Mathematical Statistics
Stockholm University

Relative Risks for General Phenotypes and Genetic Models

Azra Kurbasic and Ola Hössjer

Research Report 2005:4

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2005:4**,
http://www.math.su.se/matstat

Relative Risks for General Phenotypes and Genetic Models

Azra Kurbasic * and Ola Hössjer†

March 2005

Abstract

Many common diseases are known to have genetic components but since they are non-Mendelian, i.e. a large number of genetic factors effect the phenotype, these components are difficult to localize. These traits are often called complex and analysis of siblings is a valuable tool for mapping. It has been shown that the power of affected relative pairs method to detect linkage of a disease susceptibility locus depends on the locus contribution to increased risk of relatives compared with population prevalence [Risch, 1990a and 1990b]. In this paper we generalize calculation of relative risk to arbitrary phenotypes and genetic models but also show that the relative risk can be split into the relative risk at the main locus and the relative risk due to interaction between the main locus and loci at other chromosomes. In order to achieve power to detect linkage a certain number of relative pairs has to be collected. To be able to quantify the amount of information of the relative pair we use the effective number of meioses introduced in Hössjer [2004], which is closely related to power to detect linkage. Relative risks and effective number of meioses are computed for several genetic models with binary or quantitative phenotypes, with or without polygenic effects.

KEY WORDS: Complex diseases, relative risk, linkage analysis, effective number of meioses.

*Mathematical Stastistics, Centre for Mathematical Sciences, Lund University, Box 118, 221 00 LUND, Sweden. E-mail: azra@maths.lth.se.

†Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: ola@math.su.se. Financial support from the Swedish Research Council, contract nr. 626-2002-6286.

1 Introduction

An important class of traits for study in humans are those caused by multiple loci which, in general, have a chronic and highly prevalent nature. Some examples are cancer, epilepsy, and psychiatric disorders. They are often called quantitative since the disease phenotype is typically measured on a continuous scale or it is defined by thresholds to a continuous variable. Classical linkage analysis has low power in analyzing complex diseases since there is no one-to-one correspondence between the genotype and disease phenotype which is typical for Mendelian diseases. To be able to detect linkage to a disease susceptibility locus, large samples of pedigrees are needed, since each disease gene typically has a small effect by itself. For this reason, relative pair families, most notably sib pairs, have been a valuable tool, since they are relatively easy to collect in large quantities.

Before undertaking a linkage study, one must have some understanding of how many relatives are needed for obtaining evidence for linkage. The fundamental assumption is that the linkage information in pairs can be studied through the relationship between their identity-by-descent sharing (IBD) at the putative locus (loci) and phenotypes. In a series of papers Risch [1990a, b] studied relative risk ratios for binary traits. He showed that these risk ratios are the essential parameters for determining power to detect linkage when using affected relative pairs, in that probabilities of IBD-sharing depend on the genetic model only trough relative risk ratios. For quantitative traits, the sib pair's phenotypes were shown to play an important role. Carey and Williamson [1991] and Cardon and Fulker [1994] noticed the value of selecting sibs with extreme phenotypes. Risch and Zhang [1995, 1996] observed that sib-pairs with extremely discordant or highly concordant phenotypes provide the greatest power to detect linkage.

In this paper we study relative risk for arbitrary phenotypes and genetic models. We assume that the trait/phenotype is a result of one so called main locus but also other sources of covariation, i.e. polygenes unlinked to the main locus and shared environment. By splitting the relative risk into the risk at the main locus and the risk due to other sources, it is possible to use the relative risk component at the main locus and calculate IBD-sharing probabilities and effective number of meioses, c.f. Hössjer [2004]. It determines the amount of information for linkage that the relative pair contains. We compute relative risks and effective number of meioses for several different types of genetic models, including binary and quantitative traits, with or without polygenes.

Our approach is conditional on observed phenotypes. Many authors have realized that treating the number of alleles shared IBD at the marker locus (loci) as dependent and the sib trait values as independent variable has several advantages, since sample selection is often through trait values but

almost never through marker genotypes, see f.e. [Risch and Zhang, 1995 and 1996; Dudoit and Speed, 2000; Kraft and Thomas, 2000]. This is in contrast to the regression method of Haseman and Elston [1972] and variance components techniques [Amos, 1994], where IBD-sharing is the independent variable and phenotypes the dependent variable.

2 Conditional Variances and Covariances

Consider a trait which is influenced by a gene having two possible alleles. If the trait is related to a certain disease, we think of the two alleles as the normal (0) and disease (1) allele respectively, having probabilities q and p, $q+p=1$. Let Y_1 and Y_2 be the observed phenotypes for two individuals and $f(Y_k)$ and $f(Y_1,Y_2)$ the marginal and joint probability (density) functions respectively of Y_1 and Y_2. Each Y_k may be scalar or vector-valued, and may include covariates. The three possible genotypes (00), (01) and (11) give rise to three penetrance values $\psi_0^{(k)} = f(Y_k|(00))$, $\psi_1^{(k)} = f(Y_k|(01))$ and $\psi_2^{(k)} = f(Y_k|(11))$ for each individual k. Under Hardy-Weinberg equilibrium,

$$f(Y_k) = E(f(Y_k|G_k)) = q^2\psi_0^{(k)} + 2pq\psi_1^{(k)} + p^2\psi_2^{(k)},$$

where Y_k is kept fixed and expectation is with respect to G_k, the genotype of k. Similarly,

$$\begin{aligned}\sigma_g^{(k)} &:= \mathrm{Var}(f(Y_k|G_k)) \\ &= q^2(\psi_0^{(k)} - f(Y_k))^2 + 2pq(\psi_1^{(k)} - f(Y_k))^2 + p^2(\psi_2^{(k)} - f(Y_k))^2,\end{aligned}$$

which we refer to as the conditional genetic variance of the penetrance for k. It can be split into conditional additive and dominance genetic variance components $\sigma_g^{(k)} = \sigma_a^{(k)} + \sigma_d^{(k)}$, where

$$\begin{aligned}\sigma_a^{(k)} &= 2pq\left(p(\psi_2^{(k)} - \psi_1^{(k)}) + q(\psi_1^{(k)} - \psi_0^{(k)})\right)^2 \\ \sigma_d^{(k)} &= (pq)^2\left(\psi_2^{(k)} - 2\psi_1^{(k)} + \psi_0^{(k)}\right)^2,\end{aligned}$$

c.f. Elston et al [2002].

In order to get an expression for $f(Y_1,Y_2)$, we introduce the 3×3 matrix $\psi = \{\psi_{jl}\}_{j,l=0}^2$ of joint penetrances, i.e. ψ_{jl} is the value of $f(Y_1,Y_2|G_1,G_2)$ when G_1, 1's genotype, has j disease alleles and G_2, 2's genotype, has l disease alleles. Define

$$\sigma_g^{(12)} := E(f(Y_1,Y_2|G,G)) - E(f(Y_1,Y_2|G,G')), \tag{1}$$

where Y_1 and Y_2 are fixed and expectation is with respect to two *independent* genotypes G and G'. We refer to $\sigma_g^{(12)}$ as the conditional genetic covariance of the joint penetrances. This name is motivated by considering the special case when there is no contribution to the trait from other genes, polygenes or shared environment, i.e. $\psi_{jl} = \psi_j^{(1)}\psi_l^{(2)}$. Then $\sigma_g^{(12)} = \mathrm{Cov}(f(Y_1|G_1), f(Y_2|G_2))$ for a monozygotic twin pair (1,2) and $\sigma_g^{(kk)} = \sigma_g^{(k)}$. It is shown in the appendix that a decomposition $\sigma_g^{(12)} = \sigma_a^{(12)} + \sigma_d^{(12)}$ into additive and dominant conditional genetic covariances is possible, with

$$\begin{array}{rcl} \sigma_a^{(12)} & = & 2pq \cdot u_a \psi u_a', \\ \sigma_d^{(12)} & = & (pq)^2 \cdot u_d \psi u_d', \end{array} \tag{2}$$

$u_a = (-q, q-p, p)$, $u_d = (1, -2, 1)$ and u_a' and u_d' the transpose of u_a and u_d respectively.

3 Relative Risks

We indicate the type of relationship between 1 and 2 with R. For a pair R, the relative risk ratio

$$\lambda_R = \frac{f_R(Y_1|Y_2)}{f(Y_1)} = \frac{f_R(Y_2|Y_1)}{f(Y_2)} = \frac{f_R(Y_1, Y_2)}{f(Y_1)f(Y_2)},$$

quantifies the relative change in density for Y_1 after observing Y_2 or vice versa. Let $I \in \{0, 1, 2\}$ be the number of alleles shared identical by descent by (1,2) at the trait locus and put $f_{Ri}(Y_1, Y_2) = f_R(Y_1, Y_2|I = i)$[1]. Then decompose the relative risk as

$$\lambda_R = \lambda_R^{\text{other}} \cdot \lambda_R^{\text{main}},$$

where $\lambda_R^{\text{main}} = f_R(Y_1, Y_2)/f_{R0}(Y_1, Y_2)$ is the contribution to relative risk because of allele sharing identical by descent at the trait locus and $\lambda_R^{\text{other}} = f_{R0}(Y_1, Y_2)/(f(Y_1)f(Y_2))$ represents relative risk due to other sources of covariation, such as other genes, polygenes and shared environment. Notice that

$$f_{R0}(Y_1, Y_2) = u\psi u', \tag{3}$$

with $u = (q^2, 2pq, p^2)$, equals the second term on the RHS of (1).

[1]When R is a monozygotic twin pair or a parent-offspring pair the event '$I = 0$' has probability zero and $f_{R0}(Y_1, Y_2|I = 0)$ is not defined, but we still can use formula (3). For the monozygotic twin pair, when '$I = 1$' the probability is zero and $f_{R1}(Y_1, Y_2)$ is defined as for an ordinary sib pair. For a unilineal relationship '$I = 2$' has probability zero, and then $f_{R2}(Y_1, Y_2)$ need not be defined.

Let $\alpha_{Ri} = P(I = i)$ be the prior probability that the pair R shares i alleles IBD and $r_R = 0.5\alpha_{R1} + \alpha_{R2}$ the coefficient of relationship, [Haseman and Elston, 1972]. It is shown in the appendix that

$$\lambda_R^{\text{main}} = 1 + \frac{r_R \sigma_a^{(1,2)} + \alpha_{R2} \sigma_d^{(1,2)}}{f_{R0}(Y_1, Y_2)}, \tag{4}$$

which generalizes expressions obtained by James [1971] and Risch [1990a] for binary traits recurrence and relative risks.

When $f_{R0}(Y_1, Y_2)$, $\sigma_a^{(1,2)}$ and $\sigma_d^{(1,2)}$ are independent of R, λ_R^{main} will depend on the degree of relationship R only through r_R and α_{2R}, as shown by Risch [1990a]. This happens when the penetrance matrix ψ is independent of R, as for one-locus models with no shared environmental or polygenic effects, but also for multiplicative multilocus models, since then the common R-dependent factor in $f_{R0}(Y_1, Y_2)$, $\sigma_a^{(1,2)}$ and $\sigma_d^{(1,2)}$ cancels out in (4). For instance, if there are no dominance effects (p small) and $R = n$ denotes a relationship of degree n, ($n = 0$: MZ twins, $n = 1$: parent-offspring, siblings, $n = 2$: half-sibs, uncle-nephew, grandparent-grandchild, $n = 3$: first cousins), then $\lambda_n - 1 = 2^{-n}(\lambda_0 - 1)$.

4 Linkage Analysis for Relative Pairs

4.1 IBD probabilities and effective number of meioses

Let $z_{Ri} = P(I = i|Y_1, Y_2)$ be the posterior probability that R shares i alleles IBD. Put $\lambda_{Ri}^{\text{main}} = f_{Ri}(Y_1, Y_2)/f_{R0}(Y_1, Y_2)$. Then, applying Bayes' rule, as in Risch [1990b], we get

$$z_{Ri} = \frac{\alpha_{Ri} f_{Ri}(Y_1, Y_2)}{f_R(Y_1, Y_2)} = \frac{\alpha_{Ri} \lambda_{Ri}^{\text{main}}}{\lambda_R^{\text{main}}}, \quad i = 0, 1, 2.$$

The amount of information contained in (Y_1, Y_2) for testing linkage between the trait locus and genetic markers is to a large extent determined by how much $\{z_{Ri}\}_{i=0}^2$ depart from $\{\alpha_{Ri}\}_{i=0}^2$. The effective number of meioses m^{test} for testing is introduced in Hössjer [2004] for general pedigrees, phenotypes and genetic models. It quantifies the equivalent number of fully observed meiotic events contained in (Y_1, Y_2). It is shown in the appendix that

$$m_R^{\text{test}} = \log_2 \left(\frac{z_{R0}^2}{\alpha_{R0}} + \frac{z_{R1}^2}{\alpha_{R1}} + \frac{z_{R2}^2}{\alpha_{R2}} \right), \tag{5}$$

with the convention $0/0 = 0$ whenever $\alpha_{Ri} = 0$. For a data set with several relative pairs (of the same or different kind), m^{test} is obtained by simply adding (5) over all pairs. For a unilineal relationship of degree n, $(\alpha_{1R} = 2^{-(n-1)}, \alpha_{2R} = 0)$, the maximal possible value $m_R^{\text{test}} = n-1$ is obtained when the genetic model and (Y_1, Y_2) is such that $z_{1R} = 1$. In other words, when the genetic component at the main locus is strong, distant relationships are more informative than close ones.

4.2 Power approximation

We will now motivate the relevance of m^{test} for linkage analysis. Suppose we wish to test

$$H_0 : \tau \notin \Omega$$
$$H_1 : \tau \in \Omega,$$

where Ω is the genomic region of interest, consisting of n_Ω chromosomes of total length L_Ω Morgans and τ is the unknown position of the disease gene. Consider a set of N relative pairs, $i = 1, \ldots, N$, whose relationships R_i and phenotypes $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})$ may vary. Let m_i be the number of meioses in the pedigree corresponding to R_i and $v_i(t)$ the inheritance vector of Pedigree i at locus t. This is a binary vector of length m_i, each bit of which corresponds to a meioses, with value 0 or 1 depending on whether a grandpaternal or grandmaternal allele was transmitted [Donelly 1983]. Let $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)$ be the collection of all phenotypes and $\boldsymbol{v}(t) = (v_1(t), \ldots, v_N(t))$ the collection of inheritance vectors at locus t. The latter is a binary vector of length $m_{\text{tot}} = m_1 + \ldots + m_N$. With complete marker information, a wide class of test statistics for testing H_0 against the pointwise alternative $\tau = t$, $t \in \Omega$, is

$$Z(t) = S(\boldsymbol{v}(t)), \tag{6}$$

where $S(\boldsymbol{w}) = S(\boldsymbol{w}; \boldsymbol{Y})$ is a score function defined for all binary vectors $\boldsymbol{w} = (w_1, \ldots, w_N)$ of length m_{tot}. We assume that S is standardized so that $\sum_{\boldsymbol{w}} S(\boldsymbol{w}) = 0$ and $2^{-m_{\text{tot}}} \sum_{\boldsymbol{w}} S^2(\boldsymbol{w}) = 1$, ensuring $E_{H_0}(Z(t)) = 0$ and $\text{Var}_{H_0} Z(t)) = 1$ under complete marker information. Notice that (6) contains as special case

$$Z(t) = \sum_{i=1}^{N} \gamma_i S_i(v_i(t)), \tag{7}$$

which is a linear combinations of family scores $S_i(v_i(t))$ with weights γ_i satisfying $\sum_{i=1}^N \gamma_i^2 = 1$, and with S_i a standardized score function of Pedigree i. The class (7) includes the affected pedigree method, see e.g. Weeks and Lange [1998], Fimmers et al. [1989], Whittemore and Halpern [1994] and Kruglyak et al. [1996]. As test statistics for testing H_0 against H_1 we use

$$Z_{\max} = \sup_{t \in \Omega} Z(t),$$

and H_0 is rejected as soon as $Z_{\max}$ exceeds a given threshold z.

Since $Z(t)$ can be seen as a stationary process we can use results from extreme value theory to approximate power and significance level, see appendix for details. The power approximation (C.1) is an increasing function of the noncentrality parameter η. As shown in Hössjer [2004], the maximal noncentrality parameter for the class (6) of test statistics is

$$\eta = \sqrt{2^{m_{\text{total}}^{\text{test}}} - 1} \tag{8}$$

where $m_{\text{total}}^{\text{test}} = m_{R_1}^{\text{test}} + \ldots + m_{R_N}^{\text{test}}$ is the total effective number of meioses for testing. The maximum (8) is attained for a score function $S(\boldsymbol{w}) = (P(\boldsymbol{w}) - \mu)/\sigma$, where $P(\boldsymbol{w}) = P(\boldsymbol{v}(\tau) = \boldsymbol{w}|\boldsymbol{Y}) = \prod_{i=1}^{N} P(v_i(\tau) = w_i|\boldsymbol{Y}_i)$ is the joint conditional distribution of all inheritance vectors at the trait locus, and μ and σ are standardization constants. In general, this optimal score function requires knowledge of the genetic model, hence the maximum noncentrality parameter (8) can be regarded as an ideal upper bound when the genetic model is unknown. Choosing threshold to achieve the desired level of significance and noncentrality parameter as in (8), we get, by insertion into (C.1) an approximation of the maximal possible power that can be obtained as function of the crossover rate ρ_Z and normalized mean slope d of Z, defined in the appendix.

Figure 1 shows the power β of detecting the significant linkage on a chromosome of length 2.985 Morgans. In calculation we used $\rho_Z = 2$ and $d = 1$, although they can vary a little depending on the genetic model and the pedigree structure [Lander and Kruglyak, 1995; Hössjer, 2003b]. We calculated thresholds required to reach three 'standard' significance levels of 0.05, 0.01 and 0.001. In all three cases we need a value of $m_{\text{total}}^{\text{test}}$ between four and five to achieve a power of 0.9. As we will show below the number of relative pairs required to reach a high power depends on the genetic model and the phenotypes. In Figure 2, when the threshold z is the one required for genomwide significance, the values of $m_{\text{total}}^{\text{test}}$ needed for reaching power 0.9 increase but not remarkably much.

5 Genetic models

5.1 Gaussian phenotypes

We studied relative risks for a class of genetic models where the genetic influence is a mixture of a major gene and a number of polygenes but the major gene and polygenes are unlinked. The goal of the analysis is to map the major gene and to increase statistical efficiency we take the polygenes

into account. If G_k is the major gene's genotype for individual k and Y_k the phenotype, we assume $Y_k|G_k \in N(\mu_{|G_k|}, \sigma^2)$. Here $|G_k|$ is the number of disease alleles of G_k and μ_0, μ_1 and μ_2 are the mean phenotype values for an individual with 0, 1, and 2 disease alleles, respectively. Residual variance σ^2 is the variance caused by polygenic and/or environmental effects. If large values of the phenotype indicates disease a natural constraint is $\mu_0 \leq \mu_1 \leq \mu_2$. For a relative pair we have $(Y_1, Y_2)|(G_1, G_2) \in N(\mu, \sigma^2\Sigma)$, where $\mu = (\mu_{|G_1|}, \mu_{|G_2|})$, Σ is a 2×2 correlation matrix with ones on the diagonal and correlation coefficient $\rho_Y = r_R h_a^2 + \alpha_{2R} h_d^2$ with h_a^2 and h_d^2 additive and dominant polygenic heritabilites, respectively (i.e. the fraction of σ^2 due to additive and dominance effects). Then we have that

$$f(Y_k|G_k) = \phi((Y_k - \mu_{|G_k|})/\sigma)/\sigma$$

and

$$f(Y_1, Y_2|G_1, G_2) = \phi_2((Y_1 - \mu_{|G_1|})/\sigma, ((Y_2 - \mu_{|G_2|})/\sigma; \rho_Y)/\sigma^2,$$

where ϕ and $\phi_2(.,.,\rho_Y)$ are the univariate and bivariate standard normal densities, the latter with correlation coefficient ρ_Y. See Lynch and Walsh [1998] and Hössjer [2005] for more details about the Gaussian phenotypes.

5.2 Liability Threshold Model

In the liability threshold model the observed phenotypes are binary ($Y_k = 1$ affected and $Y_k = 0$ unaffected) but there is no simple Mendelian inheritance pattern and the probability of expressing the disorder i.e. the distribution of Y_k is modelled as a function of an underlying quantitative variable X_k, the same kind of variable as in Section 5.1. The phenotype is then defined as $Y_k = 1_{\{X_k \geq T\}}$ where T is a given threshold. For more details about liability threshold models see Todorov and Suarez [2002].

We assume $X_k|G_k \in N(\mu_{|G_k|}, 1)$ and hence $X|G \in N(\mu, \Sigma)$. We then have penetrance parameters

$$\psi_i = P(Y_k = 1||G_k| = i) = 1 - \Phi(T - \mu_i),$$

for $i = 0, 1, 2$, where Φ is the distribution function of a standard normal variable. This yields

$$f(Y_k|G_k) = \psi_{|G_k|}^{Y_i} \cdot (1 - \psi_{|G_k|})^{1-Y_i}$$

and

$$f(Y_1, Y_2|G_1, G_2) = \int_A \phi_2(x_1, x_2; \rho_X) dx$$

where $A = (T - \mu_{|G_1|}, \infty) \times (T - \mu_{|G_2|}, \infty)$ if $Y_1 = Y_2 = 1$, $A = (T - \mu_{|G_1|}, \infty) \times (-\infty, T - \mu_{|G_2|})$ if $Y_1 = 1$ and $Y_2 = 0$, $A = (-\infty, T - \mu_{|G_1|}) \times (T - \mu_{|G_2|}, \infty)$ if $Y_1 = 0$ and $Y_2 = 1$, and $A = (-\infty, T - \mu_{|G_1|}) \times (-\infty, T - \mu_{|G_2|})$ if $Y_1 = Y_2 = 0$.

6 Results

For the Gaussian mixed model, we have standardized phenotypes so that $E(Y_k) = q^2\mu_0 + 2pq\mu_1 + p^2\mu_2 = 0$ and $\text{Var}(Y_k) = 1$. After standardization, there are four essential genetic parameters when $h_d^2 = 0$, namely p, h_a^2, the displacement $\text{Disp} = (\mu_2 - \mu_0)/\sigma$ that quantifies the strength and $\text{Dom} = (2\mu_1 - \mu_0 - \mu_2)/(\mu_1 - \mu_0)$, that quantifies the degree of dominance of the main locus genetic component. We studied the relation between the relative risk and different parameters but also investigated the informativity of the different relative pairs by looking at the value of m_R^{test}. A list of genetic models is provided in Table 1. We have that $\lambda_R = \lambda_R^{\text{main}}$ when $h_a^2 = 0$, see for example Figures 3 and 4, which is a result of no interaction between the main locus and the other loci. We define the pairs as being concordant if they have the same or similar phenotypes and discordant when they have opposite phenotypes. The informativity increases with h_a^2 for discordant pairs and decreases with h_a^2 for concordant pairs, see Figure 5. The most informative are discordant sib pairs with extreme phenotype values, large h_a^2, and large h^2 (main locus heritability). Relatives with phenotypes close or equal to zero are non-informative. These are the same kind of conclusions as in Risch and Zhang [1995, 1996]. Concordant pairs for high positive and low negative values are also informative. If $p < 0.5$ the concordant positive phenotypes are more informative then the concordant negative phenotypes, see Figures 5-9. For discordant phenotypes, the close relatives are most informative but for concordant positive phenotypes, the distant relatives are more informative, since α_{R_1} is small, see Figure 6. Further, the parent-offspring pair is not informative since $\alpha_{R_1} = z_{R_1} = 1$. Informativity increases also with p $(0 < p \leq 0.5)$. When p is small, often, the dominant model ($\text{Dom} = 1$) is most informative and the recessive one ($\text{Dom} = -1$) least informative. Informativity increases with Disp, even when h_a^2 increases.

For the liability threshold model we studied risk for different pairs of relatives and different values of h_a^2, p, and (ψ_0, ψ_1, ψ_2). Figures 10-14 show some examples from which we can see that both the relative risk and the effective number of meioses are almost independent of h_a^2. Only when the disease allele frequency p is low in relation to ψ_0 we can observe some dependence. The relative risk increases (slightly) with h_a^2 when $Y_1 = Y_2$ and decreases with h_a^2 when $Y_1 \neq Y_2$. In case when $Y_1 = Y_2 = 0$ the relative risk is zero or very close to zero. On the other hand, the relative risk at the main locus decreases with h_a^2 in case when $Y_1 \neq Y_2$ and $Y_1 = Y_2 = 1$. Again, when $Y_1 = Y_2 = 0$ it is very close to zero. From the figures with the effective number of meioses we can observe that m_R^{test} slightly increases when $Y_1 \neq Y_2$ (discordant pair) and decreases when $Y_1 = Y_2 = 1$ (concordant pair). The MZ twin pair is noninformative, since $\alpha_2 = z_{R_2} = 1$, as well as a relative pair with $Y_1 = Y_2 = 0$. Distant relationships are more informative than close ones when $Y_1 = Y_2 = 1$ and vice versa when $Y_1 \neq Y_2$. We also can observe that the values of ψ_1 and ψ_2 effect m_R^{test} but not as much as the value of ψ_0.

7 Discussion

We have shown numerically how relative risks and effective number of meioses (and hence also the number of relative pairs needed for detecting linkage) depend on phenotypes for two classes of genetic models; Gaussian and liability threshold models. For Gaussian phenotypes, extreme discordant sib pairs are most powerful. This is because these pairs are unlikely to share alleles IBD for any genetic model. Sib pairs concordant for extreme values can be also useful whereas sib pairs with intermediate values are only informative when the genetic component at disease locus is strong. The two major determinants of the power to detect linkage for a locus contributing to a quantitative trait are the heritability at that locus and the additive polygenic heritability. Additive polygenic heritability also increases statistical efficiency and compensate for power loss when the heritability at the main locus is low.

The liability threshold model seems intuitively reasonable for complex diseases, including one major gene and polygenes. However, we have shown that polygenic heritability affects the effective number of meioses and power very little when the penetrance parameters ψ_i are kept fixed. This indicates that the penetrance parameters themselves (without any polygenic components) provide a good description of a wide range of genetic models for binary traits. Another possibility is to consider relative risks and effective number of meioses for oligogenic multilocus models, as in Risch [1990a].

A Derivation of (3) and (4)

Conditioning on I we have

$$f_R(Y_1, Y_2) = \sum_{i=0}^{2} \alpha_{Ri} f_{Ri}(Y_1, Y_2).$$

Notice that (4) follows if we establish

$$\begin{aligned} f_{R1}(Y_1, Y_2) &= f_{R0}(Y_1, Y_2) + 0.5\sigma_a^{(1,2)}, \\ f_{R2}(Y_1, Y_2) &= f_{R0}(Y_1, Y_2) + \sigma_a^{(1,2)} + \sigma_d^{(1,2)}. \end{aligned} \tag{A.1}$$

Assume there are f founders in the pedigree to which (1,2) belongs and let $(a_1, \ldots, a_{2f})$ be a binary vector of length $2f$ containing the founder alleles. Under random mating $\{a_i\}$ are independent random variables with $P(a_i = 1) = p$ and $P(a_i = 0) = q$. Write $G_1 = (a_j a_k)$ and $G_2 = (a_l a_n)$, where $j, k, l, n \in \{1, \ldots, 2f\}$ and $I = 1_{\{j=l\}} + 1_{\{j=n\}} + 1_{\{k=l\}} + 1_{\{k=n\}}$. Let $a = (a_j, a_k, a_l, a_n)$ and write

$$f_R(Y_1, Y_2 | G_1, G_2) = h(a), \tag{A.2}$$

regarding (Y_1, Y_2) as fixed and (G_1, G_2) as varying. Assuming no imprinting, the order of the alleles within each genotype is immaterial, and hence h is determined by the 3×3 joint penetrance matrix

$$\begin{pmatrix} h(0,0,0,0) & h(0,0,0,1) & h(0,0,1,1) \\ h(0,1,0,0) & h(0,1,0,1) & h(0,1,1,1) \\ h(1,1,0,0) & h(1,1,0,1) & h(1,1,1,1) \end{pmatrix} = \begin{pmatrix} \psi_{00} & \psi_{01} & \psi_{02} \\ \psi_{10} & \psi_{11} & \psi_{12} \\ \psi_{20} & \psi_{21} & \psi_{22} \end{pmatrix} = \psi.$$

Let $U = \mathbb{R}^3$ and $V = U \times U$ be the spaces of 1×3 vectors and 3×3 matrices respectively. Introduce the scalar product $< u, w >= q^2 \cdot u_0 w_0 + 2pq \cdot u_1 w_1 + p^2 \cdot u_2 w_2$ on U and

$$\begin{aligned} (\psi, \theta) &= q^2 \cdot q^2 \cdot \psi_{00}\theta_{00} + q^2 \cdot 2pq \cdot \psi_{01}\theta_{01} + q^2 \cdot p^2 \cdot \psi_{02}\theta_{02} \\ &+ 2pq \cdot q^2 \cdot \psi_{10}\theta_{10} + 2pq \cdot 2pq \cdot \psi_{11}\theta_{11} + 2pq \cdot p^2 \cdot \psi_{12}\theta_{12} \\ &+ p^2 \cdot q^2 \cdot \psi_{20}\theta_{20} + p^2 \cdot 2pq \cdot \psi_{21}\theta_{21} + p^2 \cdot p^2 \cdot \psi_{22}\theta_{22} \end{aligned}$$

on V respectively. An orthonormal basis on U is

$$\begin{aligned} e_1 &= (1,1,1), \\ e_2 &= \tfrac{1}{\sqrt{2pq}}(-2p, q-p, 2q), \\ e_3 &= (1/q - 1, -1, 1/p - 1), \end{aligned}$$

see Hössjer [2003a]. Similarly, an orthonormal basis on V consists of the nine matrices $\{e_{ij} = e_i' e_j;\ i, j = 1, 2, 3\}$. Let $\xi_i = (a_i - p)/\sqrt{pq}$, so that $\{\xi_i\}$ are i.i.d. random variables with zero mean and unit variance. The RHS of (A.2) can be expanded into a sum of uncorrelated terms

$$\begin{aligned} h(a) &= (\psi, e_{11}) + \tfrac{1}{\sqrt{2}}(\psi, e_{12})(\xi_l + \xi_n) + (\psi, e_{13})\xi_l\xi_n \\ &+ \tfrac{1}{\sqrt{2}}(\psi, e_{21})(\xi_j + \xi_k) + \tfrac{1}{2}(\psi, e_{22})(\xi_j + \xi_k)(\xi_l + \xi_n) + \tfrac{1}{\sqrt{2}}(\psi, e_{23})(\xi_j + \xi_k)\xi_l\xi_n \\ &+ (\psi, e_{31})\xi_j\xi_k + \tfrac{1}{\sqrt{2}}(\psi, e_{32})\xi_j\xi_k(\xi_l + \xi_n) + (\psi, e_{33})\xi_j\xi_k\xi_l\xi_n \end{aligned}$$

generalizing the corresponding expansion for U in Hössjer [2003a, Lemma 1]. See also the suplementary material of Hössjer [2005] for expansions involving more than two individuals.

When $I = 0$, the indices j, k, l and n are all different. Using the zero mean, unit variance and independence of $\{\xi_i\}$,

$$\begin{aligned} f_{R0}(Y_1, Y_2) &= E(h(a)|I = 0) = E(h(a_j, a_k, a_l, a_n)) \\ &= (\psi, e_{11}) = u\psi u', \end{aligned}$$

proving (3). If $I = 1$, we may without loss of generality assume $j = l$ and $j \neq k \neq n \neq j$. Hence

$$\begin{aligned} f_{R1}(Y_1, Y_2) &= E(h(a)|I = 1) = E(h(a_j, a_k, a_j, a_n)) \\ &= (\psi, e_{11}) + 0.5(\psi, e_{22}) \\ &= f_{R0}(Y_1, Y_2) + 0.5u_a\psi u_a' \\ &= f_{R0}(Y_1, Y_2) + 0.5\sigma_a^{(1,2)}. \end{aligned}$$

Similarly, if $I = 2$ we assume $j = l$, $k = n$ and $j \neq k$ and obtain

$$\begin{aligned} f_{R2}(Y_1, Y_2) &= E(h(a)|I = 2) = E(h(a_j, a_k, a_j, a_k)) \\ &= (\psi, e_{11}) + (\psi, e_{22}) + (\psi, e_{33}) \\ &= f_{R0}(Y_1, Y_2) + u_a \psi u_a' + u_d \psi u_d' \\ &= f_{R0}(Y_1, Y_2) + \sigma_a^{(1,2)} + \sigma_d^{(1,2)}. \end{aligned}$$

The last two displayed equations prove (A.1). Finally notice that $\sigma_g^{(1,2)} = f_{R2}(Y_1, Y_2) - f_{R0}(Y_1, Y_2) = \sigma_a^{(1,2)} + \sigma_d^{(1,2)}$. □

B Derivation of (5)

Let m be the number of meioses of the pedigree to which 1 and 2 belong and $v = (v_1, \ldots, v_m)$ the corresponding binary inheritance vector at the disease trait locus. Define $P(w) = P(v = w|Y_1, Y_2)$ for all 2^m binary vectors w of length m. This gives the posterior distribution given phenotypes of the inheritance vector at the trait locus. The general expression for the effective number of meioses for testing in Hössjer [2004] is

$$m^{\text{test}} = \log_2 \left(2^m \sum_w P^2(w) \right) \tag{B.1}$$

for one pedigree. We will show that this expression coincides with (5) for a relative pair R. Let n_i be the number of inheritance vectors that give $I = i$ alleles IBD for 1 and 2, so that $\alpha_{Ri} = n_i/2^m$. Further, let C_i be the set inheritance vectors corresponding to $I = i$ (hence $|C_i| = n_i$). Then $P(w) = 2^{-m} z_{Ri}/\alpha_{Ri}$ when $w \in C_i$, so that

$$\begin{aligned} \textstyle\sum_w P^2(w) &= 2^{-2m} \left(n_0 (z_{R0}/\alpha_{R0})^2 + n_1 (z_{R1}/\alpha_{R1})^2 + n_2 (z_{R2}/\alpha_{R2})^2 \right) \\ &= 2^{-m} \left(z_{R0}^2/\alpha_{R0} + z_{R1}^2/\alpha_{R1} + z_{R2}^2/\alpha_{R2} \right), \end{aligned}$$

Insertion of this expression into (B.1) gives (5). □

C Power approximation formula

According to Feingold et al. [1993] and Hössjer [2003b], the power $\beta = P_{H_1}(Z_{\max} \geq z)$ can be approximated as a function of the noncentrality parameter $\eta = E(Z(\tau))$ by

$$\beta \approx 1 - \Phi(z - \eta) + \phi(z - \eta)\left(\frac{2}{\eta d} - \frac{1}{\eta(2d - 1) + z}\right), \tag{C.1}$$

where $d = (-E'(Z(x)|_{x=\tau})/(2\rho\eta)$ is a normalized mean slope at the disease locus and ρ_Z is the crossover rate that measures the amount of fluctuations of the process Z. The threshold z is calculated so that the significance level $\alpha = P_{H_0}(Z_{\max} \geq z)$ attains a given value. To this end, we use the approximation

$$\alpha \approx 1 - \exp(-(1 - \Phi(z))(n_\Omega + 2\rho_Z L_\Omega z^2)). \tag{C.2}$$

defined by Lander and Kruglyak [1995].

Acknowledgement

The first author was sponsored by the National Research School in Genomics and Bioinformatics and the second by the Swedish Research Council, contract nr 626-2002-6286.

References

Amos CI. 1994. Robust variance-components approach for assesing genetic linkage in pedigrees. Am J Hum Genet 54:535-543.

Cardon LR, Fulker DW. 1994. The power of interval mapping of quantitative trait loci using selected sib pairs. Am J Hum Genet 55:825-833.

Carrey G, Williamson JA. 1991. Linkage analysis of quantitative traits-increased power by using selected samples. Am J Hum Genet 49:786-796.

Donnely P. 1983. The probability that related individuals share some section of the genome identical by descent.Theoretical Population Biology 3:34-64.

Dudoit S, Speed TP. 2000. A score test for the linkage analysis of qualitative and quantitative traits based on identity by descent data from sib-pairs. Biostatistics 1:1-26.

Elston R, Olson J, Palmer L. 2002. Biostatistical genetics and genetic epidemiology. Wiley reference in series in biostatistics.

Fimmers R, Seuhter SA, Neugebauer M, Knapp M, Baur MP. 1989. Identity-by-descent analysis using complete high-resolutions. In Multipoint mapping and linkage based on affected pedigree members, eds. Elston, R.C. et al. Genetic Analysis Workshop 6, Liss, New York, 123-128.

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

Haseman JK, Elston RC. 1972. The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:2-19.

Hössjer O. 2003a. Determining inheritance distributions via stochastic penetrances. J Amer Statist Assoc 98:1035-1051.

Hössjer O. 2003b. Spectral decomposition of score functions in linkage analysis. Report 2003:21, Mathematical Statistics, Stockholm University.

Hössjer O. 2004. Information bounds in linkage analysis. J Math Biol 50(2):208-232.

Hössjer O. 2005. Conditional likelihood score functions in linkage analysis. To appear in Biostatistics.

James JW. 1971. Frequency in relatives for an all-or-none trait. Ann Hum Genet 35:47-48.

Kraft P, Thomas DC. 2000. Bias and efficiency in family-based gene-characterization studies: conditional, prospective, retrospective, and joint likelihoods. Am J Hum Genet 66:1119-1131.

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. 1996. Parametric and nonparametric linkage analysis: A unified multipoint approach. Am J Hum Genet 42:315-326.

Lander EL, Kruglyak L. 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nature Genetics 11:241-247.

Risch N. 1990a. Linkage strategies for genetically complex traits I. Multilocus models. Am J Hum Genet 46:222-228.

Risch N. 1990b. Linkage strategies for genetically complex traits II. The power of affected relative pairs. Am J Hum Genet 46:229-241.

Risch N, Zhang H. 1995. Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science 268:1584-1589.

Risch N, Zhang H. 1996. Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations. Am J Hum Genet 58:836-843.

Todorov AA, Suarez B K. 2002. Liability model. In biostatistical Genetics and Genetic Epidemiology, Elston R, Olson J, Palmer L. (eds.), Wiley, 430-435.

Weeks D, Lange L. 1988. The affected-pedigree-member method of linkage analysis. Am J Hum Genet 42:315-326.

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

Table 1: Overview of the parameters in the study with the Gaussian phenotypes. Note that $E(Y_k) = 0$, $\text{Var}(Y_k) = 1$, and h^2 is the main locus heritability, i.e. $\text{Var}(\mu_{|G_k|})/\text{Var}(Y_k)$.

Model	Disp	Dom	h^2	h_a^2	p	μ_0	μ_1	μ_2	σ
1	2	0	0.1525	0	0.1	-0.1841	0.7365	1.6570	0.9206
2	2	0	0.1525	0.1	0.1	0.1841	0.7365	1.6570	0.9206
3	2	0	0.1525	0.5	0.1	0.1841	0.7365	1.6570	0.9206
4	2	0	0.1525	0.8	0.1	0.1841	0.7365	1.6570	0.9206
5	2	0	0.0020	0.5	0.001	-0.0020	0.9970	1.9960	0.9990
6	2	0	0.1525	0.5	0.1	-0.1841	0.7365	1.6570	0.9206
7	2	0	0.3334	0.5	0.5	-0.81656	0	0.8165	0.8165
8	2	1	0.3810	0.5	0.1	-0.2990	1.2745	1.2745	0.7867
9	2	-1	0.0381	0.5	0.1	-0.0196	-0.0196	1.9419	0.9808
10	2	0.5	0.2652	0.5	0.1	-0.2486	1.0372	1.4658	0.8572
11	2	-0.5	0.0679	0.5	0.1	-0.1062	0.3765	1.8247	0.9654
12	1	0	0.0432	0.5	0.1	-0.0978	0.3913	0.8804	0.9782
13	3	0	0.2883	0.5	0.1	-0.2531	1.0124	2.2779	0.8436
14	4	0	0.4186	0.5	0.1	-0.3050	1.2200	2.7450	0.7625

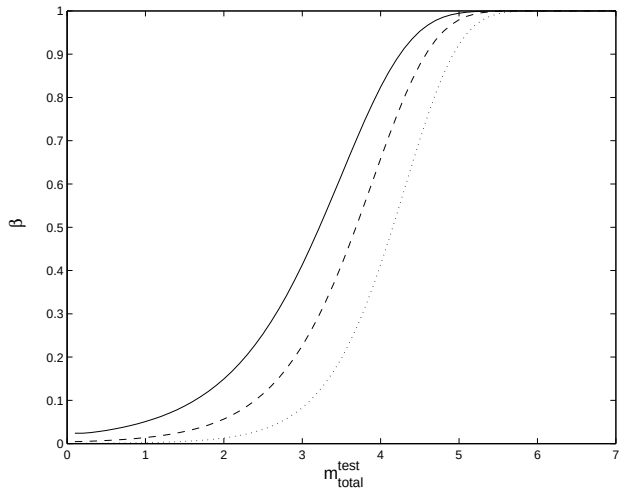

Figure 1: Approximation of the power to detect linkage β as function of the effective number of meioses $m^{\text{test}}_{\text{total}}$ for one chromosome of length $L_\Omega = 2.985$ Morgans. The threshold z and significance level α are chosen as $z = 3.375$ and $\alpha = 0.05$ (solid), $z = 3.863$ and $\alpha = 0.01$ (- -), $z = 4.455$ and $\alpha = 0.001$ $(\cdots)$.

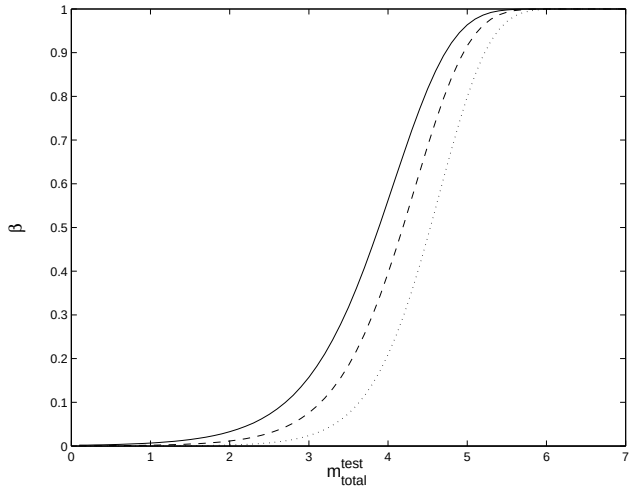

Figure 2: Approximation of the power to detect linkage β as function of the effective number of meioses $m^{\text{test}}_{\text{total}}$. The threshold z is calculated for a genomwide significance level α, i.e. $n_\Omega = 22$ and $L_\Omega = 33.5$ Morgans. In the figure $z = 4.1$ and $\alpha = 0.05$ (solid), $z = 4.5$ and $\alpha = 0.01$ (- -), $z = 5.02$ and $\alpha = 0.001$ $(\cdots)$.

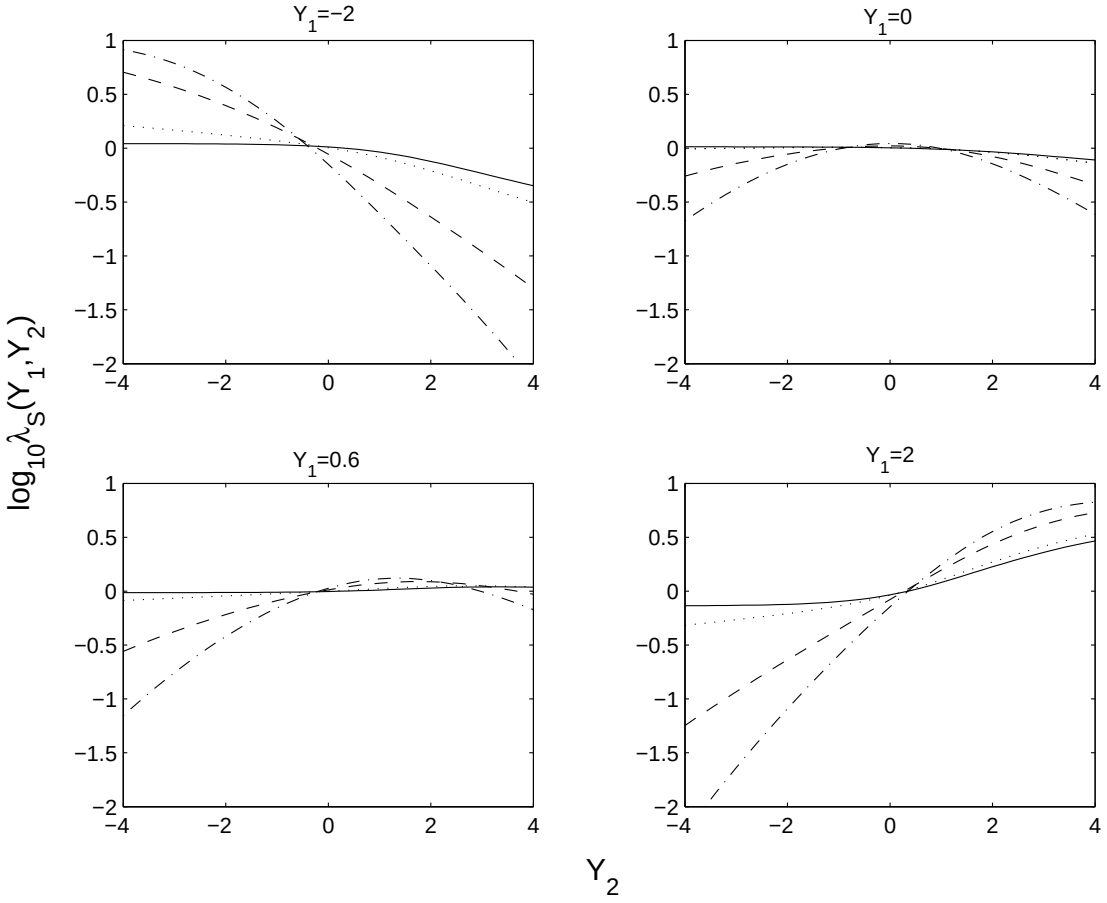

Figure 3: Relative risk λ_S as function of the values of the trait Y_1 and Y_2. The reference model is Gaussian with no dominant polygenic effects ($h_d^2 = 0$). Disease allele frequency $p = 0.1$, displacement Disp $= 2$ and dominance of the main locus genetic component Dom $= 0$. Relative risk is calculated for four different values of additive polygenic heritability; $h_a^2 = 0$ (solid), $h_a^2 = 0.1$ $(\cdots)$, $h_a^2 = 0.5$ (- -), and $h_a^2 = 0.8$, (-·-).

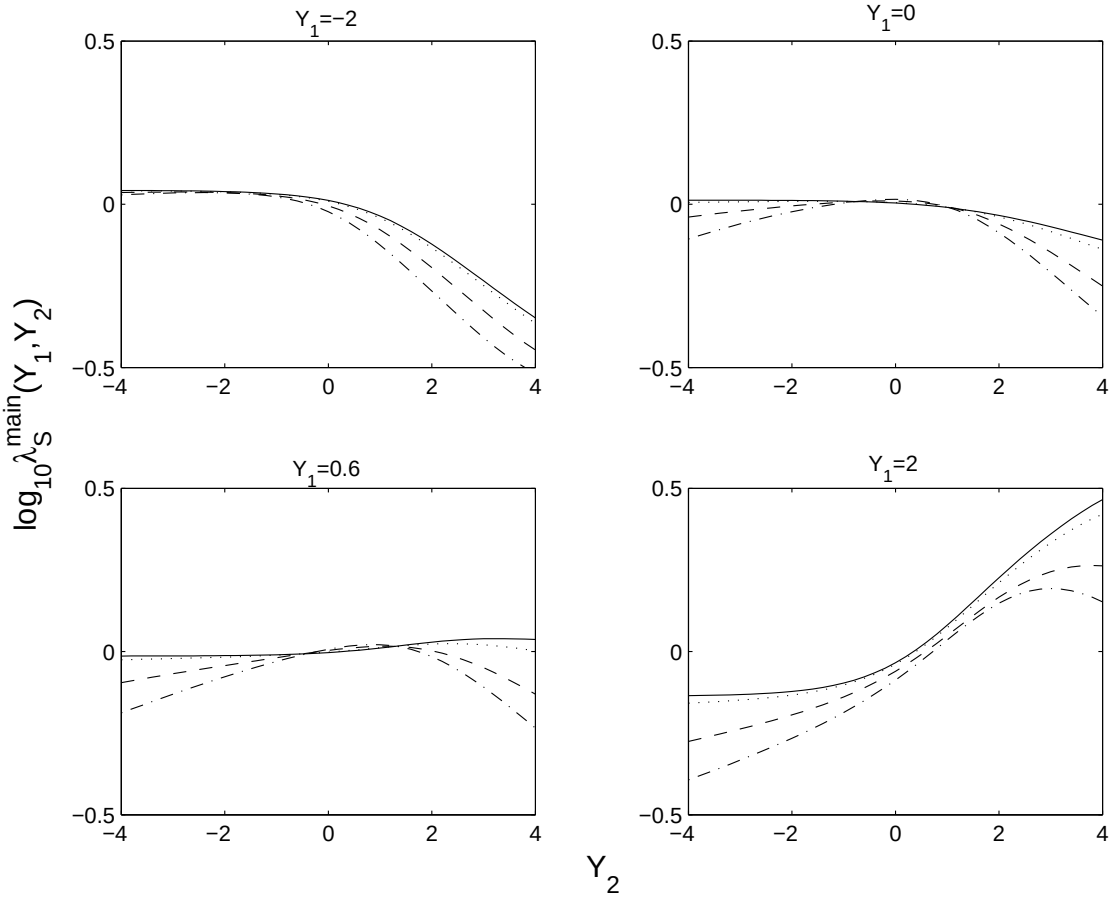

Figure 4: Relative risk due to trait locus λ_S^{main} as function of the values of the trait Y_1 and Y_2. The reference model is Gaussian with no dominant polygenic effects ($h_d^2 = 0$). Disease allele frequency $p = 0.1$, displacement Disp $= 2$ and dominance of the main locus genetic component Dom $= 0$. Relative risk is calculated for four different values of additive polygenic heritability; $h_a^2 = 0$ (solid), $h_a^2 = 0.1$ $(\cdots)$, $h_a^2 = 0.5$ (- -), and $h_a^2 = 0.8$, (-·-).

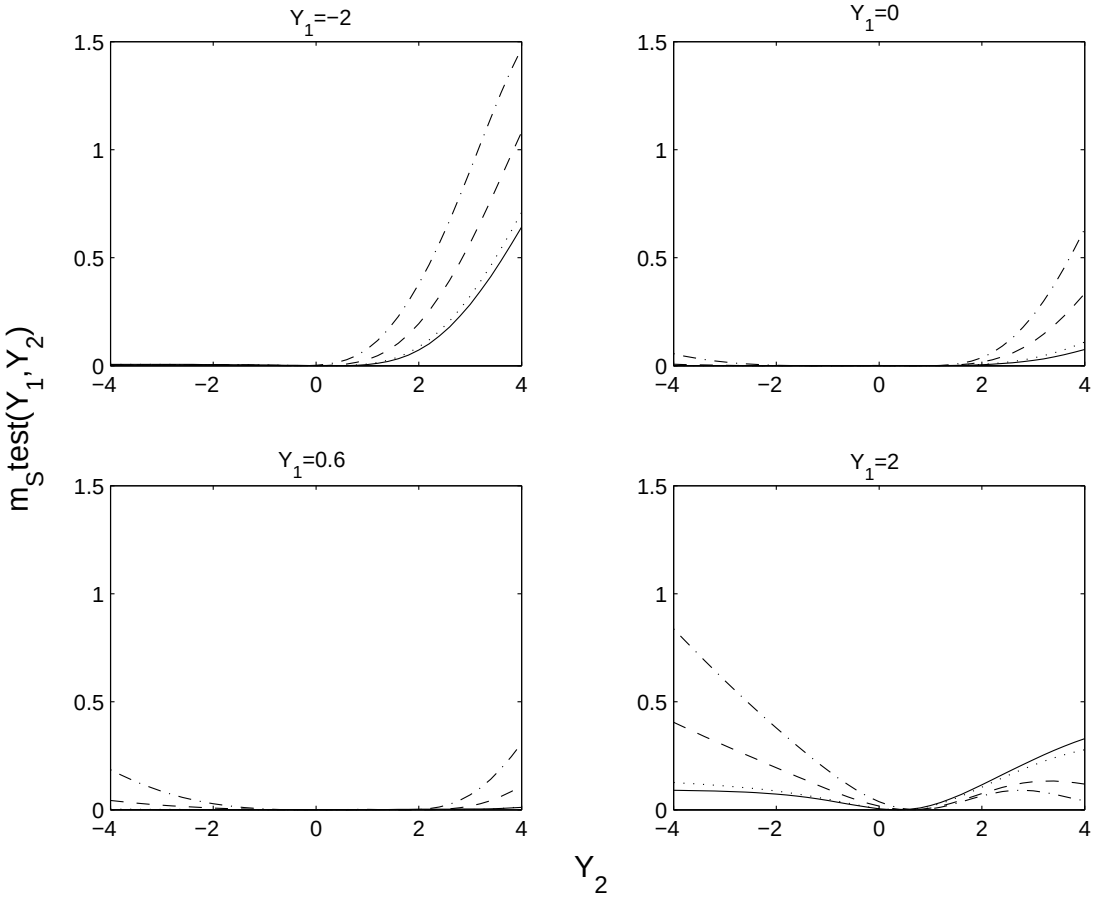

Figure 5: Effective number of meioses m_S^{test} as function of the values of the trait Y_1 and Y_2. The reference model is Gaussian with no dominant polygenic effects ($h_d^2 = 0$). Disease allele frequency $p = 0.1$, displacement Disp $= 2$ and dominance of the main locus genetic component Dom $= 0$. Relative risk is calculated for four different values of additive polygenic heritability; $h_a^2 = 0$ (solid), $h_a^2 = 0.1$ $(\cdots)$, $h_a^2 = 0.5$ (- -), and $h_a^2 = 0.8$, (-·-).

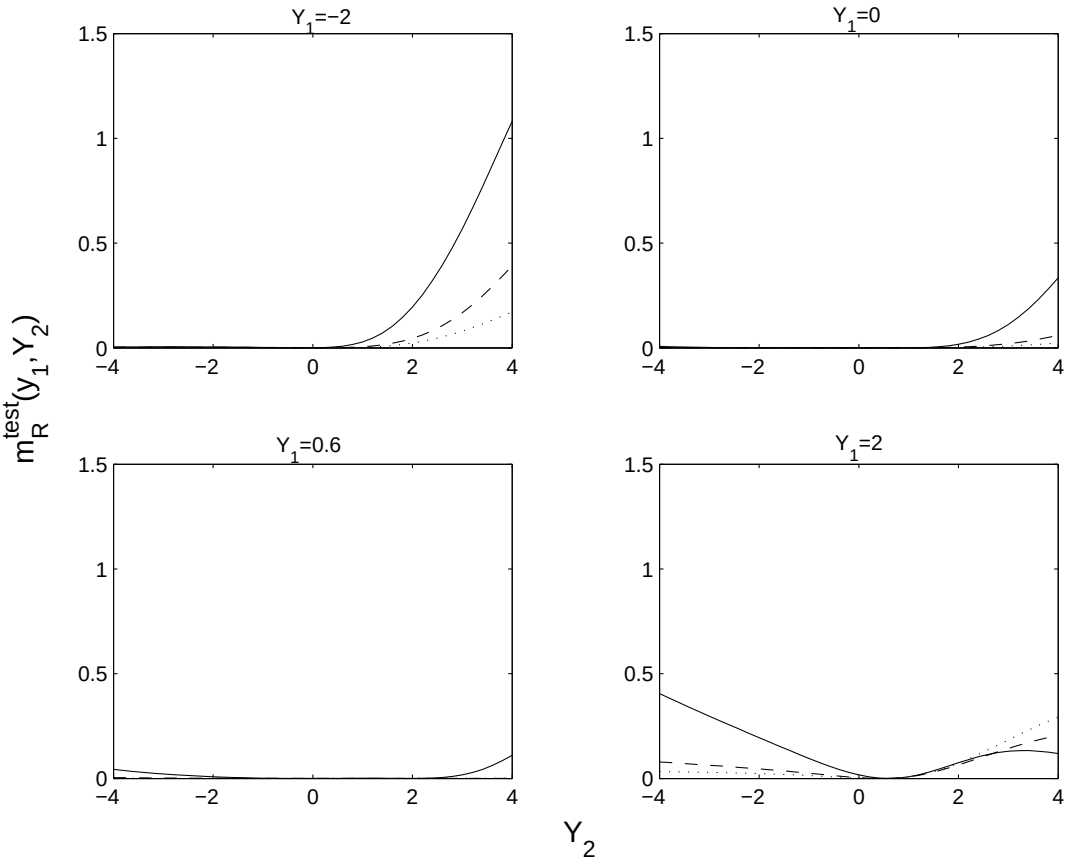

Figure 6: Effective number of meioses m_R^{test} as function of the values of the trait Y_1 and Y_2. The reference model is Gaussian with no dominant polygenic effects ($h_d^2 = 0$). Disease allele frequency $p = 0.1$, additive polygenic heritability $h_a^2 = 0.5$, displacement Disp $= 2$ and dominance of the main locus genetic component Dom $= 0$. Relative risk is calculated for three different relative pairs; sib pair (solid), grandparent-offspring (- -), and first cousins ($\cdots$). Note: For a parent-offspring pair $m^{\text{test}} = 0$ since $\alpha_{R_1} = z_{R_1}$.

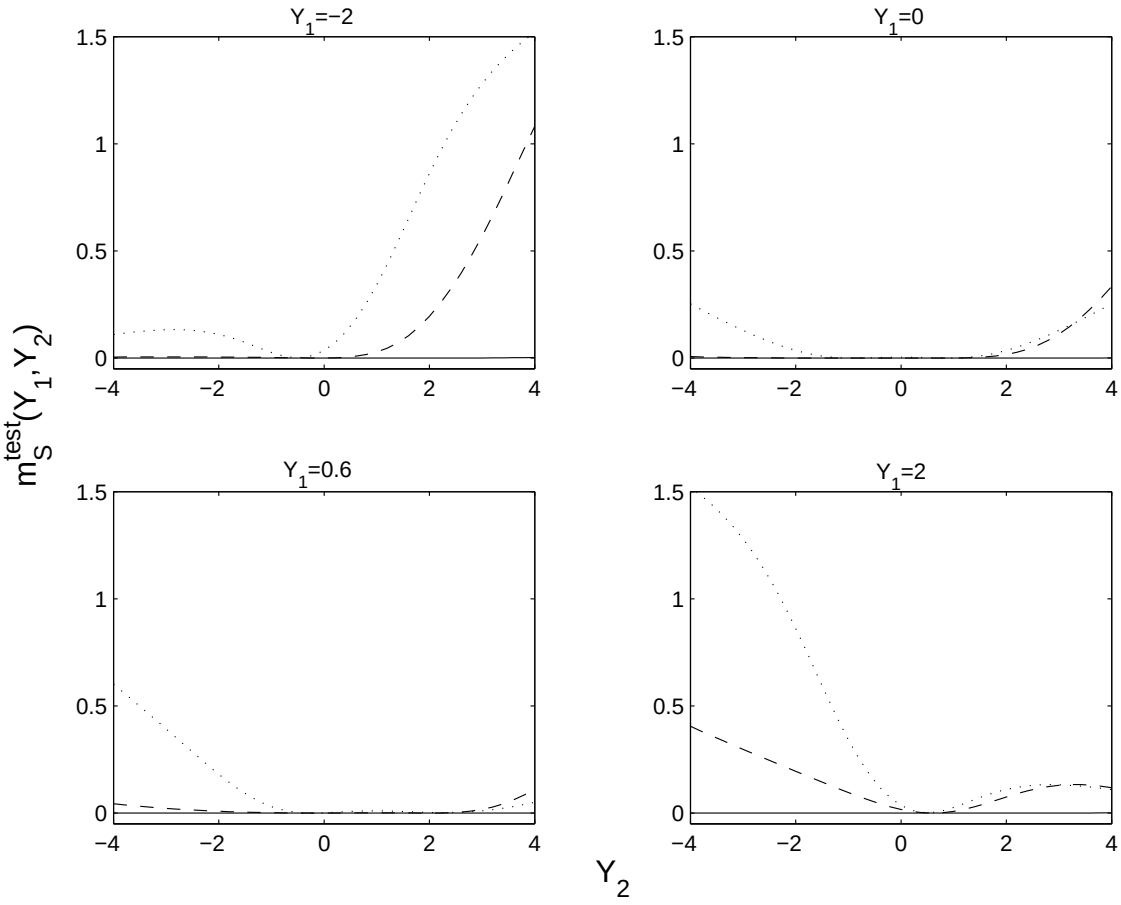

Figure 7: Effective number of meioses m_S^{test} as function of the values of the trait Y_1 and Y_2. The reference model is Gaussian with no dominant polygenic effects ($h_d^2 = 0$). Additive polygenic heritability $h_a^2 = 0.5$, displacement Disp $= 2$ and dominance of the main locus genetic component Dom $= 0$. Relative risk is calculated for three different values of disease allele frequencies, $p = 0.001$ (solid), $p = 0.1$ (- -), and $p = 0.5$ ($\cdots$). When $p = 0.5$ curve for $Y_1 = 2$ is just a mirror image of curve when $Y_1 = -2$ because of symmetry.

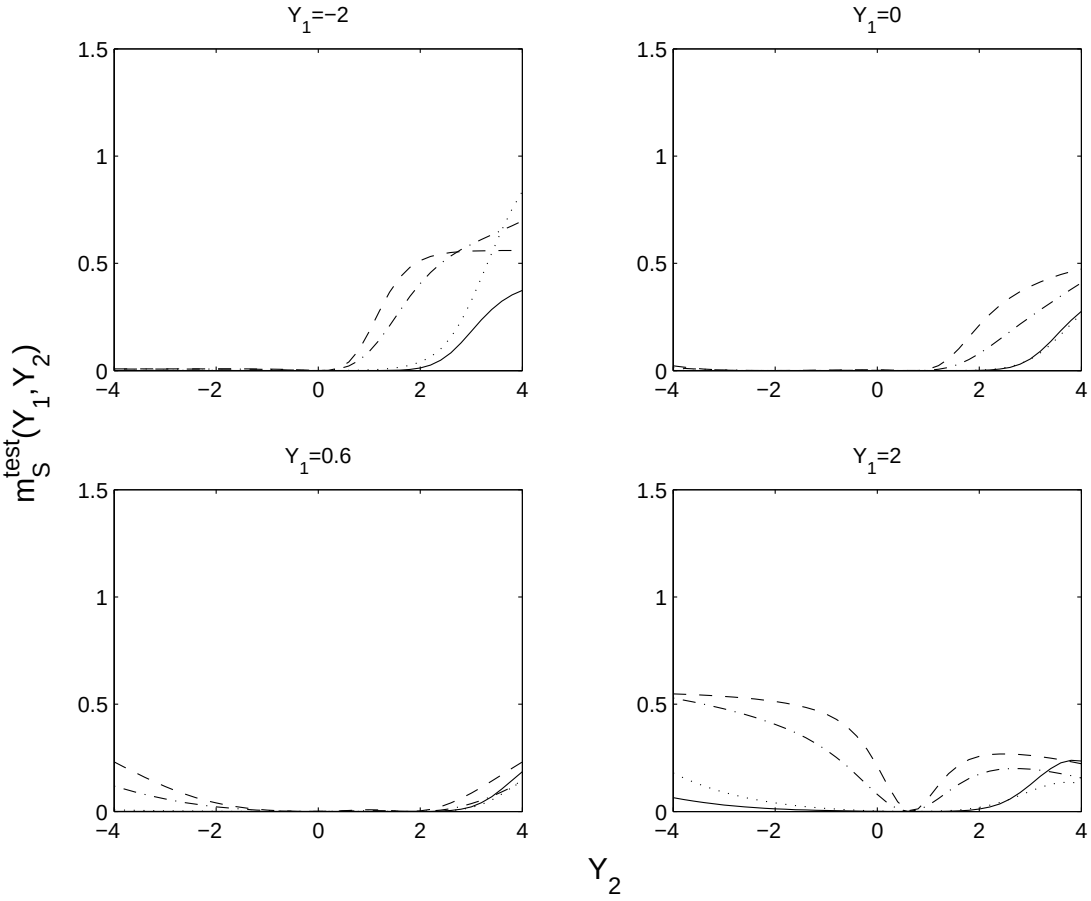

Figure 8: Effective number of meioses m_S^{test} as function of the values of the trait Y_1 and Y_2. The reference model is Gaussian with no dominant polygenic effects ($h_d^2 = 0$). Disease allele frequency $p = 0.1$, additive polygenic heritability $h_a^2 = 0.5$, and displacement Disp $= 2$. Relative risk is calculated for four different values of dominance of the main locus genetic component; Dom $= -1$ (solid), Dom $= 1$ (- -), Dom $= 0.5$ (-·-), and Dom $= -0.5$ ($\cdots$).

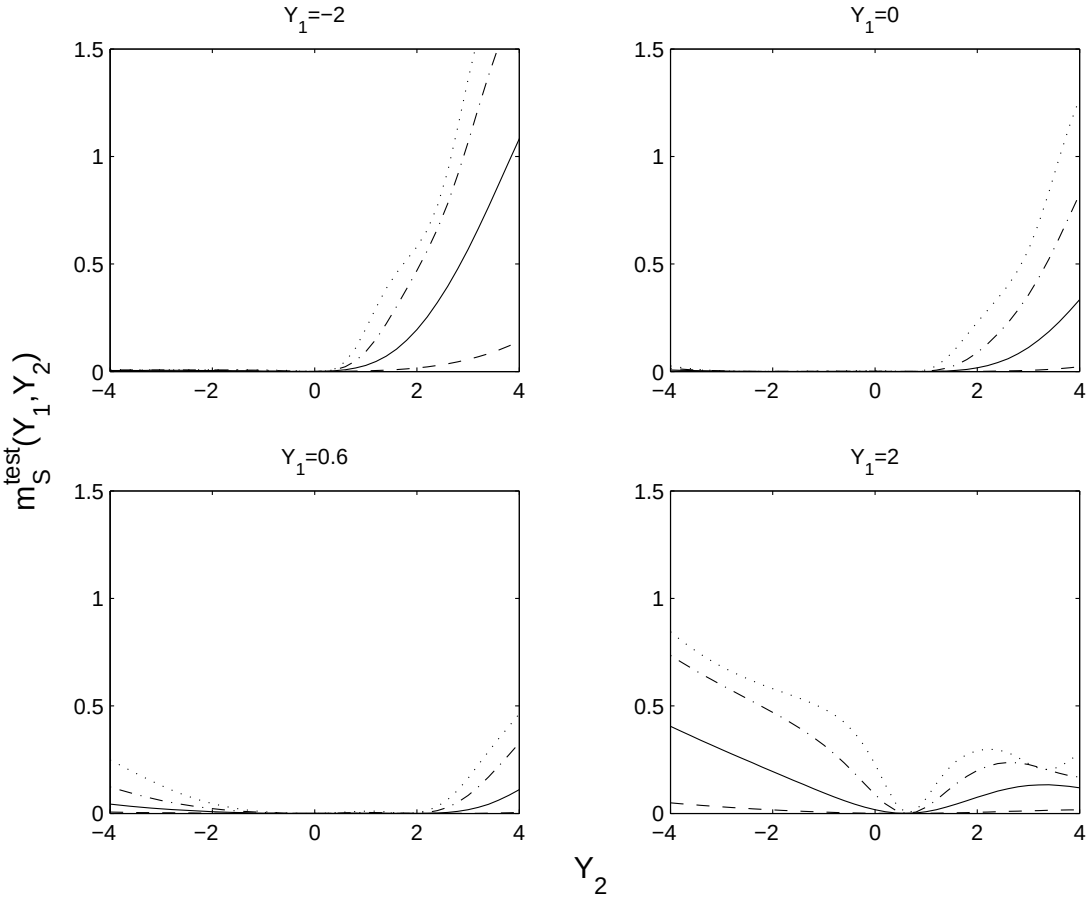

Figure 9: Effective number of meioses m_S^{test} as function of the values of the trait Y_1 and Y_2. The reference model is Gaussian with no dominant polygenic effects ($h_d^2 = 0$). Disease allele frequency $p = 0.1$, additive polygenic heritability $h_a^2 = 0.5$, and dominance of the main locus genetic component Dom $= 0$. Relative risk is calculated for four different values of displacement; Disp $= 1$ (- -), Disp $= 2$ (solid), Disp $= 3$ (-·-), and Disp $= 4$ ($\cdots$).

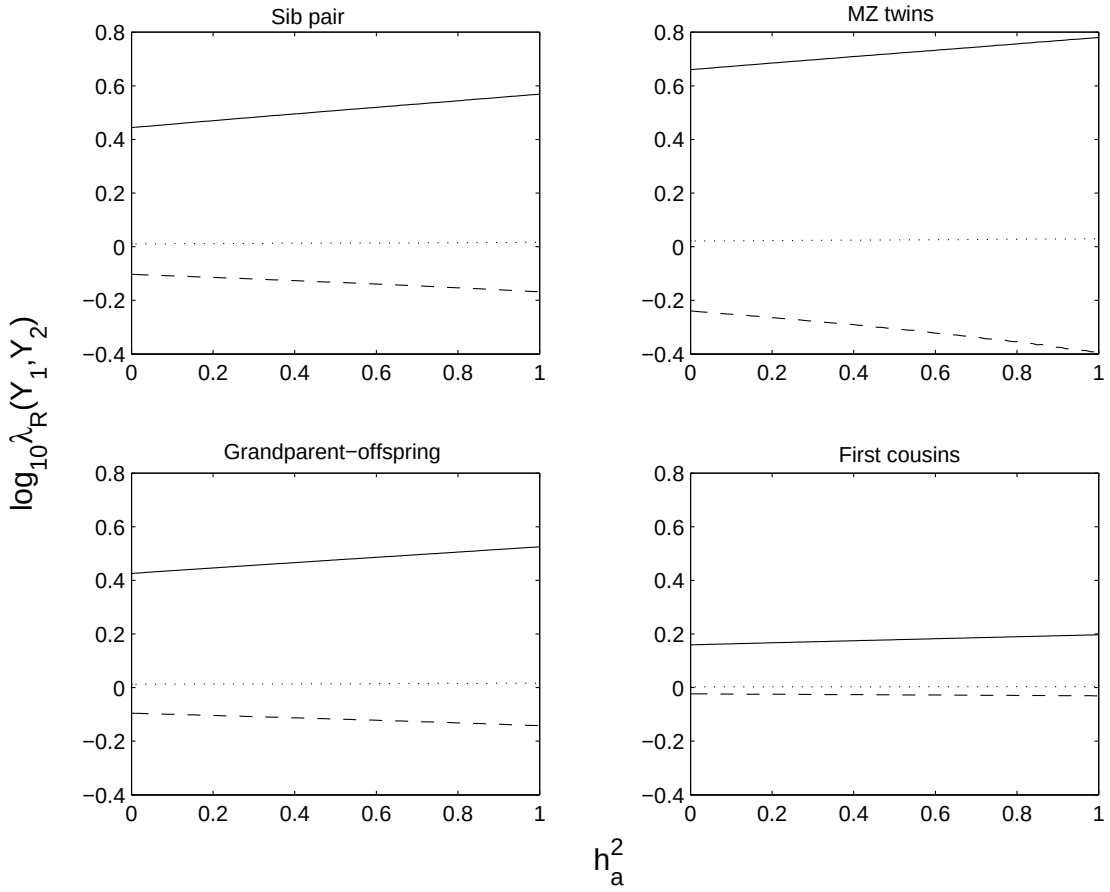

Figure 10: Relative risk λ_R for different relative pairs as function of the additive polygenic heritability h_a^2 and values of the trait Y_1 and Y_2. It is calculated for three different combinations of Y_1 and Y_2; $Y_1 = Y_2 = 1$ (solid), $Y_1 = 0$ and $Y_2 = 1$ (- -), and $Y_1 = Y_2 = 0$ $(\cdots)$. The reference model is the liability threshold model with $(\psi_0, \psi_1, \psi_2) = (0.01, 0.5, 0.8)$.

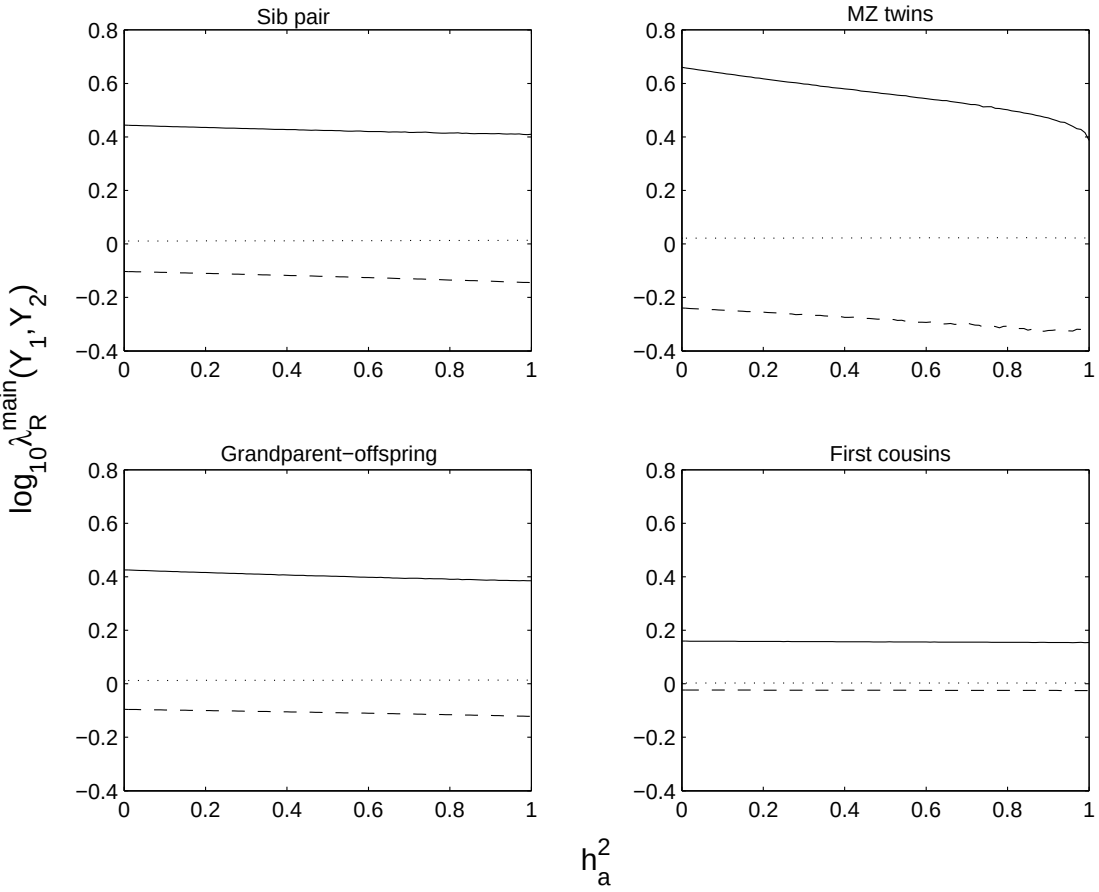

Figure 11: Relative risk at the main locus λ_R^{main} for different relative pairs as function of the additive polygenic heritability h_a^2 and values of the trait Y_1 and Y_2. It is calculated for three different combinations of Y_1 and Y_2; $Y_1 = Y_2 = 1$ (solid), $Y_1 = 0$ and $Y_2 = 1$ (- -), and $Y_1 = Y_2 = 0$ ($\cdots$). The reference model is the liability threshold model with $(\psi_0, \psi_1, \psi_2) = (0.01, 0.5, 0.8)$.

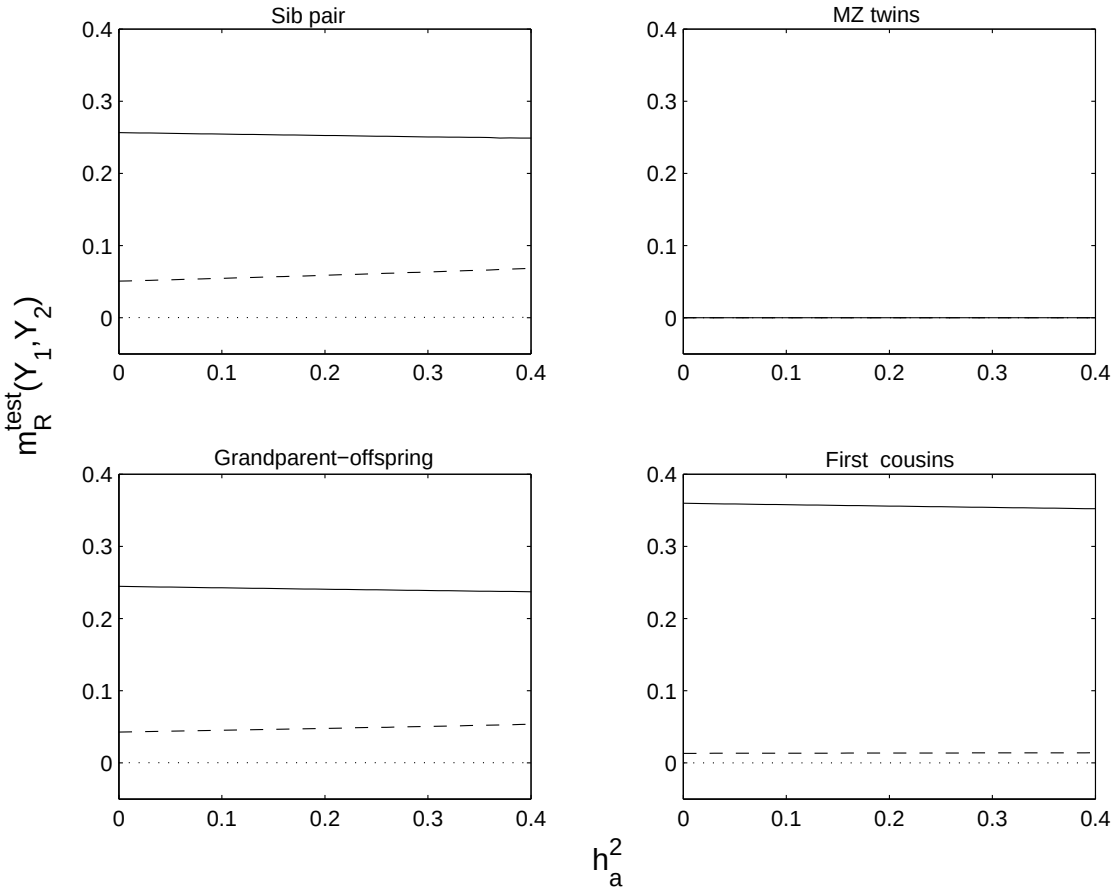

Figure 12: Effective number of meioses m_R^{test} for different relative pairs as function of the values of the additive polygenic heritability h_a^2 and values of the trait Y_1 and Y_2. It is calculated for three different combinations of Y_1 and Y_2; $Y_1 = Y_2 = 1$ (solid), $Y_1 = 0$ and $Y_2 = 1$ (- -), and $Y_1 = Y_2 = 0$ $(\cdots)$. The reference model is the liability threshold model with $(\psi_0, \psi_1, \psi_2) = (0.01, 0.5, 0.8)$.

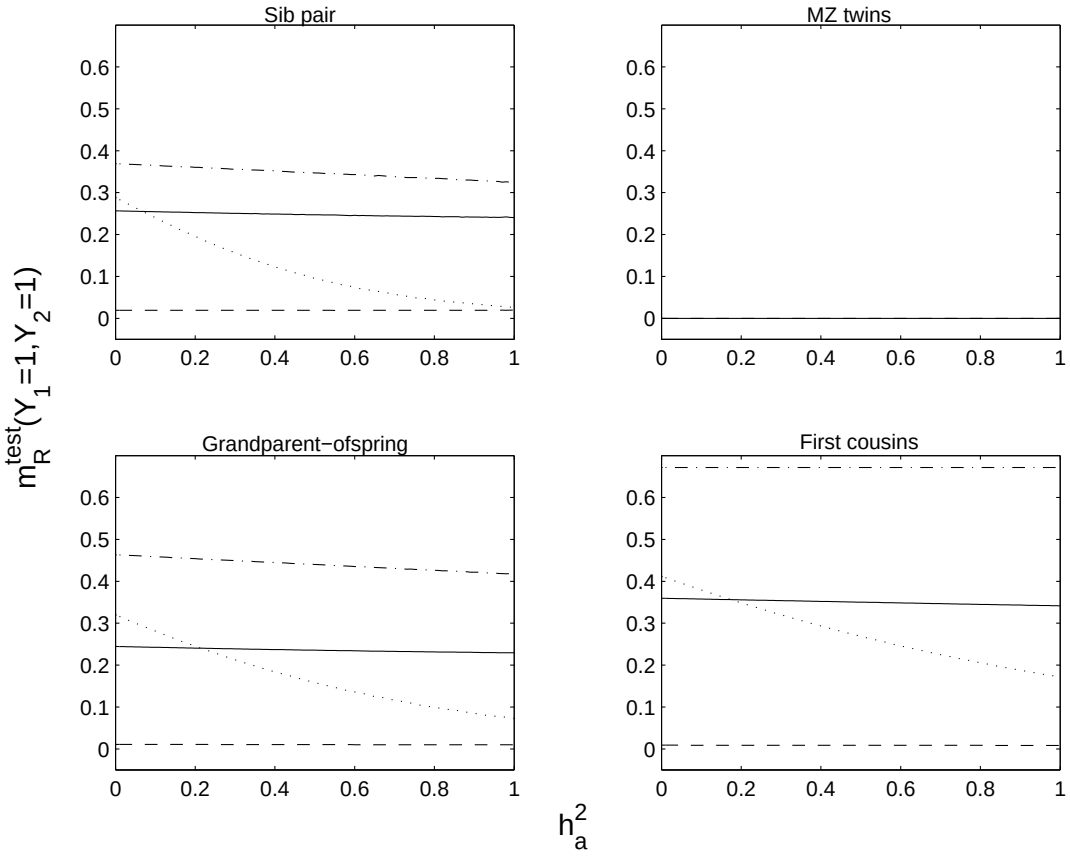

Figure 13: Effective number of meioses m_R^{test} for different relative pairs as function of the values of the additive polygenic heritability h_a^2 when $Y_1 = Y_2 = 1$. It is calculated for four different values of disease allele frequency, $p = 0.001$ ($\cdots$), $p = 0.05$ (-·-), $p = 0.1$ (solid) and $p = 0.5$ (- -). The reference model is the liability threshold model with $(\psi_0, \psi_1, \psi_2) = (0.01, 0.5, 0.8)$.

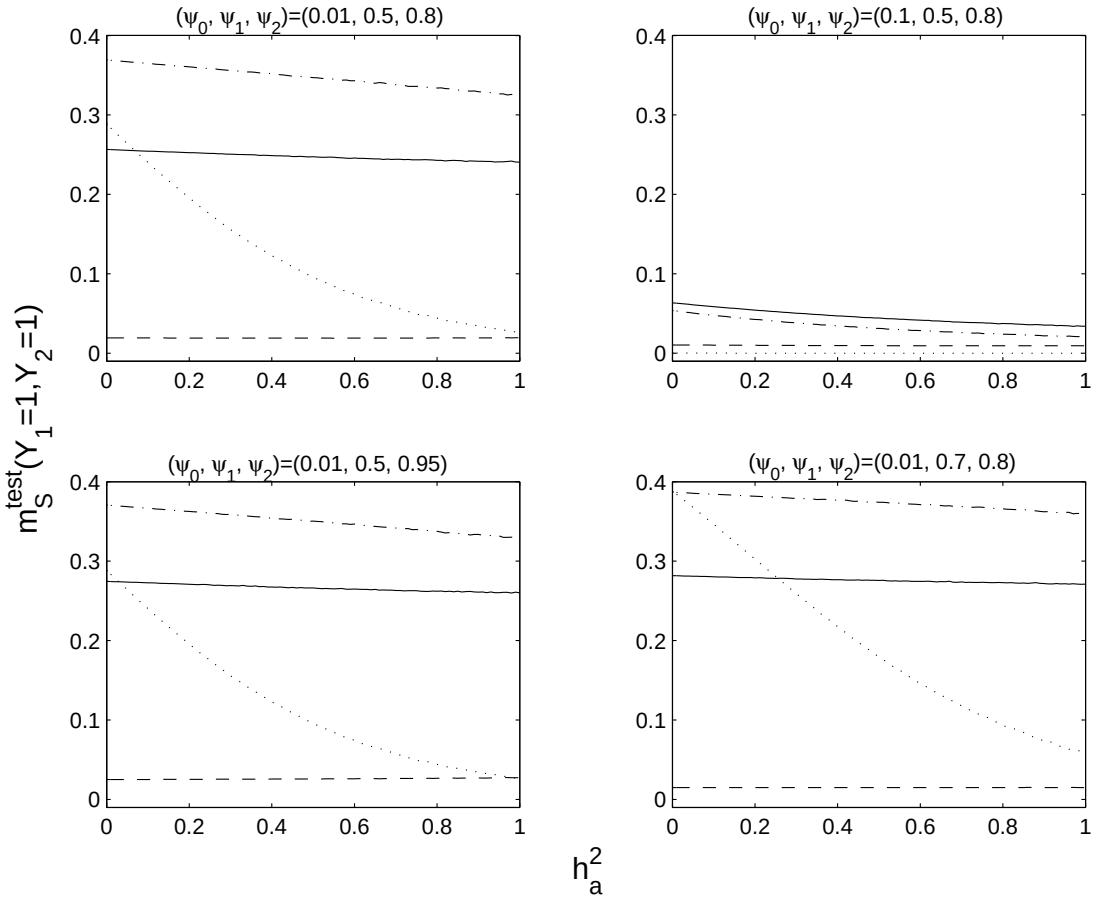

Figure 14: Effective number of meioses m_S^{test} for a sib pair as function of the values of the additive polygenic heritability h_a^2 when $Y_1 = Y_2 = 1$. It is calculated for four different values of disease allele frequency, $p = 0.001$ ($\cdots$), $p = 0.05$ (-·-), $p = 0.1$ (solid) and $p = 0.5$ (- -) and the four different penetrance vectors (ψ_0, ψ_1, ψ_2).