Original Paper

Human Heredity

Hum Hered 2004;57:207–219
DOI: 10.1159/000081448

Received: October 30, 2003
Accepted after revision: May 27, 2004

On Computation of *p*-Values in Parametric Linkage Analysis

Azra Kurbasic[a] Ola Hössjer[b]

[a]Division of Mathematical Statistics, Centre for Mathematical Sciences, Lund University, Lund, and
[b]Department of Mathematics, Stockholm University, Stockholm, Sweden

Key Words
Linkage analysis · Lod score distribution · Pointwise/genomewide *p*-value

Abstract
Parametric linkage analysis is usually used to find chromosomal regions linked to a disease (phenotype) that is described with a specific genetic model. This is done by investigating the relations between the disease and genetic markers, that is, well-characterized loci of known position with a clear Mendelian mode of inheritance. Assume we have found an interesting region on a chromosome that we suspect is linked to the disease. Then we want to test the hypothesis of no linkage versus the alternative one of linkage. As a measure we use the maximal lod score Z_{max}. It is well known that the maximal lod score has asymptotically a $(2 \ln 10)^{-1} \times (1/2\ \chi^2(0) + 1/2\ \chi^2(1))$ distribution under the null hypothesis of no linkage when only one point (one marker) on the chromosome is studied. In this paper, we show, both by simulations and theoretical arguments, that the null hypothesis distribution of Z_{max} has no simple form when more than one marker is used (multipoint analysis). In fact, the distribution of Z_{max} depends on the number of families, their structure, the assumed genetic model, marker denseness, and marker informativity. This means that a constant critical limit of Z_{max} leads to tests associated with different significance levels. Because of the above-mentioned problems, from the statistical point of view the maximal lod score should be supplemented by a *p*-value when results are reported.

Introduction

The aim of linkage analysis is to infer the position (locus) along one or several chromosomes, of a gene underlying or contributing to a certain trait, often related to a certain disease. Based on trait phenotype and DNA marker data from a number of families, this is done by estimating the relative positions of the trait loci along the chromosome(s). The DNA marker data give information about occurrence of crossovers at the regions close to the trait locus during meioses; markers cosegregate with the trait phenotypes in the families. Statistically, linkage analysis can be formulated as a hypothesis testing problem for testing the null hypothesis (H_0) that the trait locus is unlinked to the chromosomal region(s) of interest

KARGER

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

0001–5652/04/0574–0207$21.00/0

Accessible online at:
www.karger.com/hhe

Azra Kurbasic
Division of Mathematical Statistics, Centre for Mathematical Sciences
Lund University, Box 118
SE–221 00 Lund (Sweden)
Tel. +46 46 222 9272, Fax +46 46 222 4623, E-Mail azra@maths.lth.se

against the alternative (H_1) that it is located on one of the chromosomes. One can use either parametric or nonparametric methods. In this paper, a parametric method is one where the genetic model (mode of inheritance and disease allele frequency) is assumed to be known. (Although there are parametric models that are mode of inheritance free, cf. [1].) It is traditionally based on logarithms of likelihood ratios, so called lod scores, for testing H_0 against H_1. The nonparametric methods are typically based on allele sharing statistics and do not require knowledge of the genetic model. It is well known that parametric linkage analysis is more powerful than nonparametric when the underlying genetic model is known and true. On the other hand nonparametric linkage analysis is more robust against misspecification of the genetic model.

This paper addresses theoretical and practical aspects of parametric linkage analysis when we use the maximal lod score $Z_{\max}$ as test statistics. It is statistically important that the H_0 distribution of $Z_{\max}$ is (asymptotically) independent of a number of nuisance parameters such as genetic model, pedigree structures, marker data informativity and trait phenotypes. Otherwise, there is no natural correspondence between p-values and $Z_{\max}$, rendering statistical conclusions more difficult to draw when knowledge of the genetic model can be questioned. It is well known that the maximal lod scores have a well defined asymptotic limit distribution when only one marker is studied (two-point linkage analysis), cf. [2]. In this paper we show, both by theoretical arguments, simulations and a real data set, that the situation is completely different for maximal lod scores with many markers (multipoint linkage analysis). In this case, the relation between $Z_{\max}$ and the p-value depends a lot on several factors, such as number and form of pedigrees, marker informativity, and the assumed genetic model. The reason for the different statistical properties of $Z_{\max}$ in two-point and multipoint linkage analysis is that the parameter space for the parameter of interest is the recombination fraction θ in two-point analysis and disease locus position x in multipoint analysis. In the former case the parameter space is the interval [0, 0.5], with the H_0 parameter 0.5 as right end point. In the latter case the parameter space is not connected, and the H_0 parameter is an isolated point.

We also briefly discuss some alternatives to lod scores with asymptotic H_0 distributions that are independent of nuisance parameters for multipoint analysis. These include extensions of affected pedigree methods (APM) [3–6] and mod scores [7–9].

Parametric Linkage Analysis

One Marker

Let ψ be the assumed genetic model parameters (disease allele frequency and penetrance parameters) and θ the recombination fraction between the marker and disease locus. Furthermore, let Y and M denote the collection of disease phenotypes and marker data, respectively. Then the lod score

$$Z(\theta;\psi) = \log_{10} \frac{P(Y,M|\theta,\psi)}{P(Y,M|0.5,\psi)}$$

is used for testing $H_0 : \theta = 0.5$ against alternatives $\theta < 0.5$. The composite hypothesis testing problem uses the alternative $H_1 : \theta \in [0, 0.5)$. The total parameter space can be depicted as

with ◯ indicating H_0. Thus H_0 is a boundary point of the parameter space. The maximal lod score

$$Z_{\max}(\psi) = \sup_{0\le\theta\le 0.5} Z(\theta;\psi) = (2\ln 10)^{-1} 2\ln \frac{\sup_\theta P(Y,M|\theta,\psi)}{P(Y,M|0.5,\psi)} \qquad (1)$$

has asymptotically a $(2\ln 10)^{-1} \times (\frac{1}{2}\chi^2(0) + \frac{1}{2}\chi^2(1))$ distribution under H_0 as the number of pedigrees grows and when the genetic model is correctly specified. This is asymptotically independent of pedigree structure, marker informativity and disease phenotypes, although for finite samples the distribution will usually depend on such quantities to some extent, cf. Section 4.4 and 4.6 in [2]. Mixture of χ^2 distributions typically arise for log likelihood ratios under the null hypothesis when the H_0 parameter is a boundary point of a connected parameter space, cf. [10]. In (1) the same asymptotic limit distribution also appears when ψ is misspecified, cf. [11] and [12]. Hence the significance level

$$\alpha(T) = P_{H_0}(Z_{\max}(\psi) \ge T) \qquad (2)$$

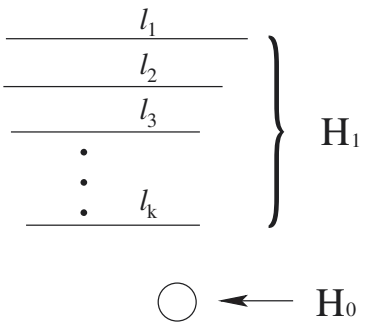
is asymptotically independent of the nuisance parameter ψ as well as ψ_{true}, the true value of the genetic model parameters. It is common practice to use a maximal lod score as a measure of significance rather than a p-value, cf. [2] and [1]. Since the asymptotic distribution of (1) is robust to model misspecification, there is asymptotically a one-to-one correspondence between $Z_{\max}$ and the p-value $\alpha(Z_{\max})$ in the single-marker case. Under mild regularity conditions, this is true also if the pedigrees are different, although larger sample sizes may then be required for the asymptotic approximation to be accurate, cf. [12].

Multipoint Linkage Analysis

Simultaneous analysis of several linked markers (multipoint analysis) is preferable to the single marker analysis when at least two markers are available on a chromosome, cf. [6]. The parameter of interest in multipoint analysis is the disease locus x measured in Morgans, and we write

$$Z(x;\psi) = \log_{10} \frac{P(Y, M|x, \psi)}{P(Y, M|\infty, \psi)} \tag{3}$$

where $x = \infty$ corresponds to H_0. The parameter space is no longer connected but can be depicted as



where K is the number of chromosomes and l_i the genetic length of chromosome i. The maximal lod score

$$Z_{\max}(\psi) = \max_{x \in H_1} Z(x, \psi) \tag{4}$$

does no longer converge to a limiting distribution under H_0 as the number of pedigrees grows. The reason is that the parameter space is no longer connected. This implies that the maximal lod score (4), as opposed to (1), has a degenerated limit distribution at $-\infty$ as the number of families $N \longrightarrow \infty$. For multipoint linkage analysis the significance (2) for a given threshold T depends heavily on ψ, the number of pedigrees, their structure, and the disease and marker phenotypes. In multipoint analysis, we need not consider ψ_{true} when computing p-values. The reason is that the H_0 distribution of $Z_{\max}$ depends only on ψ, not on ψ_{true}, since H_0 is an isolated point in parameter space.

Pointwise Distribution

The number of founders in the pedigree is denoted by f and the number of nonfounders by $n - f$, where n is the number of individuals in the pedigree. We also assume that founders are unrelated and carry $2f$ alleles that are not IBD (identical by descent). At one locus x in the genome the inheritance pattern can be represented by a *binary inheritance vector*, $v(x)$, defined as $v(x) = (p_1, m_1, p_2, m_2, ..., p_{n-f}, m_{n-f})$. The coordinates of the inheritance vector p_i and m_i describe the outcome of the paternal and maternal meioses. They are set to 0 or 1 if the i^{th} nonfounder's allele at position x originate from a grandfather or a grandmother. The probability distribution over possible inheritance vectors given marker data is referred to as the *inheritance distribution*. In the absence of any genotype information, the probability distribution of $v(x)$ is uniform over the set V of all $2^{2(n-f)}$ possible inheritance vectors (P_{uniform}). More details about inheritance vectors can be found in [6].

For two-point analysis (one marker and disease), we define the inheritance vector slightly differently. Let $v(\theta)$ be the inheritance vector of a locus at recombination fraction θ from the single marker. Although this locus is not unique (there are often two loci having the same recombination fraction to the marker), the marker gives exactly the same information about both of these inheritance vectors. Hence, as we will see below, this definition makes sense when defining two-point lod scores.

Parametric and nonparametric linkage analysis is defined in a unified framework in [6] using inheritance vectors. A scoring function S that depends on inheritance vector v ($= v(x)$ or $v(\theta)$) and the observed phenotypes Y is specified. It is a measure of compatibility between v and Y, i.e. the extent to which the phenotype vector Y can be explained by an inheritance vector v at the disease locus. In parametric linkage analysis, when the inheritance vector is known, $P(Y|v)$ is the likelihood of observed phenotypes Y in the pedigree conditioned on the inheritance vector v. The scoring function for one family with m meioses is

$$S(v) = \frac{P(Y|v)}{\sum_{w \in V} P_{\text{uniform}}(w) P(Y|w)} = \frac{P(Y|v)}{\sum_{w \in V} 2^{-m} P(Y|w)}, \tag{5}$$

where for simplicity, we omit the assumed genetic model parameters ψ in the notation. For two-point analysis, we define the distribution

$$P_\theta(w) = P(v(\theta) = w|M),$$

which quantifies the information the single marker yields at a recombination fraction θ away from it. Then the two-point lod score $Z(\theta)$ and likelihood ratio $LR(\theta)$ can be rewritten ([13]) as

$$Z(\theta) = \log_{10} LR(\theta) = \log_{10} \sum_w S(w) P_\theta(w).$$

For multipoint analysis we let

$$P_x(w) = P(v(x) = w|M)$$

quantify the information that the marker data gives about inheritance at locus x. Then the multipoint lod score and likelihood ratio can be written as

$$Z(x) = \log_{10} LR(x) = \log_{10} \sum_w S(w) P_x(w). \quad (6)$$

The total multipoint lod score for N families, finally, is obtained by summing the individual family lod scores,

$$Z(x) = \sum_{i=1}^{N} Z_i(x), \quad (7)$$

where Z_i is the i^{th} family score. For two-point analysis, (7) remains true if we replace x by θ.

The two extreme cases of marker informativity are perfect marker data ($LR(x) = S(v(x))$ for one pedigree) and no marker information at all (P_{uniform} for one pedigree). If we let E denote the expectation and $LR_p(x)$ the likelihood ratio for perfect marker data we see from (6) that

$$LR(x) = E(LR_p(x)|M).$$

From this and the fact that the H_0 distribution of $v(x)$ (in a population of many pedigrees) is P_{uniform} we get

$$E_{H_0}(LR(x)) = E_{H_0}(LR_p(x)) = 1 \quad (8)$$

$$\text{Var}_{H_0}(LR(x)) \leq \text{Var}_{H_0}(LR_p(x)) \text{ when the variances exist,} \quad (9)$$

where (9) follows from Jensen's inequality (cf. [14]). Similarly, we let $Z_p(x) = \log LR_p(x)$ be the lod score for perfect marker data. Then

$$E_{H_0}(Z_p(x)) \leq E_{H_0}(Z(x)) \leq 0, \quad (10)$$

is also deduced from Jensen's inequality. We also conjecture that in most cases

$$\text{Var}_{H_0}(Z(x)) \leq \text{Var}_{H_0}(Z_p(x)), \quad (11)$$

as a natural consequence of (9), although we have no general proof of the second inequality. In other words, there is less variation of the H_0 distribution of the lod score around the mean value for imperfect marker data. In the extreme case of no marker data $Z(x) = 0$.

For two-point analysis, there is no marker information at $\theta = 0.5$ ($Z(0.5) = 0$), whereas perfect marker information can arise only at $\theta = 0$ if the marker is fully polymorphic. Hence, $Z_{\max}$ is derived quite differently in two- and multipoint analysis. In the former case $Z(\theta)$ will be negative for most values of θ away from 0.5 under H_0 but $Z_{\max}$ is never negative. For multipoint analysis we maximize a function $Z(x)$ whose mean value under H_0 is negative for all x (unless there is no marker information somewhere), and this often implies that $Z_{\max}$ is negative as well.

Genomewide Significance Levels

In this section we describe methods for approximating the genomewide significance level $\alpha(T)$ for multipoint lod scores. The most straightforward method is to use Monte Carlo simulations, that is

$$\alpha(T) \approx \frac{1}{N_R} \sum_{i=1}^{N_R} I(Z^i_{\max} \geq T) \quad (12)$$

where $Z^i_{\max}$ are independent copies of $Z_{\max}$ under H_0 and N_R is the number of generated copies. Methods for simulating $Z^i_{\max}$ are described for example in Section 9.7 of [2]. The Monte Carlo method is very general, but can sometimes be slow, especially for large pedigrees and low marker informativity.

For perfect marker data ($Z = Z_p$) we can use analytical methods based on Gaussian extreme value theory to approximate $\alpha(T)$ as described for example in [15] and [16]. Notice first that $Z(x)$ as well as family scores $Z_i(x)$ are stationary processes under H_0 when marker information is perfect. We start assuming $P(Z(x) = -\infty) = 0$ at all x under H_0. This is typically true for most genetic models except for those with complete penetrance and no phenocopies. Let μ_i and σ_i denote the mean and standard deviation of the i^{th} family score $Z_i(x)$ under H_0. Then $\mu = \sum_{i=1}^{N} \mu_i$ and $\sigma = \sqrt{\sum_{i=1}^{N} \sigma_i^2}$ are the mean and standard deviation of $Z(x)$ under H_0. Further

$$Z(x) = \mu + \sigma \tilde{Z}(x), \quad (13)$$

where $\tilde{Z}(x) = \sum_{i=1}^{N} \sigma_i \tilde{Z}_i(x) / \sqrt{\sum_{i=1}^{N} \sigma_i^2}$ and $\tilde{Z}_i(x) = (Z_i(x) - \mu_i)/\sigma_i$. Notice that $\tilde{Z}$ and all $\tilde{Z}_i$ are stationary processes under H_0 with mean zero and variance one. For large sample sizes the central limit theorem implies that $\tilde{Z}$ approaches a Gaussian process with mean zero and unit variance. For this reason the genomewide significance level

$$\tilde{\alpha}(T) = P_{H_0}(\tilde{Z}_{\max} \geq T) \quad (14)$$

of the $\tilde{Z}$ process, with $\tilde{Z}_{\max} = \max_x \tilde{Z}(x)$, is much more stable with respect to variations of genetic model, pedigree structures, and number of pedigrees. Because of (13) we have

$$\alpha(T) = \tilde{\alpha}((T - \mu)/\sigma). \quad (15)$$

If the data set is more informative, $\mu < 0$ typically decreases and $\sigma > 0$ increases. The reason is that $LR(x) = 10^{Z(x)}$ under H_0 becomes more spread out around the expected value 1 for an informative data set. This happens, for instance, if the genetic model gets stronger, the number of pedigrees increases, the pedigrees become more informative or the marker informativity increases. (In latter case

the process Z is no longer stationary, so that $\mu = \mu(x)$ and $\sigma = \sigma(x)$ depend on x making (15) a bit oversimplified. For a fixed x however, we have the desired inequalities for $\mu(x)$ and $\sigma(x)$ for perfect versus imperfect marker data in (10) and (11).) Therefore, (15) implies that the p-value curve as function of T is flatter the more informative the data set is. Depending on T and the relation between μ and σ, a more informative data set has smaller or larger p-value than a less informative one.

Let $\tilde{T} = (T - \mu)/\sigma$. An approximate formula

$$\tilde{\alpha}(\tilde{T}) \approx 1 - e^{-\mu(\tilde{T})} \quad (16)$$

based on extreme value theory forGaussian processes,was suggested in [15] (see also [17]). Here $\mu(\tilde{T}) = (1-\Phi(\tilde{T}))(K+2\rho_{\tilde{Z}} l_{tot}\tilde{T}^2)$ approximates the average number of upcrossings of $\tilde{Z}$ over the level $\tilde{T}$, Φ is the cumulative distribution function of a standard normal $N(0,1)$, $\rho_{\tilde{Z}}$ the crossover rate of $\tilde{Z}$ and, $l_{tot} = l_1 + \ldots + l_K$ the total genomic length. The factor $(1 - \Phi(\tilde{T}))$ approximates the pointwise significance level $\alpha_{pt}(T) = P_{H_0}(Z(x) \geq T)$ (assuming that $\tilde{Z}(x) \sim N(0,1)$) and $C = \exp(K + 2\rho_{\tilde{Z}} l_{tot}\tilde{T}2)$ is a factor correcting for multiple testing. In [16], (16) was generalized in two ways. Firstly, a formula for the crossover rate $\rho_{\tilde{Z}}$ was derived for arbitrary pedigree structures. Secondly, an adjustment for non-normality in (16) was provided as follows: Let $\tilde{F}$ be the marginal distribution of $\tilde{Z}$ and $g = \tilde{F}^{-1} \circ \Phi$. The transformed process $Y(x) = g^{-1}(\tilde{Z}(x))$ then has marginal distribution Φ. The main idea of the adjusted normal approximation is that better accuracy is achieved if (16) is applied to Y rather than $\tilde{Z}$. With $Y_{\max} = \max_x Y(x)$, the adjusted normal approximation becomes

$$\tilde{\alpha}(\tilde{T}) = P_{H_0}(Y_{\max} \geq g^{-1}(\tilde{T})) \approx 1 - e^{-\mu_{adj}(g^{-1}(\tilde{T}))}, \quad (17)$$

where $\mu_{adj}(g^{-1}(\tilde{T})) = \alpha_{pt}(T)C_{adj}$, $\alpha_{pt}(T) = P_{H_0}(Y(x) \geq g^{-1}(\tilde{T})) = 1-\Phi(g^{-1}(\tilde{T}))$ is the pointwise significance level defined below (16) and $C_{adj} = \exp(K + 2\rho_Y l_{tot}(g^{-1}(\tilde{T}))^2)$ is an adjusted multiple testing factor.

In case $\varepsilon = P(Z(x) - \infty) > 0$, we proceed by defining (13) as well as μ_i and σ_i conditionally on the event $Z(x) > -\infty$. The procedure is similar except that we multiply $\mu(\tilde{T})$ and $\mu_{adj}(g^{-1}(\tilde{T}))$ by a factor $(1 - \varepsilon)$ to account for the fact that the average number of upcrossings is reduced by this factor.

Results

Standardized lod scores $\tilde{Z}(x)$ turned out very useful when calculating p-values in parametric linkage analysis. We calculated p-values using Monte Carlo simulations (12), and for perfect marker data also the normal approximation formula (16), and adjusted normal approximation formula (17). We present some figures with the p-values from the multipoint analysis of a real data set containing pedigrees ICEL80002 and ICEL80004, cf. Figure 3 and 4. These are two Icelandic breast cancer families that were part of the BRCA2 linkage studies, cf. [18] and Figure 2. For the simulated data set we used four different pedigree structures and five different genetic models, cf. Figure 2 and Table 1, in the study. Figures 5-6 and Figure 7 display the calculated genomewide p-values for Model 3 and Model 4, respectively for Pedigrees 1 - 4 as a function of threshold T. It is assumed that the genome scan consists of the 22 autosomes, with sex-averaged chromosome lengths as in [2], Table 1.2. As discussed in Section 4, formula (15) suggests that a more informative data set has a flatter p-value curve $\alpha(T)$. Comparing models in the Figure 3, we find that the stronger Model 1 with a lower rate of phenocopies flattens the p-value curve compared to Model 2. Genetic models 3 and 4 with the same penetrances but different disease allele frequencies are compared in Figures 5 and 7. The changed disease allele frequency has a little effect on the small Pedigrees 1 and 2 with only two founders. The reason is that it is very unlikely with more than one disease allele among the founders, even for the weaker Model 4 with larger disease allele frequency 0.1. For Pedigrees 3 and 4, there are several unaffected founders of low age. This makes the inheritance pattern more uncertain for Model 3, since several founders may be disease allele carriers. The analogous conclusions are valid for models 1 and 5 and the real data set containing large pedigrees, cf. Figure 3. Comparing Figures 5 and 6, with two different sample sizes ($N = 60$ versus $N = 180$), we find that increased sample size results in a flatter and lower p-value curve for all four pedigrees. It is also seen that the multigenerational Pedigree 3 results in much lower p-value curve (that is, a smaller $\alpha(T)$).

Table 1. Summary of details about the models. Affection status locus type is autosomal disease. The penetrance value f_i is the conditional probability that an individual is affected given the genotype with i disease alleles at the disease locus.

Model	Disease allele frequency	Penetrance values		
		f_0	f_1	f_2
1	0.1	0.001	0.5	0.8
2	0.1	0.2	0.5	0.8
3	0.0033	age dependent penetrance		
4	0.1	age dependent penetrance		
5	0.0033	0.001	0.5	0.8

Penetrance values from the standard genetic model for breast cancer presented in [24], see table 2.

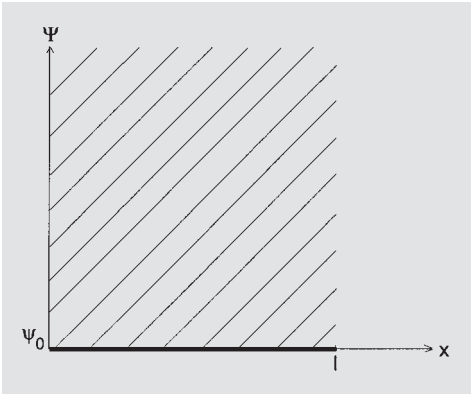


Fig. 1. The total parameter space corresponds to the shaded area in the figure and H_0 to the line $[0, l] \times \{\psi_0\}$.

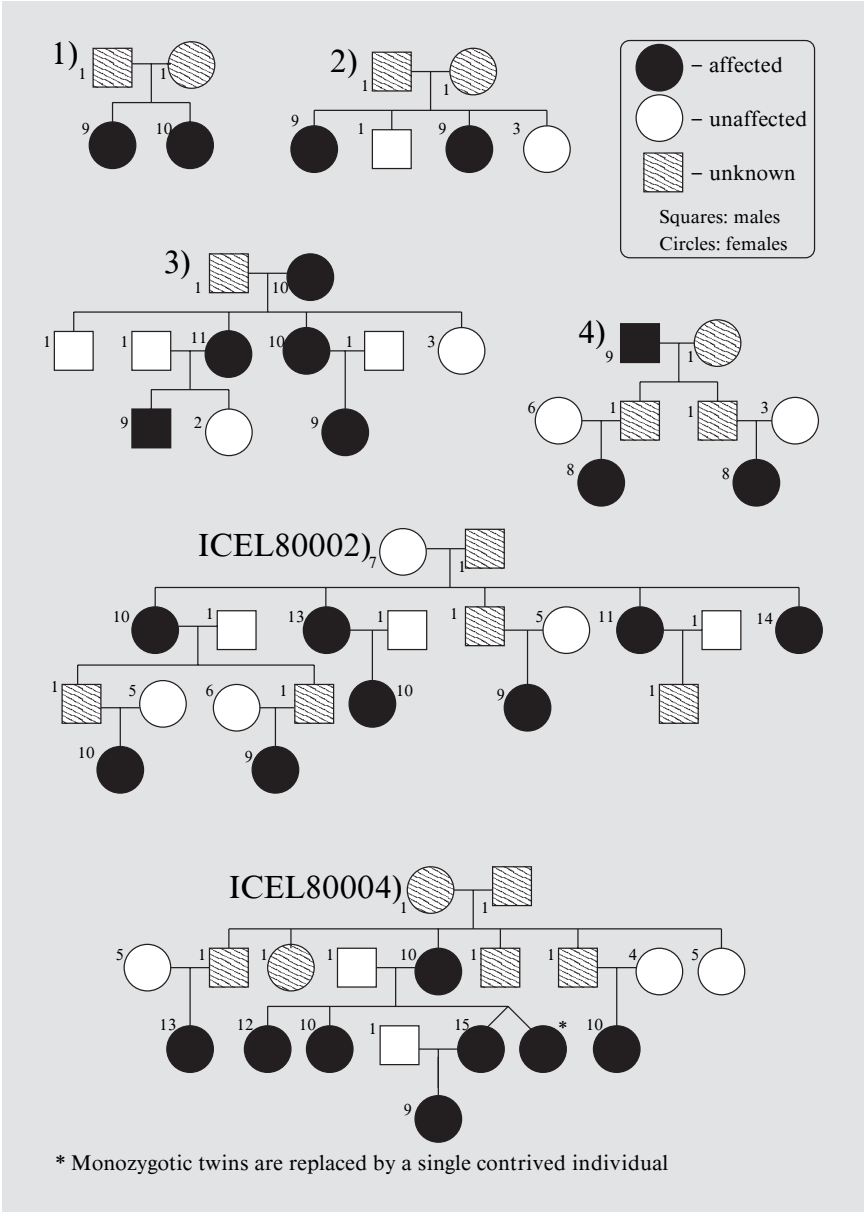


Fig. 2. Pedigrees used in the study. Each individual is assigned to one of 15 liability classes, indicated in the figure, depending on age and affection status. Individuals whose parents are members of the pedigree are called nonfounders and those whose parents are not members of the pedigree are known as founders. Pedigrees ICEL8002 and ICEL80004 are family 2 and reduced family 4, respectively on page 751 in [18].

The effect of marker informativity is a bit more delicate to interpret. In Figures 8-9, the effect of marker heterozygosity and marker spacing is shown for Models 1, 2, and 5, and Pedigree 1. For all three models, the p-value curve is wider the more informative marker data is. The p-value curve essentially gets lower with increased marker informativity for the strongest Model 5, whereas the opposite is true for the other two Models 1 and 2. This shows that perfect marker data approximations for lod score p-values can be either conservative or anticonservative, depending on the genetic model. The effect of marker informativity is stronger for the larger Pedigree 4 in Figure 10. A difference compared to Figure 8 is that Pedigree 4 is more informative than Pedigree 1 for Models 1 and 5, making the p-value curve flatter and lower.

It follows from (17) that the genomewide p-value at least for perfect marker data is approximately a function of the pointwise p-value $\alpha_{pt}(T)$. For this reason, we also studied how $\alpha_{pt}(T)$ depends on the informativity of the data set. The results are analogous to those for genomewide p-values. A technical report with results and figures can be obtained from the authors.

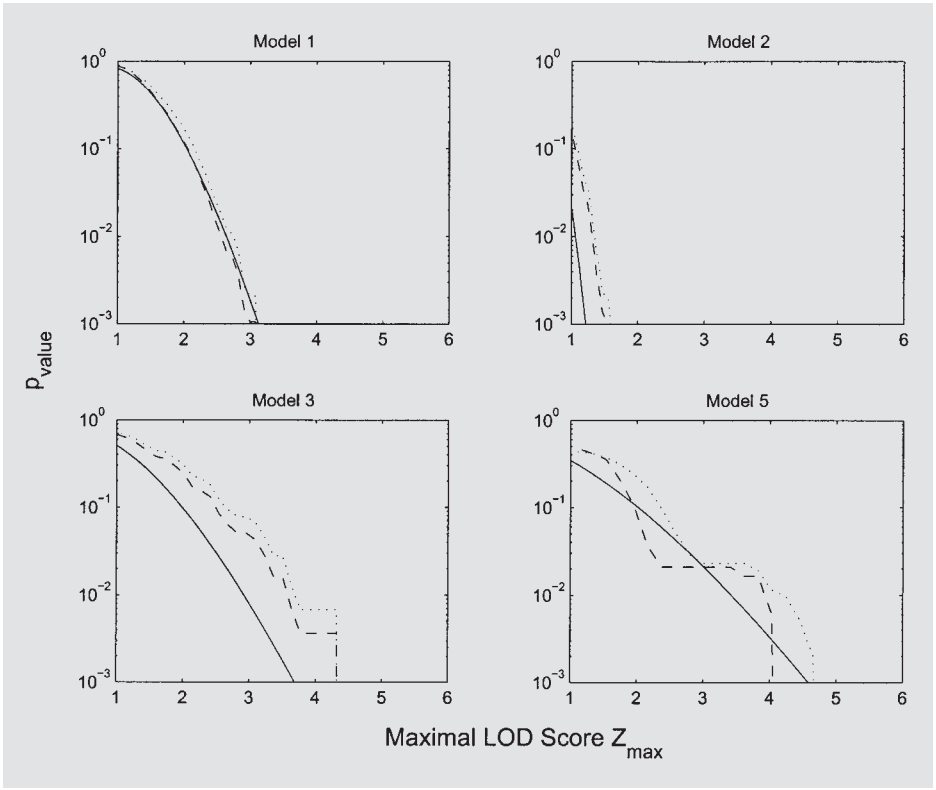


Fig. 3. Comparisons between the *p*-values in a multipoint linkage analysis plotted against the maximal lod score for the normal approximation (——), the simulation procedure (– – –) given by (12) using 100,000 replicates and the adjusted normal approximation (· · · · ·) for a data set containing families ICEL80002 and ICEL80004 and four different models. Marker data is perfect and the chromosome region is 6.43 cM long.

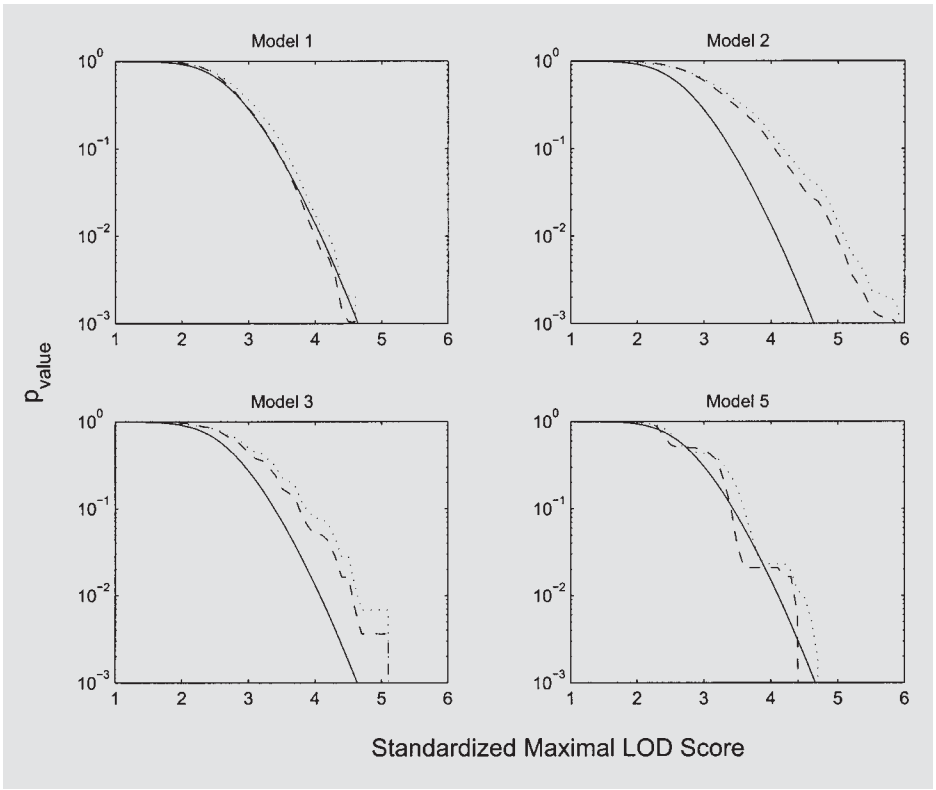


Fig. 4. Comparisons between the *p*-values in a multipoint linkage analysis plotted against the standardized maximal lod score for the normal approximation (——), the simulation procedure (– – –) given by (12) using 100,000 replicates and the adjusted normal approximation (· · · · ·) for a data set containing families ICEL80002 and ICEL80004 and four different models. Marker data is perfect and the chromosome region is 6.43 cM long.

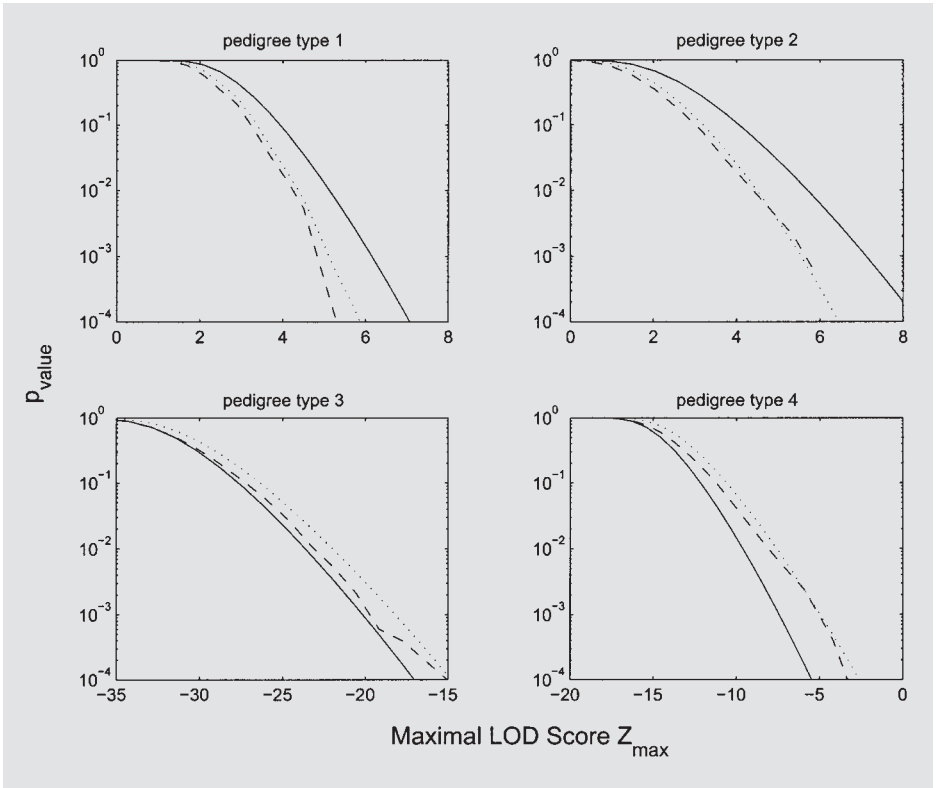


Fig. 5. Comparisons between the genomewide *p*-values for the normal approximation (——), the simulation procedure (– – –) given by (12) using 10,000 replicates and the adjusted normal approximation (· · · · ·) for Model 3 and 60 families for each pedigree type. Marker data is perfect.

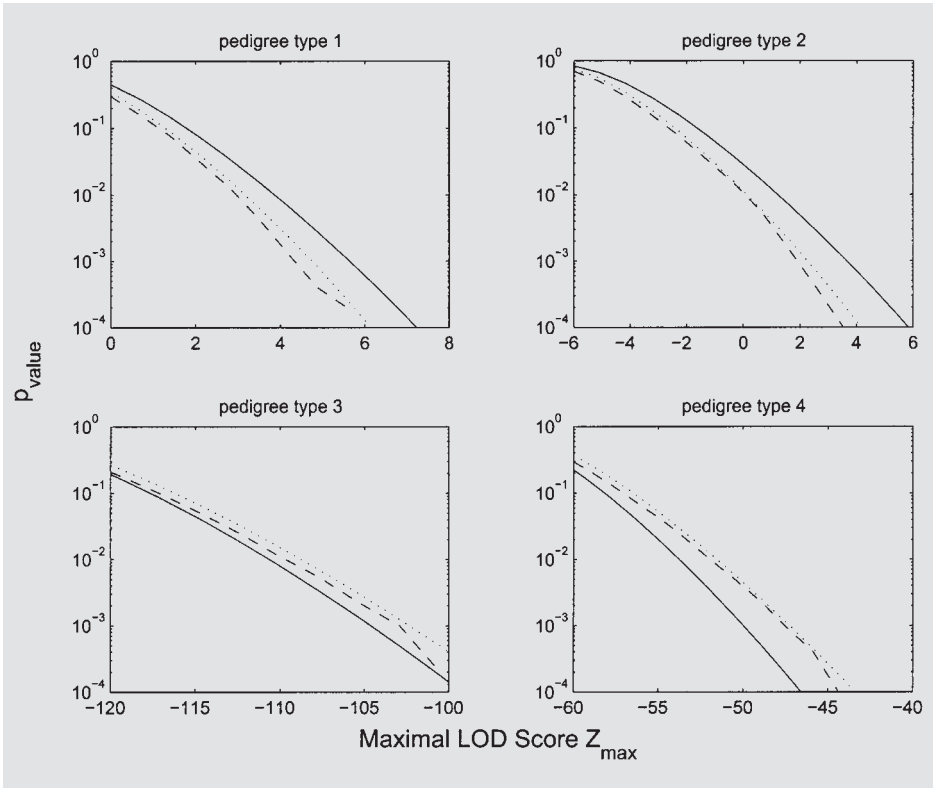


Fig. 6. Comparisons between the genomewide *p*-values for the normal approximation (——), the simulation procedure (– – –) given by (12) using 10,000 replicates and the adjusted normal approximation (· · · · ·) for Model 3 and 180 families for each pedigree type. Marker data is perfect.

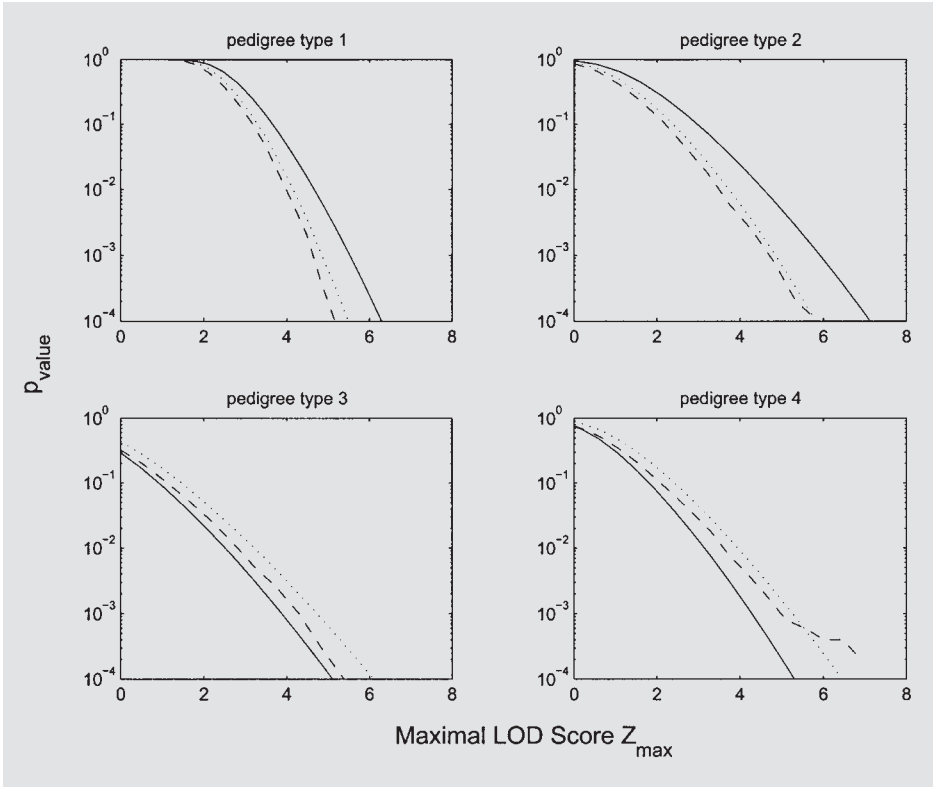


Fig. 7. Comparisons between the genomewide *p*-values for the normal approximation (——), the simulation procedure (– – –) given by (12) using 10,000 replicates and the adjusted normal approximation (· · · · ·) for Model 4 and 60 families for each pedigree type. Marker data is perfect.

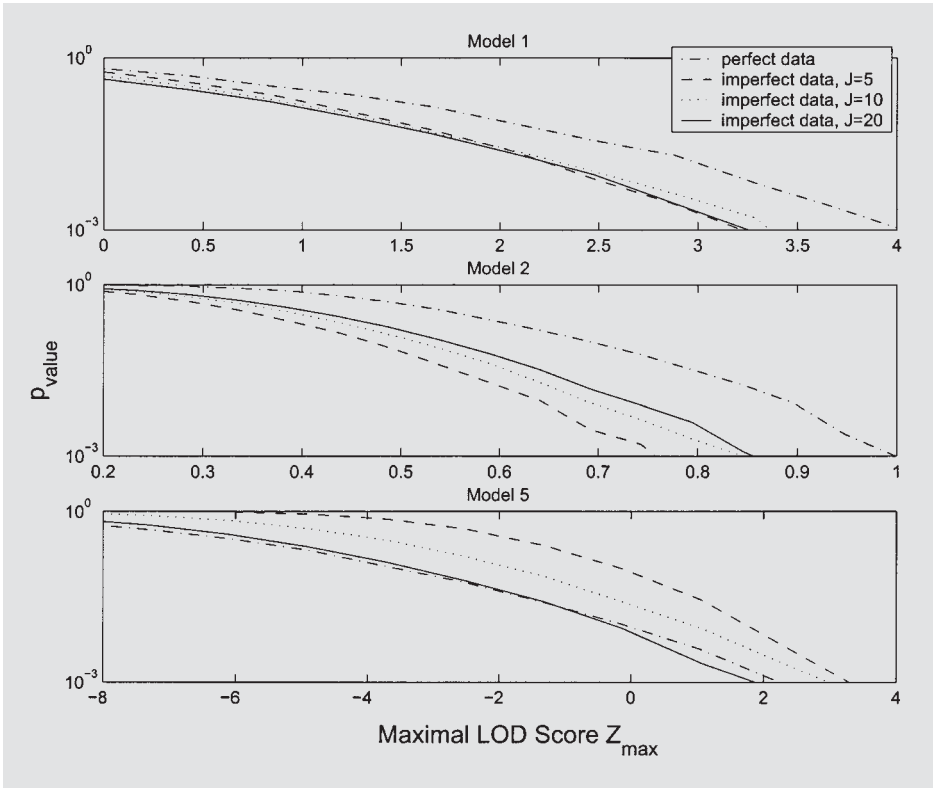


Fig. 8. Comparisons between the chromosomewide (chromosome 1, length 298.5 cM) *p*-values for perfect and imperfect data for 60 families of pedigree type 1 for the simulation procedure given by (12) using 10,000 replicates. Simulations are done for three different models. All markers have *J* possible alleles with equal probability 1/*J*. Distance between markers for imperfect marker data is 10 cM. Only nonfounders are genotyped when marker data is incomplete.

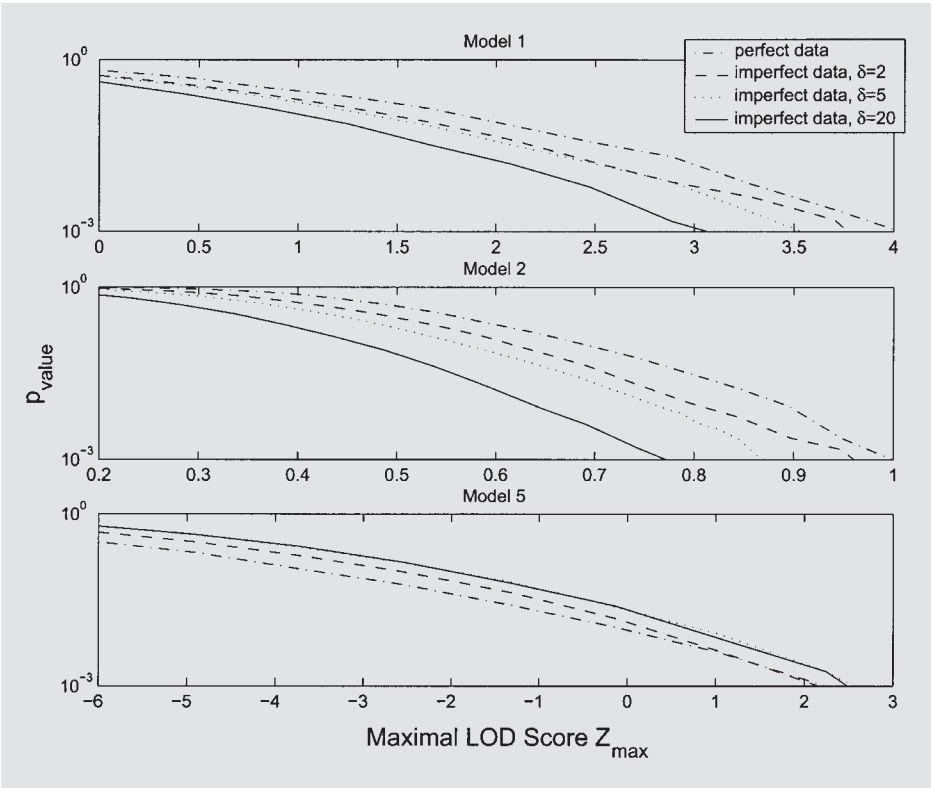


Fig. 9. Comparisons between the chromosomewide (chromosome 1, length 298.5 cM) *p*-values for perfect and imperfect data for 60 families of pedigree type 1 for the simulation procedure given by (12) using 10,000 replicates. Simulations are done for three different models. All markers have $J = 10$ possible alleles with equal probability 1/10. For imperfect data, the markers are equally spaced with distance δ cM and only nonfounders are genotyped.

Table 2. Penetrance values used in the study for models three and four. Seven age groups × two disease classifications, affected and unaffected, were used. Unaffected males and individuals with unknown affection status were assigned to liability class one. Disease and normal allele at the disease locus are denoted by *d* and *D*, respectively. They give rise to three different genotypes.

Liab. class	Age group	Penetrance of genotype		
		dd	Dd	DD
Cumulative risk for unaffected females				
1	<30	0.00009	0.008	0.008
2	30–39	0.00146	0.083	0.083
3	40–49	0.0083	0.269	0.269
4	50–59	0.0210	0.469	0.469
5	60–69	0.0390	0.616	0.616
6	70–79	0.0610	0.724	0.724
7	≥80	0.0820	0.801	0.801
Density for affected females				
8	<30	0.00002	0.00167	0.00167
9	30–39	0.00026	0.01276	0.01276
10	40–49	0.00112	0.02305	0.02305
11	50–59	0.00137	0.01711	0.01711
12	60–69	0.00226	0.01260	0.01260
13	70–79	0.00218	0.00908	0.00908
14	≥80	0.00213	0.00654	0.00654
Product of the penetrances for the monozygotic twins				
15		0.00000029	0.000294	0.000294

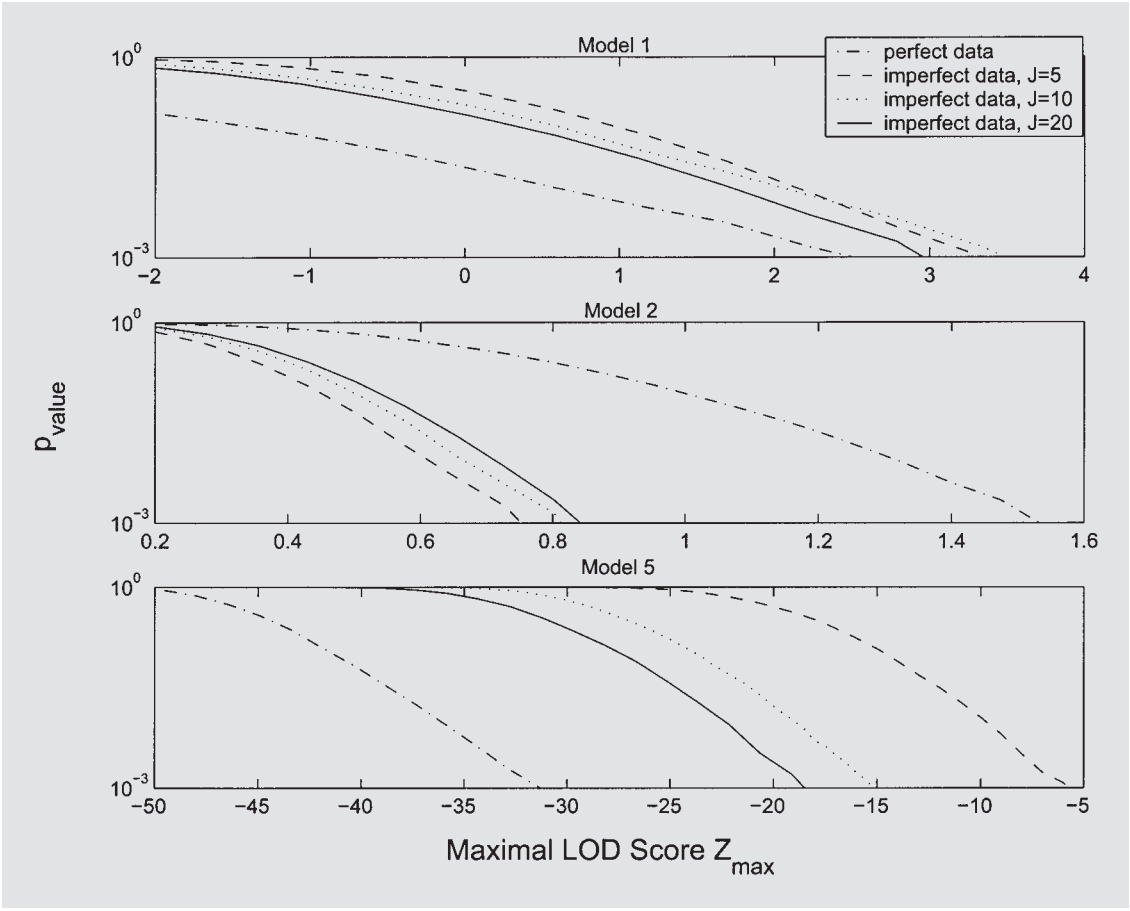


Fig. 10. Comparisons between the chromosomewide (chromosome 1, length 298.5 cM) *p*-values for perfect and imperfect data for 60 families of pedigree type 4 for the simulation procedure given by (12) using 10,000 replicates. Simulations are done for three different models. All markers have *J* possible alleles with equal probability 1/*J*. Distance between markers for imperfect marker data is 10 cM. Only nonfounders are genotyped when marker data is incomplete.

Summary and Other Approaches

An essential issue in this paper was to find out and explain the behaviour of *p*-values based on maximal lod scores. For multipoint linkage analysis we have found, both by extensive simulations and theoretical arguments, that the H_0 distribution of pointwise and maximal lod scores depend heavily on genetic model parameters, number and structure of pedigrees, phenotypes, and marker informativity. For this reason the *p*-value is a more appropriate performance measure than the maximal lod score itself.

It is possible to use alternative approaches in parametric linkage analysis than traditional lod scores whose $Z_{\max}$ distributions under H_0 are less sensitive to variations of nuisance parameters. One possibility is to define a mod score ([7], [8], [9])

$$Z(x) = \sup_{\psi} Z(x;\psi) \tag{18}$$

by maximizing (3) over a predefined set of genetic model parameters. In this case the parameter (x, ψ) contains both the disease locus x and genetic model parameters ψ, of which the latter are nuisance parameters. Let ψ_0 be a set of genetic model parameters corresponding to no genetic effect. For a chromosome $[0, l]$ of length l, the null hypothesis is no longer formulated as $x = \infty$ but rather as 'a disease locus' $x \in [0, l]$ with no genetic component $(\psi = \psi_0)$. To be precise, x is a disease locus only when there is a genetic component, but we may think of (x, ψ_0) as the limit of (x, ψ) when $\psi \to \psi_0$. This parameter space, illustrated in Figure 1, is connected if the genetic model parameter space ψ is connected. As a result, the distribu-

tion of $Z(x)$ under H_0 is asymptotically free from nuisance parameters. Depending on the set of genetic models $\{\psi\}$ locally around ψ_0 and whether or not ψ_0 is at the boundary of $\{\psi\}$, the limit distribution can be either a χ^2 distribution or a mixture of χ^2 distributions, cf. [10], [19], and [20]. Hence, at least the pointwise p-value $\alpha_{pt}(T)$ is asymptotically free from nuisance parameters.

The nonparametric linkage method in Genehunter ([6]) is based on an allele sharing score function $S(v)$ and NPL score

$$Z(x) = \sum_w S(w)P_x(w), \tag{19}$$

for one pedigree with P_x as defined in Section 3. The score function S is standardized so that $Z(x)$ has zero mean and unit variance under H_0 when the marker information is perfect. This method originally goes back to the affected pedigree method in [3], although these authors focused on identity by state (IBS) rather than IBD sharing. It is also possible to define S for genetic models with other phenotypes than affected/nonaffected, cf. [9], [20], [21] and [22] and use it for parametric linkage analysis. In fact, a large class of scores (19) can be obtained by differentiating the lod score (3) with respect to ψ at ψ_0. In that case, $Z(x)$ can be interpreted as a score test version of the mod score (18), which is a profile likelihood curve with ψ being the profiled set of parameters.

When several family scores (19) are added and standardized to have unit variance under H_0, the total NPL score is asymptotically normally distributed and quite insensitive to variation of genetic model parameters. A modified version of (19) suggested in [23] makes the H_0 distribution less sensitive to variation in marker informativeness as well. As a result, the pointwise p-value $\alpha_{pt}(T)$ is asymptotically free of nuisance parameters. However, the same is not true for the genomewide p-value $\alpha(T)$, since the effective amount of multiple testing depends for example on pedigree structure, see [17], [15], and [16].

As a practical implication of our work, we suggest, if traditional lod scores are to be used in multipoint linkage analysis, that they are reported together with p-values. An alternative is that the standardized version $\tilde{Z}$ from Section 4 is plotted and used in the analysis, cf. Figure 4. It is in fact an NPL score with S as in (5), standardized to have mean zero and unit variance under H_0 ($S \leftarrow (S-\mu_i)/\sigma_i$ for family i) and with family weights $\sigma_i/\sqrt{\sum_{i=1}^N \sigma_i^2}$, cf. (13). For imperfect marker data, we define each family score according to (19) or use the approach in [23]. How $\tilde{Z}$ compares with mod scores and NPL scores with other S in terms of power and robustness to model misspecification will depend on both the true and assumed genetic model, pedigree structure(s), sample size, etc. This is certainly a topic for future research.

As a final illustration showing the usefulness of the $\tilde{Z}$ score, we calculated p-values for the real dataset assuming marker data is perfect, cf. Figures 3 and 4, and found that the p-value curves for Z vary much more between the models than for $\tilde{Z}$, although we have only two families and the Central Limit Theorem approximation is expected to be inaccurate. In the real dataset 64% of individuals were typed for four markers, D13S1246, D13S260, D13S171, and D13S267. The maximal lod score obtained in multipoint analysis (with Genehunter, cf. [6]) for the breast cancer model (Model 3) is 2.99. Corresponding standardized maximal lod score $\tilde{Z}$ is 4.13. We approximated the p-value for this threshold with the value 0.051 obtained from the simulation procedure (12) and assuming perfect marker data. Corresponding p-values obtained for models 1, 2, and 5 and lod score 2.99 are 0.001, 0.000, and 0.021, respectively. These should be compared to the p-values 0.01, 0.08, and 0.04 for the standardized threshold 4.13. This illustrates our findings that standardized lod scores $\tilde{Z}$ are more closely related to p-values than unstandardized ones.

Acknowledgements

The first author was sponsored by the National Research School in Genomics and Bioinformatics and the second one by the Swedish Research Council, contract nr 626-2002-6286. We wish to thank Pär-Ola Bendahl for discussing and suggesting improvements on the manuscript and Adalgeir Arason and Rosa Björk Barkardottir for providing us with the real dataset. We are also greatful for two reviewers' helpful comments.

References

1 Sham P: Statistics in Human Genetics. New York, Arnold applications of Statistics, 1998.

2 Ott J: Analysis of Human Genetic linkage, ed 3. Baltimore, Johns Hopkins University Press, 1999.

3 Weeks DE, Lange K: The affected-pedigree-member method of linkage analysis. Am J Hum Genet 1988;42:315–326.

4 Fimmers R, Seuchter SA, Neugebauer M, Knapp M, Baur MP: Identity by descent analysis using all genotype solutions; in Elston RC, Spence MA, Hodge SE, MacCluer JW (eds): Multipoint Mapping and Linkage Based on Affected Pedigree Members: Genetic Analysis Workshop 6. New York, Alan R. Liss, 1989.

5 Whittemore A, Halpern J: A class of tests for linkage using affected pedigree members. Biometrics 1994;50:118–127.

6 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.

7 Risch N: Segregation analysis incorporating genetic markers. I. Single-locus models with an application to type I diabetes. Am J Hum Genet 1984;36:363–386.

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

9 Whittemore AC: Genome scanning for linkage: An overview. Am J Hum Genet 1996;59:704–716.

10 Self SG, Liang KY: Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc 1987;82:605–610.

11 Dudoit S, Speed TP: A score test for linkage using identity by descent data from sibships. Ann Stat 1999;27:943–986.

12 Williamson JA, Amos CI: On the asymptotic behaviour of the estimate of the recombination fraction under the null hypotesis of no linkage, when the model is misspecified. Genet Epidemiol 1990;7:309–318.

13 Morton NE: Seqential tests for the detection of linkage. Am J Hum Genet 1995;7:277–318.

14 Royden HL: Real Analysis, ed 2. New York, Macmillan Publishing, 1968.

15 Lander E, Kruglyak L: Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. Nat Genet 1995;11:241–247.

16 Ängquist L, Hössjer O: Improving the Calculation of Statistical Significance in Genome-Wide Scans. To appear in Biostatistics.

17 Feingold E, Brown PO, Siegmund D: Gaussian models for genetic linkage analysis using complete high resolution maps of identity by descent. Am J Hum Genet 1993;53:234–251.

18 Gundmundsson J, Johanesdottir G, Arason A, Bergthorasson JT, Ingvarsson S, Egilsson V, Barkardottir RB: Frequent occurance of BRCA2 linkage in Icelandic breast cancer families and segregation of a common BRCA2 haplotype. Am J Hum Genet 1996;58:749–756.

19 Rotnitzky A, Cox DR, Bottai M, Robins J: Likelihood-based inference with singular information matrix. Bernoulli 2000;6:243–284.

20 Hössjer O: Conditional likelihood score function in linkage analysis. To appear in Biostatistics.

21 McPeek MS: Optimal allele-sharing statistics for genetic mapping using affected relatives. Genet Epidemiol 1999;16:225–249.

22 Hössjer O: Determining inheritance distributions via stochastic penetrances. Preprints in mathematical sciences, Centre for Mathematical Sciences, Lund University, 2001. J Am Stat Assoc 2003;98:1035–1051.

23 Kong A, Cox NJ: Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 1997;61:1179–1188.

24 Claus EB, Rish NJ, Thompson WD: Age at onset of familial risk of breast cancer. Am J Epidemiol 1990;131:961–967.