informative for the marker. For each heterogeneity model, the proportion of each disease form in the population was always 50%, although the final proportion in the dataset could vary depending on the ascertainment criteria. We generated four genetic heterogeneity models (table 1): both loci dominant (D + D); both loci recessive (R + R); linked locus dominant, unlinked locus recessive (D + R); and linked locus recessive, unlinked locus dominant (R + D). We examined the effect of varying ascertainment and penetrance in these models. We simulated 100 datasets each containing 200 families.

In the data-generating step, we labeled families as type 1, type 2, or 'mixed'. Families labeled as 'type 1' were expressing only the form of disease caused by the locus linked to the marker, while families labeled as 'type 2' were expressing only the form of disease not linked to the marker. Families were labeled as 'mixed' if both disease genotypes were expressed in the family. This labeling allowed us to compute the exact proportion of type 1 and type 2 families, as well as allowing us to observe the effect of intrafamilial heterogeneity, i.e. mixed families. Following from the definition given above, α corresponds to the proportion of type 1 families in the *dataset*, i.e. α = n(type 1 families)/n(type 1 + type 2 + mixed families).

Data generation and analysis was divided into four sections each of which was designed to answer a different question: (1) How does $\hat{\alpha}$ behave under D + D, R + R, D + R, R + D models? (2) How does $\hat{\alpha}$ behave when selection criteria for ascertainment are varied? (3) How does $\hat{\alpha}$ behave when the generating or analysis phenocopy frequency is varied? (4) How does $\hat{\alpha}$ behave when the penetrances of the two loci are reduced simultaneously or independently? The various parameters specified for generating data and ascertaining families are summarized in table 2.

(1) We first investigated the case of a heterogeneity model in which all genetic parameters were in accordance with the assumptions of the Admixture test, i.e. full penetrance, both loci of the same dominance model, low gene frequency (0.001 for dominant, 0.01 for recessive) with no intrafamilial heterogeneity. We ascertained nuclear families if at least one offspring was affected. We then extended this investigation to find the estimates of $\hat{\alpha}$ in four different heterogeneity models: (D + D; R + R; D + R; R + D). In these examples, we specified the generating penetrance (f) as 0.9 at each locus, used slightly higher gene frequencies (table 2), and ascertained nuclear families if at least one offspring was affected. Families in which both parents were affected were not selected for analysis.

(2) We examined the effect of different ascertainment schemes on $\hat{\alpha}$ selecting ≥ 1 to ≥ 6 affected offspring. We refer to ascertainment schemes requiring increasing numbers of affected offspring as being more 'stringent'. Families were always ascertained through affected offspring. To study ascertainment, we generated nuclear families for each of the four heterogeneity models described above, and again fixed $f = 0.9$. We also examined three-generation pedigrees for the D + D model.

(3) We tested the effect of varying the a) generating (which we denote as s) and b) analysis (denoted s') phenocopy frequency on the estimate of $\hat{\alpha}$. The phenocopy frequency is defined as the probability of being affected given that the individual does not have a genetic form of disease. (a) We generated a model with a linked dominant locus, an unlinked dominant locus, and sporadic (S) forms (D + D + S), with both genetic loci having $f = 0.9$ and gene frequencies = 0.2. The ascertainment scheme was fixed at ≥ 1 affected offspring. We specified the following generating phenocopy values: 0.0001, 0.001, 0.01, 0.1. (b) We also tested the effect on $\hat{\alpha}$ of changing the analysis phenocopy frequency (between 0.0001, 0.001, 0.01, 0.1) while generating a zero phenocopy frequency. We generated D + D, R + R, D + R, R + D models with all loci having $f = 0.9$, and the ascertainment scheme was fixed at ≥ 1 affected offspring.

(4) Fourthly, we examined the effect of reduced penetrance (f) on the estimate of $\hat{\alpha}$ in D + D, R + R, D + R, R + D models, specifying f for both dominant and recessive loci as 0.9, 0.7, 0.5, or 0.3. We fixed the generating penetrance of each locus to be the same. We also

Table 2. Specified parameters for generating data and ascertaining families

Locus 1 (q1)	Locus 2 (q2)	Family: N = nuclear P = pedigree	Ascertainment criteria: affected offspring	Penetrances $f1, f2$	Generating phenocopy frequency	Analysis phenocopy frequency
Heterogeneity models						
D (0.001)	D (0.001)	N	≥ 1 affd	1.0, 1.0	0.0	0.0
R (0.01)	R (0.01)	N	≥ 1 affd	1.0, 1.0	0.0	0.0
D (0.043)	D (0.043)	N, P	≥ 1 affd	0.9, 0.9	0.0	0.0
D (0.043)	R (0.29)	N	≥ 1 affd	0.9, 0.9	0.0	0.0
R (0.29)	R (0.29)	N	≥ 1 affd	0.9, 0.9	0.0	0.0
R (0.29)	D (0.043)	N	≥ 1 affd	0.9, 0.9	0.0	0.0
Ascertainment						
D (0.043)	D (0.043)	N	≥ 1 to ≥ 6 affd	0.9, 0.9	0.0	0.0
D (0.043)	R (0.29)	N	≥ 1 to ≥ 6 affd	0.9, 0.9	0.0	0.0
R (0.29)	R (0.29)	N	≥ 1 to ≥ 6 affd	0.9, 0.9	0.0	0.0
R (0.29)	D (0.043)	N	≥ 1 to ≥ 6 affd	0.9, 0.9	0.0	0.0
Phenocopy frequency						
D (0.043)	D (0.043)	N	≥ 1 affd	0.9, 0.9	0.0	0.0001–0.1
D (0.043)	R (0.29)	N	≥ 1 affd	0.9, 0.9	0.0	0.0001–0.1
R (0.29)	R (0.29)	N	≥ 1 affd	0.9, 0.9	0.0	0.0001–0.1
R (0.29)	D (0.043)	N	≥ 1 affd	0.9, 0.9	0.0	0.0001–0.1
D (0.2)	D (0.2)	N	≥ 1 affd	0.9, 0.9	0.0001–0.1	0.0
Penetrance						
D (0.043)	D (0.043)	N	≥ 1 affd	0.3–0.9, 0.3–0.9	0.0	0.0
D (0.043)	R (0.29)	N	≥ 1 affd	0.3–0.9, 0.3–0.9	0.0	0.0
R (0.29)	R (0.29)	N	≥ 1 affd	0.3–0.9, 0.3–0.9	0.0	0.0
R (0.29)	D (0.043)	N	≥ 1 affd	0.3–0.9, 0.3–0.9	0.0	0.0

Parameters for generating data and ascertaining families. Dominance models D, dominant and R recessive at locus 1 and 2. Respective gene frequencies are given in parentheses, q1 and q2. Families were generated either with nuclear (N) or pedigree (P) structure. A family is selected for analysis depending on the minimum number of affected offspring (≥ 1 etc.).

Families are not selected if both parents are affected. The penetrance (f) at locus 1 and 2 is either fixed at 0.9, or varied between the values 0.3, 0.5, 0.7, 0.9, locus 1 always having the same penetrance as locus 2. The phenocopy frequency in the analysis is also varied or fixed as shown, but the phenocopy frequency in data generation was fixed at zero except for one example. Analysis was performed using two-point lod score methods, specifying the same parameters for dominance, gene frequency, and penetrance as those used to generate the linked locus, except of course when the analysis phenocopy frequency was varied.

examined the effect of varying the penetrance of each locus ($f1$, $f2$) independently, in the D + D model only (e.g. $f1 = 0.7$, $f2 = 0.5$).

We used the 'true' parameters of the linked locus in the analysis, except of course when exploring the effect of varying analysis phenocopy frequency. Linkage calculations were performed with LIPED [19]. We used the program HETEROTEST to estimate α [2]. We calculated mean maximum lod scores and mean $\hat{\alpha}$ for the 100 datasets as a whole, and also separately for type 1, type 2, and mixed families in the datasets. In the figures, we plotted mean $\hat{\alpha}$ vs. the parameters of interest (penetrance, ascertainment criterion), the mean being calculated over the 100 datasets. We also plotted the mean of the actual proportion of type 1, type 2, and mixed families in the datasets for comparison with $\hat{\alpha}$. We included values of the lod score in some graphs.

Results

Our original purpose in this study was to assess what factors affect the estimate of α in a data set. Recall that in the Admixture test α is the proportion of linked families in the population. Whilst we do not know the theoretical relationship to the proportion of linked families in the data set, we expected that such a relationship could be empirically determined. In general, we found that $\hat{\alpha}$ was a poor estimator of α. The size and direction of the bias varied according to the generating genetic parameters and specified ascertainment criteria. We found that increasing

the number of affected family members required for ascertainment biased the estimate of α in an upward direction. Estimates of α were more inaccurate when intrafamilial heterogeneity was present and this bias worsened as the proportion of mixed families rose. Estimates of α in datasets of three-generation pedigrees were substantially worse than in datasets consisting of nuclear families. Varying the analysis phenocopy frequency had negligible effect on the estimate of α in D + R but had a notable effect in the other three models. The effect of reduced penetrance also led to substantial upward bias in the estimate of α, which was reduced by constraining the HLOD to θ = 0.01.

Influence of Mode of Inheritance on Estimate of α

First, in the full penetrance ($f1 = f2 = 1.0$), same dominance models, we found that $\hat{\alpha}$ was 0.52 (95% CI:0.50–0.54) in the D + D model, and 0.56 (95% CI:0.54–0.59) in the R + R model, when the true mean proportion of type 1 families was 0.49. This was true despite the absence of mixed families in the datasets. We found that the estimate of α was biased in both of these cases because families which have the trait caused by the unlinked trait locus still had a chance of showing positive evidence for linkage by chance alone, although at high θ (e.g. 0.2–0.3). The linked families were generated with a recombination fraction between trait and marker of zero and were thus incapable of showing negative lod scores by chance. This asymmetric situation caused the maximum heterogeneity lod score to be upwardly biased in the estimate of α.

With the generating penetrance of each locus fixed at 0.9, and demanding at least one affected offspring, estimates of alpha were increased 5–7% over the actual proportion of linked families in the dataset in the D + D, R + R, and R + D models (fig. 1a–c). In the D + R model (fig. 1d), the overestimate was much more pronounced (17%).

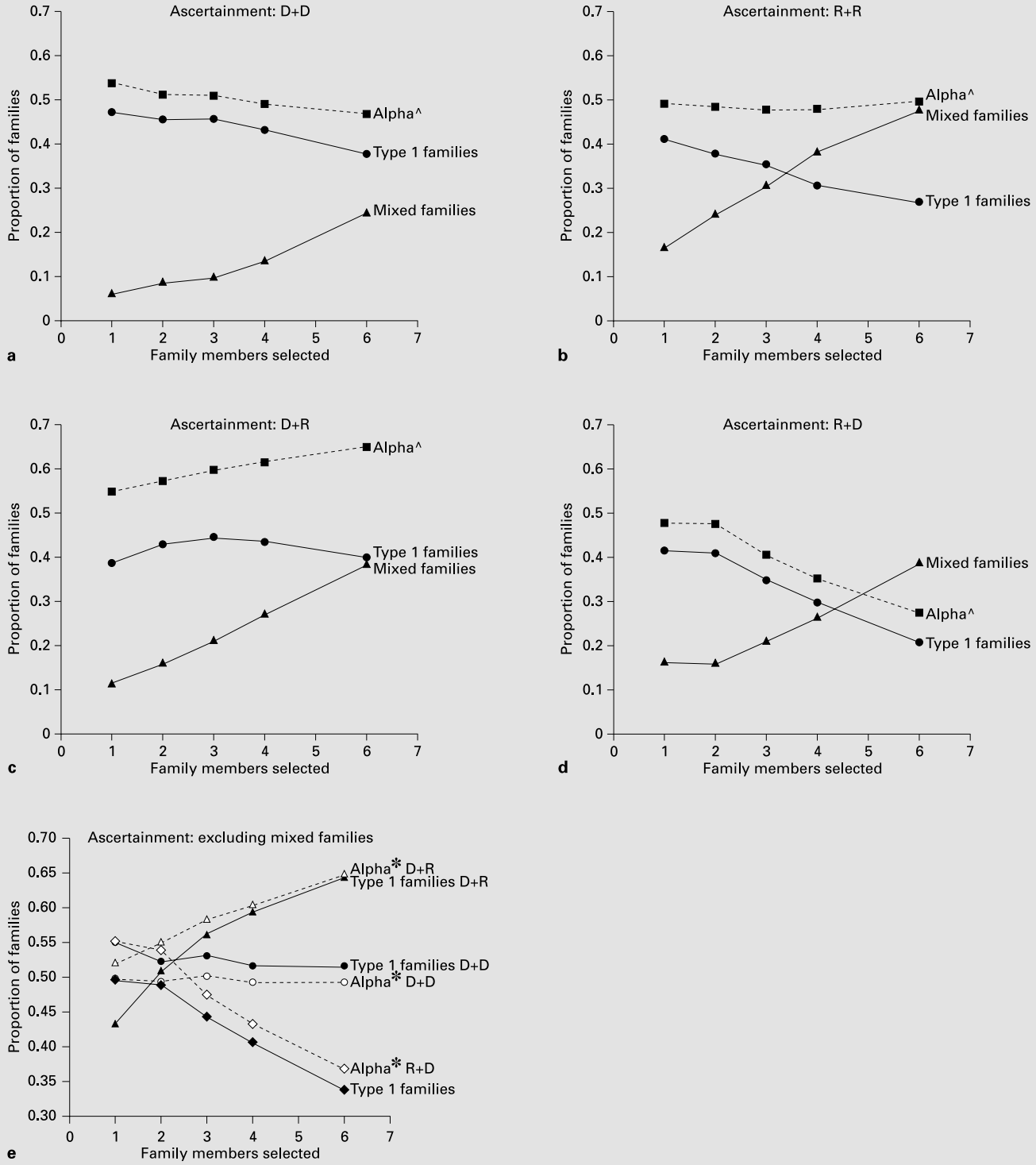
Influence of Ascertainment on Estimate of α

We next examined the effect of changing the ascertainment scheme by requiring more affected family members. As we demanded more affected family members, the error in estimating α increased in all except the R + D model (fig. 1a–d). For example, when six or more affected offspring were selected, the bias was 10%, 22% and 24% higher than the true α in the D + D, R + R, and D + R models, respectively. The divergence of $\hat{\alpha}$ from α with more stringent ascertainment was similar in D + D and R + R models (fig. 1a, b).

The bias in the R + D model was a constant 7% regardless of whether ≥ 1 or ≥ 6 offspring were selected. With increasing numbers of affected offspring, there was a concurrent rise in mixed families, in which members with both forms of disease occurred. Note from figure 1a–d that $\hat{\alpha}$ is not equivalent to the proportion of all families made up of type 1 plus mixed families. Neither does it reflect the count of evidence for linkage in the dataset: if we count type 1 families and mixed families that show evidence for linkage (lod score >0), the number of these families, as a proportion of the total families, does not correspond to $\hat{\alpha}$ (data not shown).

When we eliminated mixed families, the bias in estimation of α was significantly attenuated. Figure 1e shows estimates of α, with mixed families excluded (see the dashed lines labelled α*), the bias has fallen to 1% (D + R model) or 4% (D + D, R + R, R + D models) when ≥ 6 affected offspring are selected. (R + R is not graphed for simplicity). This result indicated that intrafamilial heterogeneity was a major source of the overestimate of α with requirement for increasing number of affected offspring. To further test this, we generated disease at low (q_L = 0.001) and high (q_H = 0.2) gene frequencies in the D + D model. We anticipated more intrafamilial heterogeneity in the high gene frequency model, and consequently a greater bias in the estimate of α. We indeed found this to be the case: selecting one or more affected offspring in the q_H example resulted in a 20% overestimate of α with 30%

Fig. 1. a True and estimated proportions of linked families, and true proportions of mixed families obtained under different ascertainment criteria. Each point represents a mean calculated from 100 datasets of 200 nuclear families each generated under a two-locus D + D heterogeneity model, each locus having 90% penetrance. The mean maximum lodscores, for lowest (≥ 1) and highest (≥ 5) ascertainment schemes, are adjacent to the curves for type 1 and mixed families. The ascertainment criteria are varied to require greater or equal to one, two, three, four or six affected offspring. Data were analyzed under two-point lodscore methods assuming the generating parameters of the linked locus. The top curve (boxes) represents the mean estimate of heterogeneity $\hat{\alpha}$ obtained from the admixture test (Alpha^). The middle curve (circles) represents the mean proportion of type 1 families in each dataset. The bottom curve (triangles) represents the mean proportion of mixed families in each dataset. b As for 1a in the R + R model. c As for 1a in the D + R model. d As for 1a in the R + D model. e Solid lines reproduce the mean proportions of type 1 families in D + R (triangles), D + D (circles), and R + D (diamonds) models according to ascertainment criteria shown in figures 1a, 1c and 1d. The dashed lines alongside each solid curve represent the estimate of α from the admixture test (Alpha*) after the exclusion of mixed families from the datasets. The R + R model, which closely resembles the D + D model, is not shown for clarity of presentation.


Ascertainment: D+D
Proportion of families
Family members selected
Alpha^
Type 1 families
Mixed families
a
Ascertainment: R+R
Alpha^
Mixed families
Type 1 families
b
Ascertainment: D+R
Alpha^
Type 1 families
Mixed families
c
Ascertainment: R+D
Mixed families
Alpha^
Type 1 families
d
Ascertainment: excluding mixed families
Alpha* D+R
Type 1 families D+R
Type 1 families D+D
Alpha* D+D
Alpha* R+D
Type 1 families
e

of the families being mixed; on the other hand in the low gene frequency (q_L) disease, there was a 4% overestimate in α and a negligible number of mixed families. We also found greater bias in $\hat{\alpha}$ in three-generation pedigrees under the D + D, $q1 = q2 = 0.043$ model: $\hat{\alpha}$ was 11% higher than the true α when one or more affected member was selected (cf. nuclear families: $\hat{\alpha}$ was 5% higher than true α). Excluding the mixed families significantly reduced the bias in both the high gene frequency case (from 20–36% to 2–3%), and in pedigrees (from 11–27% to 1–2%), confirming the hypothesis that intrafamilial heterogeneity was a major source of systematic error.

The effect of increasingly stringent ascertainment criteria (ie selecting for more affected family members) on $\hat{\alpha}$ is influenced by the mode of inheritance at the second, unlinked locus. If this unlinked locus is recessive and the linked is dominant, there is significant bias, even when ascertainment criteria are not stringent (eg 17% when ≥ 1 affected member selected). When more affected offspring are required for ascertainment, there is a consistently strong bias (eg 22–25% when ≥ 6 affected offspring selected) whenever the unlinked locus is recessive, regardless of the inheritance of the linked locus. Selecting for more affected members will always result in a preference for dominant disease forms if dominant and recessive forms have equal population prevalence. In our examples, dense families or pedigrees (regardless of the heterogeneity model) are saturated with the dominant disease form.

Sensitivity of Alpha to Phenocopy Frequency

(1) Generating phenocopies. We found, in a D + D + S model, that the effect on $\hat{\alpha}$ of varying the generating s between 0.0001 to 0.1 (fixing $s' = 0.0$) was very small: mean $\hat{\alpha}$ changed 17 from 0.55 (sd 0.079) when s was 0.0001, to 0.58 (sd 0.085) when s was 0.1.

(2) Analysis phenocopies. To test the proposition that the estimate of α is sensitive to the misspecification of the phenocopy frequency s [8], we varied the analysis phenocopy frequency s' from 0.00001 to 0.01 in the four heterogeneity models. In the D + D, R + R, R + D models, estimates of α were unaffected by change in s': when assumed sporadic frequencies between zero and 0.01 were specified, estimates of α changed by a mean of 1%. The mean maximum lod scores were almost unchanged by varying s' in these models. In the D + R model, $\hat{\alpha}$ was biased by up to 11%, rising from 0.55 (at $s' = 0.0$) to 0.66 (at $s' = 0.01$), with a parallel increase in mean maximum lod score from 21.4 at $\theta = 0.04$ ($s' = 0.0$) to $Z_{max} = 24.0$ at $\theta = 0.04$ ($s' = 0.01$).

The Effect of Reduced Penetrance on the Estimate of Alpha

(1) Both loci share same penetrance. We found that, in all models, the lower the penetrance f, the larger was the overestimate of α (fig. 2a–d for D + D model). For example, at $f = 0.3$, the bias in estimating α was 33%, 22%, 35%, and 19% in D + D, R + R, D + R, and R + D models respectively. Intrafamilial heterogeneity was not the explanation for this finding. Low generating penetrance did not lead to a marked increase in mixed families. Nor did excluding mixed families improve the estimate of α, as we had found in the ascertainment example above. As penetrance fell, the information available for estimating heterogeneity also fell. In a heterogeneous dataset with reduced penetrance, the heterogeneity lodscore (HLOD) typically maximized at an estimated recombination fraction larger than the true disease-marker distance of 0.0 (e.g. Z_{max} at $\hat{\theta} = 0.2$), with an accompanying upward bias in α. We already know that the two parameters θ and α are highly correlated [9], and so reasoned that by constraining analysis to a fixed, small value of θ, this might result in attenuation of the upward bias in $\hat{\alpha}$. We therefore constrained the HLOD to assess α at lodscore values corresponding to only $\theta = 0.01$. Constraining the θ value to $\theta = 0.01$ resulted in consistently more accurate estimates of α, as long as the evidence for linkage exceeded a lod score of 3.0. We successfully used this tactic to improve the estimate of α in conditions of reduced penetrance. The estimate of heterogeneity using an analysis θ constrained to be 0.01 is now biased by 9%, 9%, 15%, and 5% in the D + D, R + R, D + R and R + D models (whereas before it was 33%, 22%, 35%, 19% respectively). Figure 2a–d shows the D + D, R + R, D + R and R + D models respectively (dotted lines labeled Alpha^^ represent test estimates constrained to theta = 0.01). Using the HLOD constrained to $\theta = 0.01$ did not reduce the bias in $\hat{\alpha}$ resulting from stringent ascertainment schemes.

(2) Varying penetrance of each locus independently. In the D + D model, we found that the bias in $\hat{\alpha}$ was greater when the linked locus had a higher penetrance than the unlinked locus (table 2). This bias was always positive and increased both as the penetrance of the linked locus fell (as we found in 4a), and as the disparity between penetrances rose.

For example, with $f1 = 0.9$ and $f2 = 0.3$, the bias was 14%; when $f1 = 0.9$ and $f2 = 0.7$, the bias was 11%; when $f1 = 0.5$ and $f2 = 0.3$, the bias was 30%. Constraining the HLOD to $\theta = 0.01$ resulted in less positive bias when the unlinked locus had lower penetrance than the linked locus, for example with $f1 = 0.5$ and $f2 = 0.3$, the bias fell

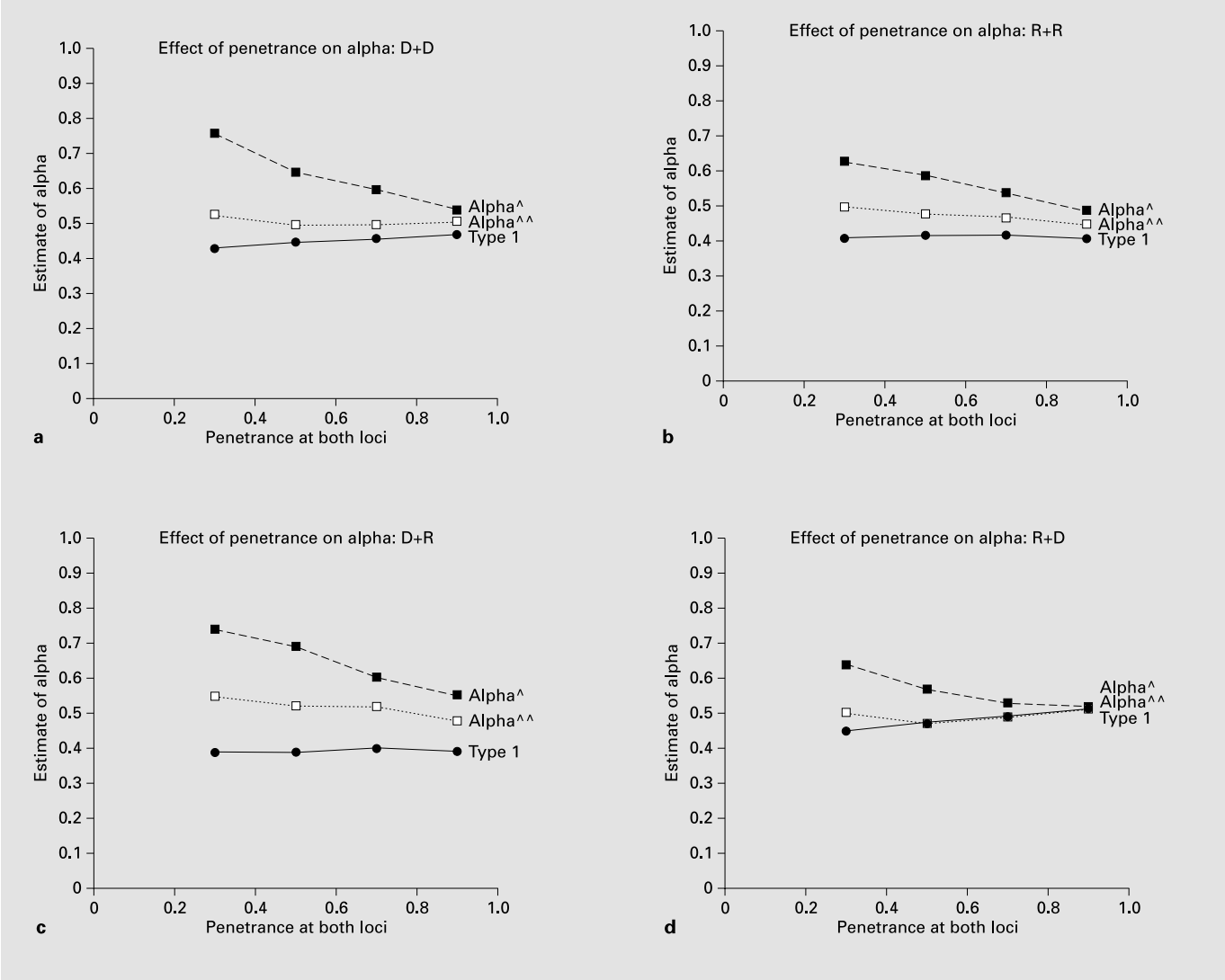


Fig. 2. a True and estimated proportions of linked families obtained when both loci have reduced penetrance varying from 0.3 to 0.9. Each point represents a mean calculated from 100 datasets of 200 nuclear families each generated under a two-locus D + D heterogeneity model. The ascertainment criteria are fixed to require greater or equal to one affected offspring. Data were analyzed under two-point lodscore methods assuming the generating parameters of the linked locus. The bottom solid curve (circles) represents the mean proportion of type 1 families. The top dashed curve (squares) represents the estimate of α from the admixture test (Alpha^). The middle dotted curve (open squares) represents the estimate of α produced by the admixture test (Alpha^^) when constrained to θ = 0.01. b as a for R + R model; c as a for D + R model; d as a for R + D model.

from 30% to 15%. When the unlinked locus had higher penetrance than the linked locus, there was a negative bias in $\hat{\alpha}$. For example, when $f1 = 0.5$ and $f2 = 0.9$, then the bias was –7% in the constrained test, compared to +9% when the unconstrained test was used (table 3).

Discussion

In complex diseases, $\hat{\alpha}$ can be viewed as a nuisance parameter when the HLOD is used to maximise evidence for linkage in the presence of heterogeneity. The Admixture test retains an important role in this regard, and we know that in the presence of heterogeneity, the heterogeneity lod score (HLOD) has superior power to detect link-

Table 3. Bias in estimate of α in the D + D heterogeneity model when each locus has different penetrance

Linked locus (D) penetrance		Unlinked locus (D) penetrance			
		0.3	0.5	0.7	0.9
0.3	$\hat{\alpha}$	0.33	0.14	0.18*	0.23*
	$\bar{\alpha}$	0.10	−0.09	−0.13*	−0.19*
0.5	$\hat{\alpha}$	0.30	0.19	0.16	0.09
	$\bar{\alpha}$	0.15	0.05	−0.05	−0.07
0.7	$\hat{\alpha}$	0.24	0.23	0.14	0.08
	$\bar{\alpha}$	0.14	0.13	0.04	−0.002
0.9	$\hat{\alpha}$	0.12	0.12	0.11	0.07
	$\bar{\alpha}$	0.12	0.08	0.06	0.04

Both loci have the same gene frequency (0.043). The admixture test is applied to estimate α without constraining θ ($\hat{\alpha}$), and also constrained to θ = 0.01 ($\bar{\alpha}$). Analyses were performed assuming 'correct' parameters of the linked locus. * Maximum lod score ≥ 3.0.

age over homogeneity lods and so called model-free methods [20, 21]. Even when proportions of linked families vary across different datasets of ASPs, a simple adaptation of the HLOD yields higher power than both the homogeneity LOD and certain nonparametric tests [22, 23]. The usefulness of the HLOD as an estimator of heterogeneity in complex disorders is more open to question. In the range of models and parameters we have studied, $\hat{\alpha}$ overall performs poorly in estimating the proportion of type 1 families in a dataset. In analysis of real datasets, the true trait parameters and ascertainment schemes are usually unknown, so it is difficult to predict exactly how accurate $\hat{\alpha}$ will be. In complex diseases, $\hat{\alpha}$ corresponds neither to the proportion of type 1 families (α), nor to the proportion of (type 1 and mixed) families in the dataset that give evidence for linkage. The estimate of α is biased when the assumptions of the Admixture test (i.e. disease forms share the same genetic model and there is no intrafamilial heterogeneity) are violated. Vieland and Logue [6] reported conclusions consistent with ours, that when linked and unlinked loci differ in their trait models, the HLOD is based on the wrong likelihood and therefore returns incorrect estimates of the trait parameters.

In this paper, we wanted to understand the detailed behavior of $\hat{\alpha}$ using the HLOD. We were interested to know how ascertainment, penetrance and the assumed phenocopy frequency would affect $\hat{\alpha}$. We found that the bias arises from three sources: intrafamilial heterogeneity, low information content, and difference in linkage evidence between linked and unlinked groups, which are all in turn influenced by genetic parameters and ascertainment strategies. For example, the probability of intrafamilial heterogeneity rises as the gene frequency of the alleles at the two disease loci increases. Demanding many affected members has the result of enriching for more than one disease form in the family. Selecting pedigrees for study, as compared to nuclear families, also increases the probability of having two or more disease forms in the family. The influence of ascertainment on $\hat{\alpha}$ applies whether or not the two disease forms have the same or different modes of inheritance, but is more pronounced when the unlinked locus is recessive.

Bias in estimating α also arises from low linkage information content of the data. Low information content arises either when: (1) there is reduced penetrance at the disease loci, or (2) when marker and disease loci are not tightly linked (i.e. $\theta > 0.01$). When the two disease loci have the same dominance model and reduced penetrance, we observed a positive bias in $\hat{\alpha}$; the size of which depends on the penetrance of the linked locus and the difference in penetrance between linked and unlinked forms. The bias arising from low information content, which we found applies to both the estimate of θ and α in the HLOD, could be partially corrected experimentally by constraining the HLOD to $\theta = 0.01$.

Using the constrained test does not improve the biased estimate of α resulting from stringent ascertainment schemes, but removing mixed families does improve it. Conversely, eliminating mixed families, which improves the estimate of α in the presence of intrafamilial heterogeneity does not improve the biased estimate of α when the disease loci have reduced penetrance. This implies that ascertainment and reduced penetrance have separate and independent influences on the estimate of α. When phenocopies are truly present in the dataset, assuming a zero phenocopy frequency in the analysis has very little effect on either the maximum lod score, or the estimate of α. Whittemore and Halpern stated that estimates of heterogeneity are sensitive to misspecification of the analysis phenocopy frequency [8]. (We note that they appeared to assume extraordinarily high proportions of phenocopies in the population). We found this to be true under the D + R model, but not the other models we studied, using a realistic range of values for non-genetic disease forms (i.e. phenocopy frequency). In the D + R model, families with the disease caused by the unlinked locus give evidence for linkage that is less negative when one assumes a non-zero phenocopy frequency, but the lod score of the linked fami-

lies also decreases [24]. When the modes of inheritance and penetrance of the linked and unlinked forms are the same, the loss and gain cancel each other out, and so the overall lod score and estimate of α are unchanged. However, in the D + R case, assuming a non-zero phenocopy frequency results in a better representation of the model, which increases the overall lod score and estimate of α. We had previously proposed that the analysis phenocopy rate differentially influences evidence for linkage in recessive disease, because families with recessive disease are more 'sporadic-like' [24]. This assumption is supported in this study, since a higher analysis sporadic frequency seems to remove evidence against linkage in the unlinked recessive disease forms which, in the D + R case, increases the lod score and biases the estimate of α. Can the estimate of α be improved sufficiently to be useful in the genetic analysis of complex diseases? Bias can be reduced by several means, but the estimate of α still appears to be unreliable. Firstly, recall the fact that α and θ are highly correlated. Thus even when there is no heterogeneity, if the true θ is greater than about 0.1, there will appear to be evidence for heterogeneity. With modern genome scans, θ can be made quite small and certainly less than 0.1. We have also shown that for certain circumstances, i.e. when reduced disease penetrance exists, the estimate of α can be further improved by constraining the HLOD to θ = 0.01. A second way to improve the estimate of α is to avoid the use of pedigrees or dense families with three or more affected offspring. Several drawbacks about the use of pedigrees in linkage analysis of complex diseases have already been highlighted [11, 25]. Larger pedigrees may of course contribute significantly towards evidence for or against linkage, whereas nuclear families may contribute less information or evidence about linkage. Therefore avoiding dense pedigrees will usually necessitate a larger sample size of families to demonstrate evidence for linkage. We have not explored the effect of misspecifying analysis values other than the phenocopy frequency. We are, however, aware that parameter misspecification has limited effect on the maximum lod score, except when dominance is incorrectly specified [26, 27]. Since the estimation of α in a two-locus heterogeneity model is dependent on the relative magnitude of lod scores contributed by the linked and unlinked disease forms, parameter misspecification might be expected to influence $\hat{\alpha}$ in a manner predicted by its effect on the lod scores of the linked and unlinked disease forms. Nevertheless, despite all these measures, the estimate of α, in the range of situations we have studied, can still be expected to diverge from the true α by 15–20%. The estimate only rarely approaches the exact proportion of type 1 families in the dataset. The bias in using the HLOD to estimate α has further implications for research in complex genetic diseases, both for data collection and for reaching conclusions about molecular associations after linkage has been established. Since linkage plus heterogeneity is difficult to detect by statistical means, strenuous attempts must be made to identify and control heterogeneity at the phenotyping or data collection stage. This will require detailed analysis of perhaps subtle clinical features, which may yield useful clues for subclassification in linkage analysis. This strategy will also yield dividends in terms of evidence for linkage, at the same time making clearer genotype-phenotype correlations.

After demonstrating linkage, the investigator might perform an association study with alleles from candidate genes in linked families (i.e. families from the dataset that give evidence for linkage). If assumptions about the homogeneity of the dataset are biased, ($\hat{\alpha}$ indicating more homogeneity than is actually present) then association results will be confusing. Some affected family members might have the mutation of interest, whereas others will not (e.g. because of intrafamilial heterogeneity from genetic or non-genetic causes, or reduced penetrance), so making it difficult to conclusively demonstrate clear genotype-phenotype relationships with putative candidate genes. This conundrum is particularly common in large dense pedigrees, in which it is often wrongly assumed that only a single disease form segregates.

We conclude that the behaviour of $\hat{\alpha}$ as an estimator of the proportion of linked families in a complex disease dataset is open to considerable bias. Unfortunately, when studying complex diseases, we have little, if any, prior knowledge of the true genetic parameters and so cannot predict the degree of bias in the estimate of α. Although we have suggested several means by which the bias may be reduced, $\hat{\alpha}$ remains an unreliable estimator of linked families in a dataset. However, we strongly disagree with Whittemore and Halpern's [8] suggestion that the estimation of linked families 'be postponed until the relevant genes have been identified'. For in order to find the genes, we must make efforts to identify and reduce heterogeneity in the first place. More important than improving the estimate of α is the fact that the actual value of α is not usually crucial in moving closer to the identification of the disease locus. What is important is i) to take heterogeneity into account, with α as a nuisance parameter, to increase evidence in favour of linkage; and ii) to use the families in the dataset with positive evidence for linkage to understand the origin of heterogeneity based on clinical, epide-

miologic, family, ethnic or other data. Once the origin of heterogeneity is understood, not only can gene identification proceed with more understanding of the disease, but also new hypotheses can be tested about how to differentiate disease forms or symptoms.

Acknowledgements

D.K.P. was supported by a Royal Society-Fulbright Distinguished Postdoctoral Scholarship and by the Dunhill Medical Trust; D.A.G. was supported by NIH grants NS31775, NS27941 and MH48858. We thank Professor Susan Hodge for her comments on the manuscript.

References

1 Smith CAB: Testing for heterogeneity of recombination fraction values in human genetics. Ann Hum Genet 1963;27:175–182.
2 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.
3 Ott J: Linkage analysis and family classification under heterogeneity. Ann Hum Genet 1983;47:311–320.
4 Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES: Parametric and nonparametric linkage analysis: A unified multipoint approach. Am J Hum Genet 1996;58:1347–1363.
5 Martinez M, Goldin LR: Power of the linkage test for a heterogeneous disorder due to two independent inherited causes: A simulation study. Genet Epidemiol 1990;7:219–230.
6 Vieland VJ, Logue M: HLODs, trait models, and ascertainment: Implications of admixture for parameter estimation and linkage detection. Hum Hered 2002;53:23–35.
7 Chiano MN, Yates JR: Linkage detection under heterogeneity and the mixture problem. Ann Hum Genet 1995;59:83–95.
8 Whittemore AS, Halpern J: Problems in the definition, interpretation, and evaluation of genetic heterogeneity. Am J Hum Genet 2001;68: 457–465.
9 Abreu PC, Hodge SE, Greenberg DA: Quantification of type I error probabilities for heterogeneity LOD scores. Genet Epidemiol 2002;22: 156–169.
10 Greenberg DA, Abreu PC: Determining trait locus position from multipoint analysis: Accuracy and power of three different statistics. Genet Epidemiol 2001;21:299–314.
11 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.
12 Cavalli-Sforza LL, King MC: Detecting linkage for genetically heterogeneous diseases and detecting heterogeneity with linkage data. Am J Hum Genet 1986;38:599–616.
13 Chakravarti A, Badner JA, Li CC: Tests of linkage and heterogeneity in Mendelian diseases using identity by descent scores. Genet Epidemiol 1987;4:255–266.
14 Clerget-Darpoux F, Babron MC, Bonaiti-Pellie C: Power and robustness of the linkage homogeneity test in genetic analysis of common disorders. J Psychiatr Res 1987;21:625–630.
15 Goldin LR, Gershon ES: Power of the affected-sib-pair method for heterogeneous disorders. Genet Epidemiol 1988;5:35–42.
16 Lander ES, Botstein D: Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. Proc Natl Acad Sci USA 1986; 83:7353–7357.
17 Ott J: The number of families required to detect or exclude linkage heterogeneity. Am J Hum Genet 1986;39:159–165.
18 Greenberg DA: Inferring mode of inheritance by comparison of LOD scores. Am J Hum Genet 1989;34:480–486.
19 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.
20 Goldin LR, Weeks DE: Two-locus models of disease: comparison of likelihood and nonparametric linkage methods. Am J Hum Genet 1993;53:908–915.
21 Abreu PC, Greenberg DA, Hodge SE: Direct power comparisons between simple LOD scores and NPL scores for linkage analysis in complex diseases. Am J Hum Genet 1999;65: 847–857.
22 Huang J, Vieland VJ: Comparison of 'model-free' and 'model-based' linkage statistics in the presence of locus heterogeneity: Single data set and multiple data set applications. Hum Hered 2001;51:217–225.
23 Vieland VJ, Wang K, Huang J: Power to detect linkage based on multiple sets of data in the presence of locus heterogeneity: Comparative evaluation of model-based linkage methods for affected sib pair data. Hum Hered 2001;51: 199–208.
24 Durner M, Greenberg DA, Hodge SE: Phenocopies versus genetic heterogeneity: Can we use phenocopy frequencies in linkage analysis to compensate for heterogeneity? Hum Hered 1996;46:265–273.
25 Greenberg DA: There is more than one way to collect data for linkage analysis. What a study of epilepsy can tell us about linkage strategy for psychiatric disease. Arch Gen Psychiatry 1992; 49:745–750.
26 Greenberg DA, Abreu P, Hodge SE: The power to detect linkage in complex disease by means of simple LOD-score analyses. Am J Hum Genet 1998;63:870–879.
27 Pal DK, Durner M, Greenberg DA: Effect of misspecification of gene frequency on the two-point lod score. Eur J Hum Genet 2001;9:855–859.