An Examination of Hardy-Weinberg Disequilibrium and Statistical Testing in Genetic Association Studies

by

Vaneeta Kaur Grover

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
October 2010

DALHOUSIE UNIVERSITY

DEPARTMENT OF MATHEMATICS AND STATISTICS

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "An Examination of Hardy-Weinberg Disequilibrium and Statistical Testing in Genetic Association Studies" by Vaneeta Kaur Grover in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

Dated: June 18, 2010

External Examiner: ______________________
Dr. J Concepcin Loredo-Osti

Research Supervisor: ______________________
Dr. David Hamilton

Examining Committee: ______________________
Dr. Chris Field

____
Dr. Bruce Smith

Departmental Representative: ______________________
Dr. Ed Susko

DALHOUSIE UNIVERSITY

Date: June 18, 2010

Author:	Vaneeta Kaur Grover
Title:	An Examination of Hardy-Weinberg Disequilibrium and Statistical Testing in Genetic Association Studies
Department or School:	Department of Mathematics and Statistics

Degree: PhD Convocation: October Year: 2010

Signature of Author

dedicated in particular...

... to my parents, sister and husband,

... and in general...

... to all those who ...

... inspired me to take on the project,

... walked with me through out or in part,

... encouraging during the downs,

... celebrating during achievements,

... supported me to reach the goals,

and

... touched my life through course of the journey.

Table of Contents

List of Tables

List of Figures

Abstract

Departure from HWE (HWD) in a sample may indicate genotyping error, population stratification, selection bias, or some combination thereof. Therefore, loci exhibiting HWD are often excluded from association studies. However, it has been shown that in case-control studies HWD can result from a genetic effect at the locus, and HWD at a marker locus can be interpreted as evidence for association with a disease.

In an unpublished study in Toronto it was observed that cases were in Hardy-Weinberg equilibrium at a locus whereas their family members were in HWD. It has been shown that the HWD coefficient for a multiplicative genetic model is zero. This led to an investigation of relatives of affected individuals to see whether the multiplicative model could be revealed by a nonzero HWD coefficient in relatives. Genotypic frequencies and HWD coefficients were derived for affected individuals and their affected and unaffected relatives. A substantial HWD was found in both individuals in dominant and recessive genetic models but HWD is only slightly nonzero for additive and multiplicative models. Methods were also developed to test for association using data from affected individuals and their relatives. Parameter estimates for these models can be obtained using maximum likelihood methods, and estimates provide valuable information regarding the mode of inheritance of the disease. The methods were applied to 112 discordant sib pairs with Alzheimer's disease typed for the ApoE polymorphism and a significant association was observed between the ε_4 ApoE allele and Alzheimer's disease.

Case-control studies may indicate spurious association with a marker locus in a stratified population. Methods were developed to determine if the HWD observed in a data set from a stratified population can be explained by both genetic association and stratification. Parameter estimates for these models can be obtained using maximum likelihood methods, and used to deduce the mode of inheritance of the disease. Applying the model to the R990G SNP of the CASR gene, it was found that the HWD was adequately explained by a recessive genetic association and a stratification proportion of 10%, consistent with the population of Toronto.

Acknowledgements

My sincere and heartfelt thanks to my thesis supervisor, Dr. David Hamilton, without whose expertise, direction, supervision, guidance and time I could not have accomplished the completion of thesis. Dr. Hamilton, thanks for having faith in me and for your persistent support. I would also like to thank Dr. David Cole for the discussions, comments, and providing data set and ideas for the thesis. I would also like to thank the Thesis Committee for their support and guidance.

It would not have been possible for me to finish thesis if it was not for Dr. Chris Field and the admission committee who admitted me to the program. I would like to thank the School of Graduate Studies, the Department of Mathematics and Statistics and the Lett Bursary for the providing the financial support. Special thanks to Dr. Swaminathan and Dr. Gupta for their constant encouragement and providing me spiritual and moral support through out the duration.

I would also like to thank and acknowledge the financial, emotional and moral support provided by my mom, dad and sister, Velma, who have always inspired me to journey farther than the stars in search of excellence all of your presence in my life, as mentors, family and friend, is a treasured gift. You all inspired me to do PhD and provided me the structure and encouragement to keep going until I reached the finish line. Thanks for being there in my life.

Thanks to my parents-in-law, sisters-in-law and brothers-in-law for providing moral support through out the journey of obtaining the degree. Last but not least thanks to my husband, Himangshu, who tolerated my crazy working style and time; my mood swings and separation for last five years. If it was not for your love, understanding, determination, support and encouragement, I could not have started or finished the degree.

Chapter 1

Introduction

1.1 Introduction

Genetic association studies involve determining whether a genetic variant is associated with a disease. If the allele is associated with disease, it will occur more often than expected by chance in a group of affected individuals. The most common study design for genetic association with a disease is case-control, where allelic or genotypic frequencies at a locus are compared between the case and control groups. The goodness of fit statistic can be used, which has an approximate χ^2 distribution in large samples. The case-control design suffers from several shortcomings. If the case and control groups are not well matched for age, ethnicity, sex, the test could lead to false positive association.

To overcome these problems, Spielman et al. (1993) proposed a transmission disequilibrium test (TDT) which is a family based test to detect genetic linkage only in the presence of genetic association. The test uses the genetic information on a case-parent trio, an affected child and their parents, and it measures the over-transmission of an allele from heterozygous parents to an affected offspring. Some modifications and extension of method like the sib-TDT (Ewens and Spielman, 1995; Spielman and

Ewens, 1998; Horvath and Laird, 1998), and the TDT1 (Fengzhu et al., 1999) among others have been proposed to overcome the limitations of matching and population stratification. However it can be difficult to obtain genotypic information on parents or family members, especially for late onset disease. Association studies are more powerful than linkage studies (Risch and Merikangas, 1996).

A genome-wide association (GWA) study involves examination of markers across a complete set of DNA to identify genetic association with a particular disease. In recent years the cost of genotyping has been reduced to the extent that it has become possible to genotype hundreds of thousands of SNPs in cases and controls. GWAs search the entire genome for association rather than focussing on a small number of candidate genes. However, GWAs involve performing a large number of statistical tests, which leads to the requirement of adjusting the level of significance at each locus. One of the biggest limitations of GWAs is that the results of association are often not replicated in other populations (Hirschhorn et al., 2002; Morgan et al., 2007).

In this thesis a heterogeneous disease model for genetic association is explored that can be used to model observed genotypic counts. The heterogeneous model involves allele frequency and penetrances (probability of disease for a given genotype) and is suitable for complex diseases because it allows phenocopies and non fully penetrant diseases. Models are constructed and applied where both case and control groups are sampled from a population made up of two strata. A model is also developed for family based studies, for use with genotypic data on an affected individual and a relative who is either affected or unaffected.

1.2 Hardy-Weinberg Equilibrium

Hardy-Weinberg equilibrium (HWE) states that in the absence of mutation, migration and selection in a random mating population, both allelic and genotypic frequencies remain constant from one generation to the next and the genotypic frequencies have the same distribution as the frequencies obtained by the random sampling of alleles. For a biallelic locus, with wild and variant type alleles A and a, there are three possible genotypes, AA, Aa and aa, labelled as 0, 1 and 2. Under HWE, the genotypic frequencies, P_0, P_1 and P_2, and allelic frequencies, p_A and p_a, are related as

$$P_0 = p_A^2,$$

$$P_1 = 2p_Ap_a,$$

$$P_2 = p_a^2.$$

HWE is based on the assumptions of large population, no migration, mutation or selection and random mating. If any of the conditions fail to be satisfied, a deviation from HWE, denoted by HWD, may occur. HWD is measured by the Hardy-Weinberg disequilibrium coefficient, which is defined as the difference in the observed and expected genotypic frequency assuming HWE. For a biallelic locus, the disequilibrium coefficient is (Weir, 1996)

$$D = P_{aa} - P_a^2,$$

If there is a deficiency of homozygotes and an excess of heterozygotes $D < 0$, and if there is an excess of homozygotes and a deficiency of heterozygotes $D > 0$. The bounds on the coefficient are

$$\max(-p_A^2, -p_a^2) \leq D \leq p_A p_a,$$

taking into account possible genotypic probabilities (Weir, 1996).

Loci exhibiting HWD are often excluded from association studies, because HWD may be variously interpreted to indicate genotyping error, population stratification, selection bias, or some combination thereof. For case-control studies, Wittke-Thompson et al. (2005) investigated the Hardy-Weinberg coefficient at a biallelic locus A, under a variety of genetic models. In a large number of data sets they found that HWD can be explained by genetic association at the locus. Nielsen et al. (1998) and Lee (2003) also showed that HWD at a marker locus in affected patients can be interpreted as evidence for association with a disease.

1.3 A Heterogeneous Model For Case-Control Studies

Denote the disease status, $d = D$ for patients and $d = C$ for control and denote the sample sizes n_D for cases and n_C controls. For a biallelic locus the data can be summarized as in Table 1.1

In a general disease model, the penetrance, defined as the conditional probability that an individual with genotype AA at the disease-susceptibility locus has the disease, is $\phi_0 = P(D|0)$. Similarly for the genotypes Aa and aa the penetrances are

Table 1.1: Data for a case-control study at a biallelic marker

Genotype	0	1	2	Total
Cases	n_{0D}	n_{1D}	n_{2D}	n_D
Controls	n_{0C}	n_{1C}	n_{2C}	n_C
Total	n_0	n_1	n_2	n

$\phi_1 = P(D|1)$ and $\phi_2 = P(D|2)$ respectively. The baseline penetrance of disease in homozygotes without a risk allele at the locus is $\alpha = \phi_0$, the heterozygote relative risk is $\beta = \phi_1/\phi_0$ and the homozygote relative risk is $\gamma = \phi_2/\phi_0$. The prevalence of disease, K_P is

$$K_P = P_0\phi_0 + P_1\phi_1 + P_2\phi_2,$$

where P_i is the genotypic probability of genotype i, $i = 0, 1, 2$. Assuming random mating and HWE in the population, the genotypic frequencies are $P_0 = (1-q)^2$, $P_1 = 2pq$ and $P_2 = q^2$. The disease prevalence, K_P, can be written as

$$K_P = (1-q)^2\phi_0 + 2q(1-q)\phi_1 + q^2\phi_2.$$

The genotypic probabilities conditional on the affection status d can be written in terms of penetrance for genotype, genotypic probabilities and the prevalence of disease (Wittke-Thompson et al., 2005) as

$$P_{id} = \frac{P(d|i)P_i}{P(d)},$$

and those probabilities are summarized in Table 1.2, assuming HWE.

Table 1.2: Genotypic probabilities for a general disease model in case-control studies

Genotype	0	1	2
Cases	$\frac{\phi_0(1-q)^2}{K_P}$	$\frac{2\phi_1 q(1-q)}{K_P}$	$\frac{\phi_1 q^2}{K_P}$
Controls	$\frac{(1-\phi_0)(1-q)^2}{1-K_P}$	$\frac{2(1-\phi_1)q(1-q)}{1-K_P}$	$\frac{(1-\phi_1)q^2}{1-K_P}$

The parameters of the model are the overall prevalence of the disease in the population, K_P, the genotypic penetrances, ϕ_0, ϕ_1, and ϕ_2 and the minor allele frequency, q. The overall prevalence K_P is assumed to be known, so one of the penetrances can be eliminated from the set of parameters requiring estimation. For example, ϕ_0 can be expressed as

$$\phi_0 = \frac{K_P - 2q(1-q)\phi_1 - q^2\phi_2}{(1-q)^2}. \tag{1.1}$$

As a result, the model is a function of only three parameters, two penetrances, ϕ_1, ϕ_2, and the minor allele frequency, q.

A general lack of fit test was proposed to assess whether the observed genotypic counts were in agreement with the model. Equivalently this test determines whether the HWD at the susceptibility locus can be explained by association with the disease. The test statistic is

$$X^2 = \sum_{d=D}^{C} \sum_{i=0}^{2} \frac{(n_{id} - n_d\hat{P}_{id})^2}{n_d\hat{P}_{id}},$$

where the $\hat{P}_{id}$ are the maximum likelihood estimates of P_{id} under HWE.

There are four degrees of freedom and three parameters to be estimated, so X^2 is asymptotically distributed as χ^2 with one degree of freedom. A small p-value provides evidence that the disease model is a poor fit to the data and the observed HWD is due to another source, such as genotyping error, chance, population stratification, inbreeding or selection.

The genotypic counts for each of the random samples from the affected and unaffected sub-populations have multinomial distributions and give the likelihood function

$$L(q, \mathbf{\Phi}) = \prod_{d=\{D,C\}} \prod_{i=0}^{2} P_{id}^{n_{id}},$$

where $\mathbf{\Phi}$ is a vector containing the penetrances, P_{id} is the genotypic frequency for genotype i and disease status $d = C$ for control and $d = D$ for cases, and n_{id} is the corresponding disease count.

The parameter estimates are obtained by maximizing the multinomial likelihood function L or equivalently by maximizing the natural logarithm of the likelihood $ln(L)$,

$$ln(L) = \sum_{d=\{D,C\}} \sum_{i=0}^{2} n_{id} ln(P_{id}).$$

The likelihood function does not have an explicit analytic solution for the MLEs of the penetrances and allele frequency so they are obtained numerically. Wittke-Thompson et al. (2005) obtained estimates by minimizing the goodness of fit statistic χ^2 , which is asymptotically equivalent. The standard errors for the estimates can be obtained by evaluating the inverse of information matrix at the maximum likelihood estimates. The information matrix is the negative expectation of the Hessian matrix, the matrix of second partial derivatives. Approximate standard errors of the parameter estimates can also be obtained using the non-parametric bootstrap. In this approach, a large number of random samples are selected with replacement from the cases and controls. The model parameters are estimated for each dataset and their standard deviations are calculated.

A likelihood ratio test can be used to assess the difference in fit between two models - a general model and one of the specific disease models (dominant, recessive additive or multiplicative). Testing of the reduced models gives insight into the mode of genetic inheritance of a particular disease. The likelihood ratio test statistic is

$$\Lambda = 2(lnL_1 - lnL_2),$$

where L_1 is the likelihood for a complex model and L_2 corresponds to that of the simplified or reduced model and the likelihoods are evaluated at the maximum likelihood estimates. The statistic Λ has an approximate χ^2 distribution with degrees of freedom equal to the difference between the number of parameters in the two models. A small p-value of the test gives evidence that the more complex model explains the data better than the reduced model.

1.4 Outline Of Thesis

This chapter introduced some of the topics that are central to the thesis. The genotypic frequencies and HWD coefficients for an affected individual and its affected or unaffected relative (sib, parent or grandparent) are derived in Chapter 2. The HWD coefficients under specific genetic models are also discussed. Testing for association between the disease and locus for genotypic data from relative pairs is discussed in

Chapter 3. The model is extended to a stratified population in Chapter 4, where the genotypic frequencies and HWD coefficients are derived. The last chapter, Chapter 5 gives an overall summary of the findings and some future research avenues including the modelling of penetrance as a function of a continuous variable such as age.

Chapter 2

Hardy-Weinberg Disequilibrium Due To Association In Affected Individuals And Their Relatives

2.1 Introduction

In an unpublished study in Toronto it was observed that cases were in Hardy-Weinberg equilibrium (HWE) at a locus whereas their family members were in Hardy-Weinberg disequilibrium (HWD). For case-control studies, HWD can result from a genetic effect at the locus (Wittke-Thompson et al., 2005). Nielsen et al. (1998) and Lee (2003) also showed that HWD at a marker locus can be interpreted as evidence for association with a disease. Wittke-Thompson et al. (2005) also showed that the HWD coefficient in cases is zero for a multiplicative genetic model.

The aim of this chapter is to investigate the HWD coefficient in family members (siblings, parents and grand-parents) of an affected individual under a general genetic model to determine whether HWD in relatives can provide extra information about disease association. The family member could be affected or not affected with the disease under study. For the multiplicative genetic model, the HWD coefficient is found to be non-zero for both the affected individual and its parent and grandparent, and for both siblings in a discordant sibling pair. Any observed departure from HWE in relative pairs could indicate association, once other reasons like genotyping error or stratification has been ruled out.

In Section 2.2 the genotypic frequencies of a relative pair, in general, are derived where the relative of the affected person could be affected or unaffected. In Section 2.3 the genotypic frequencies as well as the HWD coefficient for a pair of siblings are derived and examined for specific genetic models. The genotypic frequencies and HWD coefficient for child-parent and grandchild-grandparent pairs are discussed in Appendix A and Appendix B, respectively. Testing for HWD is discussed in Section 2.4 followed by a comparison of the HWD coefficient and the power of the test for departure from HWD among different relative pairs in Section 2.5. Finally Section 2.6 examines the results obtained by Li and Leal (2008) for discordant sibling pairs and Section 2.7 discusses the results.

2.2 Genotypic Frequencies For A Relative Pair

Consider an affected person with the disease of interest and their relative (sibling, parent or grandparent). Suppose that the genotype of the affected person ($j = 1$) and their relative ($j = 2$) are denoted by g_j, $j = 1, 2$, where $g_j = 0$ (AA), 1 (Aa) or 2 (aa), and that the disease status is denoted by d_j, where $d_j = \mathcal{A}_j$ for affected and $\bar{\mathcal{A}}_j$ for unaffected, $j = 1, 2$. For the affected person, $d_1 = \mathcal{A}_1$ but the relative could have $d_2 = \mathcal{A}_2$ or $\bar{\mathcal{A}}_2$ depending on its disease status. The joint probability of the genotypes of the two relatives, conditional on the disease status of both relatives is

$$\begin{aligned} P(g_1 \cap g_2 | \mathcal{A}_1 \cap d_2) &= \frac{P(g_1 \cap g_2 \cap \mathcal{A}_1 \cap d_2)}{P(\mathcal{A}_1 \cap d_2)} \\ &= \frac{P(\mathcal{A}_1 \cap d_2 | g_1 \cap g_2) P(g_1 \cap g_2)}{P(d_2|\mathcal{A}_1) P(\mathcal{A}_1)}. \end{aligned}$$

The term $P(\mathcal{A}_1)$ is the probability that a person 1 is affected, and is same as the prevalence of the disease, K_P, given by

$$K_P = (1-q)^2\phi_0 + 2q(1-q)\phi_1 + q^2\phi_2 \tag{2.1}$$

where $\phi_i = P(\mathcal{A}|i)$ is the penetrance of the disease for genotype i, $i = 0, 1, 2$ and q is the minor allele frequency.

The disease status of an individual is assumed to depend only on their genotype, therefore,

$$P(g_1 \cap g_2 | \mathcal{A}_1 \cap d_2) = \frac{P(\mathcal{A}_1|g_1) P(d_2|g_2) P(g_1 \cap g_2)}{P(d_2|\mathcal{A}_1) K_P}. \tag{2.2}$$

The joint probability of the genotypes, $P(g_1 \cap g_2)$, depends on the relationship between the two relatives and the allele frequencies.

2.3 Sibling Pairs

Consider an affected person and its sibling. The genotypic frequencies for a pair of siblings are derived below as are the Hardy-Weinberg coefficients.

2.3.1 Genotypic Frequencies

The joint probability of the genotypes of two siblings, $P(g_1 \cap g_2)$, can be obtained by conditioning on the parental mating type, h_i, which is the pair of genotypes of the parents,

$$P(g_1 \cap g_2) = \sum_{i=1}^{m} P(g_1 \cap g_2 | h_i) P(h_i),$$

where m is the total number of mating types. Given the parental mating type, the genotypes of the two siblings are independent of each other, so

$$P(g_1 \cap g_2) = \sum_{i=1}^{m} P(g_1|h_i) P(g_2|h_i) P(h_i). \tag{2.3}$$

Table 2.1 gives the possible parental mating types, their frequencies in the population under HWE and the probabilities of the offspring of different genotypes.

Table 2.1: Possible parental mating types, their probabilities with probabilities of possible offspring genotypes

Mating type h_i			Probability of mating type $P(h_i)$	Conditional Probability of offspring genotype $P(g_j\|h_i)$		
				AA	Aa	aa
AA	×	AA	p^4	1	0	0
AA	×	Aa	$4\ p^3\ q$	0.5	0.5	0
AA	×	aa	$2\ p^2\ q^2$	0	1	0
Aa	×	Aa	$4\ p^2\ q^2$	0.25	0.5	0.25
Aa	×	aa	$4\ p\ q^3$	0	0.5	0.5
aa	×	aa	q^4	0	0	1

Using Table 2.1 the joint probability of $g_1 = g_2 = AA$ is

$$\begin{aligned} P(AA, AA) &= P(AA|AA \times AA)P(AA|AA \times AA)P(AA \times AA) \\ &\quad +P(AA|AA \times Aa)P(AA|AA \times Aa)P(AA \times Aa) \\ &\quad +P(AA|Aa \times Aa)P(AA|Aa \times Aa)P(Aa \times Aa) \\ &= 1.1.p^2p^2 + \frac{1}{2}.\frac{1}{2}2p^2 2pq + \frac{1}{4}\frac{1}{4}.2pq2pq \\ &= \frac{p^2}{4}(4p^2 + 4pq + q^2) \\ &= \frac{p^2}{4}(2p + q)^2 \\ &= \frac{p^2}{4}(p + 1)^2. \end{aligned}$$

Similarly, the other joint probabilities can be obtained (Table 2.2).

Table 2.2: Joint probability of genotypes of a pair of siblings

	Sibling			
Affected	AA	Aa	aa	Total
AA	$\frac{1}{4}p^2(1+p)^2$	$\frac{1}{2}p^2q(1+p)$	$\frac{1}{4}p^2q^2$	p^2
Aa	$\frac{1}{2}p^2q(1+p)$	$pq(1+pq)$	$\frac{1}{2}pq^2(1+q)$	$2pq$
aa	$\frac{1}{4}p^2q^2$	$\frac{1}{2}pq^2(1+q)$	$\frac{1}{4}q^2(1+q)^2$	q^2
Total	p^2	$2pq$	q^2	1

Substituting the joint probability from equation (2.3) in equation (2.2), gives the genotypic frequency conditional on the disease status of both siblings

$$P(g_1 \cap g_2 | \mathcal{A}_1 \cap d_2) = \frac{P(\mathcal{A}_1|g_1)P(d_2|g_2)}{P(d_2|\mathcal{A}_1)K_P} \sum_{i=1}^{m} P(g_1|h_i)P(g_2|h_i)P(h_i). \tag{2.4}$$

Affected Sibling Pair

When both the siblings are affected, $d_2 = \mathcal{A}_2$ and the quantity $P(\mathcal{A}_2|\mathcal{A}_1)$ in equation (2.4) is the sibling recurrence risk denoted by K_S. The joint genotypic probability of an affected sibling pair can be computed using equation (2.4), (Table 2.3). For example, the joint probability of the genotypes AA and AA is

$$\begin{aligned} P(AA \cap AA|\mathcal{A}_1 \cap \mathcal{A}_2) &= \frac{P(\mathcal{A}_2|AA)}{P(\mathcal{A}_2|\mathcal{A}_1)} P(AA \cap AA|\mathcal{A}_1) \\ &= \frac{\phi_0}{K_S}\frac{\phi_0}{4K_P} p^2(p+1)^2 \\ &= \frac{\phi_0^2 p^2}{4K_P K_S}(p+1)^2. \end{aligned}$$

Table 2.3: Joint genotypic probabilities of a pair of affected siblings

	Sibling			
Affected	AA	Aa	aa	Total
AA	$\frac{\phi_0^2 p^2}{4K_P K_S}(1+p)^2$	$\frac{\phi_0\phi_1 p^2 q}{2K_P K_S}(1+p)$	$\frac{\phi_0\phi_2 p^2 q^2}{4K_P K_S}$	$\frac{\phi_0 p^2}{4K_P K_S} S_{AA}$
Aa	$\frac{\phi_0\phi_1 p^2 q}{2K_P K_S}(1+p)$	$\frac{\phi_1^2 pq}{K_P K_S}(1+pq)$	$\frac{\phi_1\phi_2 pq^2}{2K_P K_S}(1+q)$	$\frac{\phi_1 pq}{2K_P K_S} S_{Aa}$
aa	$\frac{\phi_0\phi_2 p^2 q^2}{4K_P K_S}$	$\frac{\phi_1\phi_2 pq^2}{2K_P K_S}(1+q)$	$\frac{\phi_2^2 q^2}{4K_P K_S}(1+q)^2$	$\frac{\phi_2 q^2}{4K_P K_S} S_{aa}$
Total	$\frac{\phi_0 p^2}{4K_P K_S} S_{AA}$	$\frac{\phi_1 pq}{2K_P K_S} S_{Aa}$	$\frac{\phi_2 q^2}{4K_P K_S} S_{aa}$	1

For simplicity of presentation, the expressions have been abbreviated using

$$S_{AA} = \phi_0(p+1)^2 + 2\phi_1(1+p)q + \phi_2 q^2,$$

$$S_{Aa} = \phi_0 p(p+1) + 2\phi_1(1+pq) + \phi_2 q(1+q),$$

and

$$S_{aa} = \phi_0 p^2 + 2\phi_1 p(1+q) + \phi_2(1+q)^2.$$

Using the fact that the probabilities in Table 2.3 sum up to one, the recurrence risk, K_S can be obtained as

$$\begin{aligned} K_S = \; & \frac{1}{4K_P}[p^2(1+p)^2\phi_0^2 + 4(1+p)p^2q\phi_0\phi_1 + 2p^2q^2\phi_0\phi_2 + 4pq(1+pq)\phi_1^2 \\ & +4pq^2(1+q)\phi_1\phi_2 + q^2(1+q)^2\phi_2^2]. \end{aligned} \tag{2.5}$$

In this case, both siblings are affected so Table 2.3 is symmetric.

Discordant Sibling Pair

The joint genotypic probabilities can be also obtained for the case of discordant sibling pair (Table 2.4) using Table 2.1 and equation (2.4) with $P(d_2|\mathcal{A}_1) = P(\bar{\mathcal{A}}_2|\mathcal{A}_1) = 1 - K_S$. For example, the joint probability of the genotypes AA and AA of the two discordant siblings is

$$\begin{aligned} P(AA \cap AA|\mathcal{A}_1 \cap \bar{\mathcal{A}}_2) &= \frac{P(\bar{\mathcal{A}}_2|AA)}{P(\bar{\mathcal{A}}_2|\mathcal{A}_1)} P(AA \cap AA|\mathcal{A}_1) \\ &= \frac{(1-\phi_0)}{(1-K_S)} \frac{\phi_0}{4K_P} p^2(p+1)^2 \\ &= \frac{\phi_0(1-\phi_0)p^2}{4K_P(1-K_S)}(p+1)^2. \end{aligned}$$

Table 2.4: Joint probability of genotypes of the discordant sibling pair

	Sibling			
Affected	AA	Aa	aa	Total
AA	$\frac{\phi_0(1-\phi_0)p^2(1+p)^2}{4K_P(1-K_S)}$	$\frac{\phi_0(1-\phi_1)p^2q(1+p)}{2K_P(1-K_S)}$	$\frac{\phi_0(1-\phi_2)p^2q^2}{4K_P(1-K_S)}$	$\frac{\phi_0p^2}{4K_P(1-K_S)}SN_{AA}$
Aa	$\frac{\phi_1(1-\phi_0)p^2q(1+p)}{2K_P(1-K_S)}$	$\frac{\phi_1(1-\phi_1)pq(1+pq)}{K_P(1-K_S)}$	$\frac{\phi_1(1-\phi_2)pq^2(1+q)}{2K_P(1-K_S)}$	$\frac{\phi_1pq}{2K_P(1-K_S)}SN_{Aa}$
aa	$\frac{\phi_2(1-\phi_0)p^2q^2}{4K_P(1-K_S)}$	$\frac{\phi_2(1-\phi_1)pq^2(1+q)}{2K_P(1-K_S)}$	$\frac{\phi_2(1-\phi_2)q^2(1+q)^2}{4K_P(1-K_S)}$	$\frac{\phi_2q^2}{4K_P(1-K_S)}SN_{aa}$
Total	$\frac{(1-\phi_0)p^2}{4K_P(1-K_S)}S_{AA}$	$\frac{(1-\phi_1)pq}{2K_P(1-K_S)}S_{Aa}$	$\frac{(1-\phi_2)q^2}{4K_P(1-K_S)}S_{aa}$	1

For simplicity of presentation, the expressions have been abbreviated using S_{AA}, S_{Aa} and S_{aa} described above, and

$$SN_{AA} = (1-\phi_0)(1+p)^2 + 2(1-\phi_1)q(1+p) + (1-\phi_2)q^2,$$

$$SN_{Aa} = (1-\phi_0)p(1+p) + 2(1-\phi_1)(1+pq) + (1-\phi_2)q(1+q),$$

and

$$SN_{aa} = (1-\phi_0)p^2 + 2(1-\phi_1)p(1+q) + (1-\phi_2)(1+q)^2.$$

2.3.2 Hardy-Weinberg Coefficient

The Hardy-Weinberg coefficient, D, measures the excess homozygosity and is given by

$$D = P_{aa} - q^2,$$

where the minor allele frequency can be obtained from the genotypic frequencies using the relationship

$$q = P_{aa} + \frac{1}{2} P_{Aa}.$$

Affected Sibling Pair

For an affected sibling pair (Table 2.3), the minor allele frequency is the same for both siblings and is

$$q_{i\mathcal{A}} = \frac{\alpha^2 \gamma q^2}{4K_P K_S}[p^2 + 2\beta p(1+q) + \gamma(1+q)^2] + \frac{\alpha^2 \beta pq}{4K_P}[p(1+p) + 2\beta(1+pq) + \gamma q(1+q)],$$

where α, β and γ are the baseline, heterozygote and homozygote relative risks. It simplifies to

$$q_{1A} = q_{2A} = \frac{\alpha^2 q}{4K_P K_S}\{q(1+q)^2\gamma^2 + p^2 q\gamma + 3pq(1+q)\beta\gamma + p^2(1+p)\beta + 2p(1+pq)\beta^2\}.$$

The HWD for both siblings is also the same, and is given by

$$\begin{aligned} D_{i\mathcal{A}} = \frac{\alpha^4 p^2 q^2}{16K_P^2 K_S^2}(& -4(1+pq)^2\beta^4 + \{4(1+pq)[q(1+q)\gamma + p(2-q)]\beta \\ & + q^2(1+q)^2\gamma^2 + p^2(2-q)^2\}(\gamma - \beta^2) \\ & + 2pq(2-q)(1+q)\gamma\beta^2 + (4 + 2q^4 - 4q^3 - 2q^2 + 4q)\gamma^2) \end{aligned}$$

Discordant Sibling Pair

For the discordant sibling pair (Table 2.4), minor allele frequencies of the affected person and unaffected sibling are denoted by $q_{1\bar{\mathcal{A}}}$ and $q_{2\bar{\mathcal{A}}}$. For the affected sibling, the minor allele frequency is

$$\begin{aligned}q_{1\bar{A}} =&\frac{\alpha\gamma q^2}{4K_P(1-K_S)}[(1-\alpha)p^2+2(1-\alpha\beta)p(1+q)+(1-\alpha\gamma)(1+q)^2]\\&+\frac{\alpha\beta pq}{4K_P(1-K_S)}[(1-\alpha)p(1+p)+2(1-\alpha\beta)(1+pq)+(1-\alpha\gamma)q(1+q)]\end{aligned}$$

which can also be written as

$$\begin{aligned}q_{1\bar{A}} = \frac{-\alpha^2 q}{K_P(1-K_S)}&\{2p(1+pq)\beta^2+[3pq(1+q)\gamma+p^2(2-q)]\beta+\gamma qp^2+q(1+q)^2\gamma^2\\&-\frac{4}{\alpha}(\gamma q+\beta p)\},\end{aligned}$$

and the minor allele frequency for its unaffected sibling is

$$\begin{aligned}q_{2\bar{A}} = \frac{-\alpha^2 q}{K_P(1-K_S)}&\{2p(1+pq)\beta^2+[3pq(1+q)\gamma+p^2(2-q)]\beta+p^2q\gamma+q(1+q)^2\gamma^2\\&-\frac{2}{\alpha}[q(1+q)\gamma-p(2q+1)\beta-p^2]\}.\end{aligned}$$

The HWD coefficients for the two siblings are

$$\begin{aligned}D_{1\bar{A}} = \frac{-\alpha^4p^2q^2}{16K_P^2(1-K_S)^2}\quad&\Big(-q^2(1+q)^2\gamma^3+[q^2(1+q)^2\beta^2-4q(1+q)(1+pq)\beta\\&+2q^2(pq+q+1)-4(1+q)]\gamma^2+[4q(1+q)(1+pq)\beta^3\\&-2qp(2-q)(1+q)\beta^2-4p(2-q)(1+pq)\beta\\&-p^2(2-q)^2]\gamma+4(1+pq)^2\beta^4+4p(2-q)(1+pq)\beta^3\\&+p^2(2-q)^2\beta^2+\frac{4}{\alpha}\{(2q^2+1+2q)\gamma^2+[-2q(1+q)\beta^2\\&+2(2pq+1)\beta+(2q^2-6q+5)]\gamma-2p(2-q)\beta^2\\&-4(1+pq)\beta^3\}+\frac{16}{\alpha^2}(\beta^2-\gamma)\Big)\end{aligned}$$

and

$$\begin{aligned} D_{2\bar{A}} = \frac{-\alpha^4 p^2 q^2}{16K_P^2(1-K_S)^2}\Big(& - q^2(1+q)^2\gamma^3 + [4q^3 + (2pq^2-4)(1+q) \\ & + q^2(1+q)^2\beta^2 - 4q(1+q)(1+pq)\beta]\gamma^2 \\ & + 4p(2-q)(1+pq)\beta^3 + [4q(1+q)(1+pq)\beta^3 \\ & - 2pq(2-q)(1+q)\beta^2 - q^4 - 4p(2-q)(1+pq)\beta \\ & + (3q-2)(1+p-2pq)]\gamma + 4(1+pq)^2\beta^4 + p^2(2-q)^2\beta^2 \\ & + \frac{1}{\alpha}\{4 + [4 + 2pq(3q+2)(1+q)\beta \\ & + q(1+q)(3q^2-5q+4)]\gamma^2 + q^2(1+q)^2\gamma^3 \\ & + [-4q^2(1+q)(5-3q)\beta^2 - 10q^3 \\ & + (-8q + 24q^3 - 4q^2 - 12q^4 + 8)\beta + 11q^2 + 3q^4 + 8p]\gamma \\ & + 2pq(5-3q)(2-q)\beta - 8(1+pq)^2\beta^3 \\ & - q(3-q)(q^2-3q+4) - 4p^2(3q+2)(2-q)\beta^2\} \\ & + \frac{4}{\alpha^2}(\beta^2-\gamma)\Big). \end{aligned}$$

Unrelated Affected Individuals

The HWD coefficient for an unrelated, affected individual is given in by the expressions for cases in

$$D_P = \frac{\alpha^2 p^2 q^2}{K_P^2}(\gamma - \beta^2).$$

Unrelated Unaffected Individuals

The HWD coefficient for an unrelated, unaffected individual is given in by the expressions for controls in

$$D_C = \frac{\alpha p^2 q^2}{(1-K_P)^2}(2\beta - 1 - \gamma - \alpha\beta^2 + \alpha\gamma).$$

In order to understand the magnitude and direction of the HWD coefficient D, it was studied and graphed under some specific genetic models discussed in the next section.

2.3.3 Specific Genetic Models

The HWD coefficients for a pair of siblings were studied in specific genetic models for both cases regarding the disease status of sibling of the affected person. For each model, the HWD coefficient is plotted for two different values of the homozygote relative risk, γ, 1.5 and 3.

Dominant Model

In the dominant model, the homozygote and heterozygote relative risks are equal implying the penetrances for the genotype Aa and aa to be the same i.e., $\phi_2 = \phi_1$ or $\beta = \gamma, \gamma > 1$.

Affected Sibling Pair

When both siblings are affected, the HWD coefficient is the same for both siblings

$$\begin{aligned} D_{1_{\mathcal{A}}} = D_{2_{\mathcal{A}}} = -\frac{\alpha^4 p^2 q^2 \gamma(\gamma-1)^2}{4K_P^2 K_S^2}\{&-(\gamma-1)q^3(6-q) + q[(5\gamma-13)q + 12(\gamma+1)] \\ &+ \frac{4}{(\gamma-1)}(\gamma^2+3\gamma+1)\}. \end{aligned}$$

Discordant Sibling Pair

For a pair of discordant siblings the HWD coefficients are

$$D_{1\bar{A}} = \frac{-\alpha^4 p^2 q^2 \gamma(\gamma-1)}{16K_P^2(1-K_S)^2}\{-(\gamma-1)^2 q^3(6-q) + (\gamma-1)q[(5\gamma-13)q + 12(\gamma+1)] \\ + 4(\gamma^2+3\gamma+1) - \frac{4}{\alpha}[2(\gamma-1)q(3-q) + 4\gamma + 5] + \frac{16}{\alpha^2}\},$$

and

$$D_{2\bar{A}} = \frac{-\alpha^4 p^2 q^2}{16K_P^2(1-K_S)^2}\{\gamma(5\gamma-13)(\gamma-1)^2 q^2 + 4\gamma(\gamma+1)(\gamma^2+\gamma-1) \\ + \gamma(\gamma-1)^3 q^4 + 12\gamma(\gamma+1)(\gamma-1)^2 q - 6\gamma(\gamma-1)^3 q^3 - 8\gamma^2 \\ + \frac{1}{\alpha}[-(\gamma-1)^3 q^4 - (5\gamma-13)(\gamma-1)^2 q^2 - (12(\gamma+1))(\gamma-1)^2 q \\ - (4(\gamma-1))(2\gamma+1)(\gamma+1) + (-18\gamma^2 + 18\gamma + 6\gamma^3 - 6)q^3] \\ + \frac{4}{\alpha^2}\gamma(\gamma-1)\}.$$

Unrelated Affected Individuals

The HWD coefficient for unrelated affected individuals is, Wittke-Thompson et al. (2005)

$$D_P = \frac{-\alpha\gamma p^2 q^2}{K_P^2}(\gamma-1),$$

which is negative.

Unrelated Unaffected Individuals

The HWD coefficient for unrelated unaffected individuals is, Wittke-Thompson et al. (2005)

$$D_C = \frac{\alpha p^2 q^2}{(1-K_P)^2}(\gamma-1)(1-\alpha\gamma),$$

which is positive.

Figure 2.1 illustrates the direction and magnitude of the HWD coefficient in the dominant model for the affected person and its sibling. Also shown are the coefficients for unrelated affected and unaffected individuals.

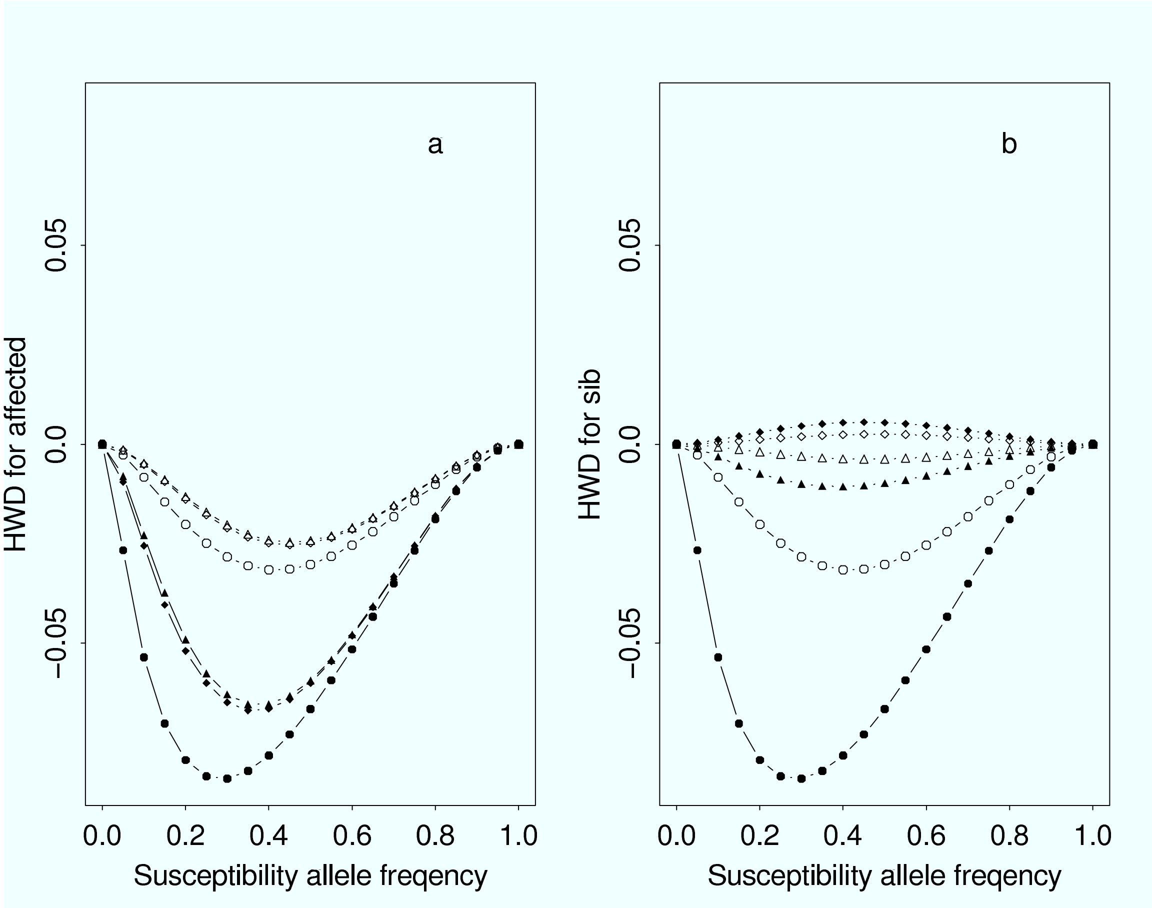


Figure 2.1: HWD coefficients for the dominant genetic model as a function of the susceptibility-allele frequency for an affected person (a) and their sibling or unrelated unaffected person (b).
$K_P = 0.1$, disease status of sibling (affected, unaffected) = ($\circ$, $\triangle$), unrelated affected/unaffected $\Diamond$, open/filled symbol for $\gamma = 1.5/3$.

The HWD coefficient is always negative except for unrelated unaffected individuals for whom the coefficient is positive. The coefficient increases with γ and is largest in magnitude at q between .3 and .5 for both siblings in an affected sibling pair ($\circ$).

Panel (a) shows that the HWD coefficient for the affected person with an unaffected sibling ($\triangle$) is very similar to the unrelated affected individuals ($\diamondsuit$). When the sibling is unaffected (panel (b)), its HWD coefficient is smaller in magnitude than that of the affected person (panel (a)).

Recessive Model

Under a recessive model, having two copies of the variant allele leads to an increased risk of disease susceptibility and the heterozygote relative risk is one, i.e., $\phi_0 = \phi_1$, $\phi_2 > \phi_1$ or $\beta = 1, \gamma > 1$.

Affected Sibling Pair

The HWD coefficients for a pair of affected siblings are the same, and are

$$D_{1\mathcal{A}} = D_{2\mathcal{A}} = \frac{\alpha^4 p^2 q^2 (\gamma - 1)}{16 K_P^2 K_S^2} \{(\gamma - 1)^2 q^3 (q + 2) + (\gamma - 1) q [(\gamma + 7) q + 8] + 4(\gamma + 4)\}.$$

The HWD coefficient is always positive in this case.

Discordant Sibling Pair

For a pair of discordant siblings the HWD coefficients are

$$\begin{aligned} D_{1\bar{\mathcal{A}}} = \frac{\alpha^4 p^2 q^2 (\gamma - 1)}{16 K_P^2 (1 - K_S)^2} \{&(\gamma - 1)^2 q^3 (q + 2) + (\gamma - 1) q [(\gamma + 7) q + 8] + 4(\gamma + 4) \\ &- \frac{4}{\alpha} [2(\gamma - 1) q (1 + q) + (\gamma + 8)] + \frac{16}{\alpha^2}\} \end{aligned}$$

and

$$\begin{aligned} D_{2\bar{\mathcal{A}}} = \frac{-p^2 q^2 \alpha^3 (1 - \alpha)(\gamma - 1)}{16 K_P^2 (1 - K_S)^2} \{&(\gamma - 1)^2 q^3 (q + 2) + (\gamma - 1) q [(\gamma + 7) q + 8] \\ &+ 4(\gamma + 4) - \frac{4}{\alpha}\}. \end{aligned}$$

Unrelated Affected Individuals

The HWD coefficient is (Wittke-Thompson et al., 2005)

$$D_P = \frac{\alpha^2 p^2 q^2}{K_P^2}(\gamma - 1)$$

which is positive.

Unrelated Unaffected Individuals

The HWD coefficient is (Wittke-Thompson et al., 2005)

$$D_C = \frac{-\alpha p^2 q^2}{(1 - K_P)^2}(\gamma - 1)(1 - \alpha)$$

which is negative.

Figure 2.2 illustrates the direction and magnitude of HWD in the recessive genetic model for the affected person and its sibling. Also shown are the coefficients for the unrelated affected and unaffected individuals. The HWD coefficient always increases in magnitude with γ and reaches a maximum when q is between 0.3 and 0.5. The HWD coefficients for the affected person is largest when the sibling is also affected (○). Panel (a) shows that the HWD coefficients for the affected person with an unaffected sibling (△) is very similar to that of unrelated affected individuals (◊). When the sibling is unaffected (panel (b), △), its HWD coefficient is smaller in magnitude than that of the affected person (panel (a), △) and is negative. Among all cases, the affected sibling pair shows largest deviation from HWE. The HWD shown in Figure 2.2 is similar to that in Figure 2.1 for the dominant model but with the signs reversed.

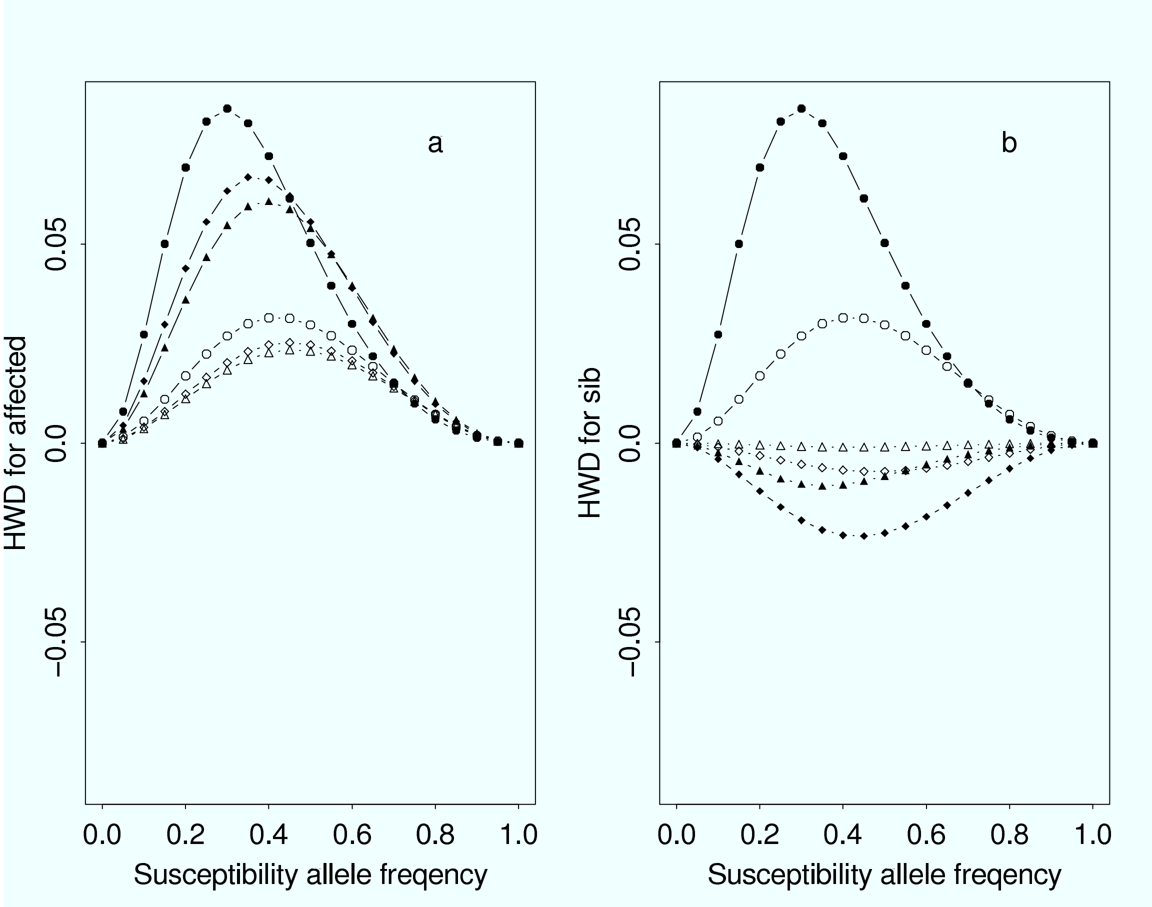


Figure 2.2: HWD coefficients for the recessive genetic model for an affected person (a) and their sibling or unrelated unaffected person (b).
$K_P = 0.2$, disease status of sibling (affected, unaffected) = ($\circ$, $\triangle$), unrelated affected/unaffected $\Diamond$, open/filled symbol for $\gamma = 1.5/3$.

Additive Model

In an additive genetic model, the difference between the homozygote and heterozygote penetrance is the same as the difference between the baseline and homozygote penetrance, i.e., $\phi_2 - \phi_1 = \phi_1 - \phi_0$ or $\phi_2 = 2\phi_1 - \phi_0$ so that homozygote relative risk, γ, is $\gamma = 2\beta - 1$, for $\beta > 1$. Wittke-Thompson et al. (2005) defined the additive model as $\gamma = 2\beta$ which is different from this definition.

Affected Sibling Pair

For a pair of affected siblings the HWD coefficients are both equal to

$$D_{1\mathcal{A}} = D_{2\mathcal{A}} = -\frac{\alpha^4 p^2 q^2 (\gamma-1)^2}{64 K_P^2 K_S^2}\{4(\gamma-1)^2 q^2 + 4(\gamma+3)(\gamma-1)q + (\gamma+9)(\gamma+1)\}.$$

which are always negative.

Discordant Sibling Pair

For a pair of discordant siblings the HWD coefficients are

$$\begin{aligned} D_{1\bar{\mathcal{A}}} = \frac{-\alpha^4 p^2 q^2 (\gamma-1)^2}{64 K_P^2 (1-K_S)^2} \quad & \{4(\gamma-1)^2 q^2 + 4(\gamma+3)(\gamma-1)q + (\gamma+9)(\gamma+1) \\ & -\frac{8}{\alpha}[2(\gamma-1)q + \gamma + 3] + \frac{16}{\alpha^2}\}, \end{aligned}$$

and

$$\begin{aligned} D_{2\bar{\mathcal{A}}} = \frac{-\alpha^4 p^2 q^2 (\gamma-1)^2}{64 K_P^2 (1-K_S)^2} \quad & \{4(\gamma-1)^2 q^2 + 4(\gamma+3)(\gamma-1)q + (\gamma+9)(\gamma+1) \\ & +\frac{4}{\alpha^2}(1-\alpha\gamma-\alpha)\}. \end{aligned}$$

Unrelated Affected Individuals

The HWD coefficient for unrelated affected individuals is

$$D_P = \frac{-\alpha^2 p^2 q^2}{4 K_P^2}(\gamma-1)^2,$$

which is always negative.

Unrelated Unaffected Individuals

The HWD coefficient for unrelated unaffected individuals is

$$D_C = \frac{-\alpha^2 p^2 q^2}{4(1-K_P)^2}(\gamma-1)^2,$$

which is always negative.

Figure 2.3 illustrates the direction and magnitude of the HWD coefficient in the additive model for the affected person and its sibling. Also shown are the coefficients for the unrelated affected and unaffected individuals.

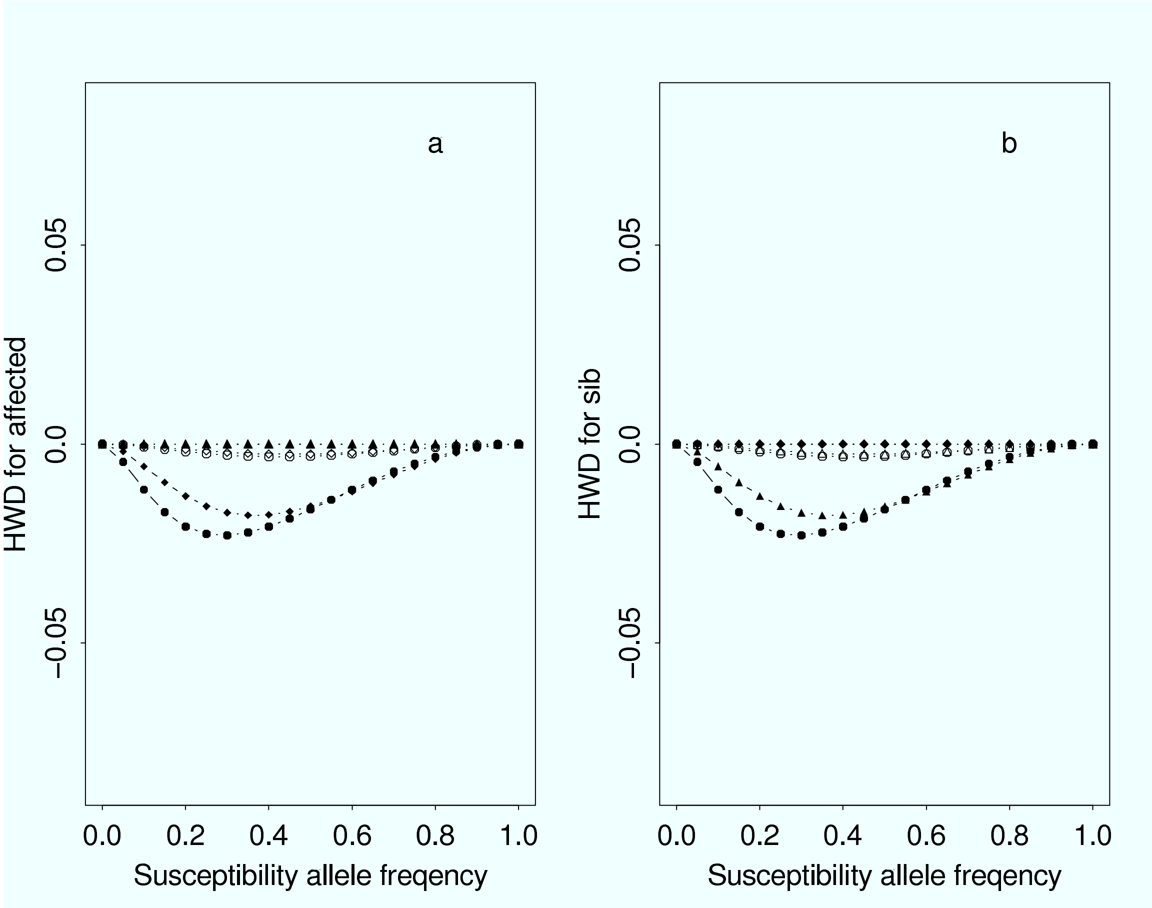


Figure 2.3: HWD coefficients for the additive genetic model for an affected person (a) and their sibling or unrelated unaffected person (b).
$K_P = 0.01$, disease status of sibling (affected, unaffected) = (○, △), unrelated affected/unaffected ◊, open/filled symbol for $\gamma = 1.5/3$.

The HWD coefficient is always negative indicating an excess of heterzygote genotypes. The largest magnitude is obtained for both siblings for q between .3 and .5

when both are affected. When the sibling is unaffected (panel (b), $\triangle$), its HWD coefficient is larger in magnitude than that of the affected person (panel (a), $\triangle$). Note that the overall magnitude of the HWD coefficient is smaller for the additive model than for the dominant or recessive models.

Multiplicative Model

In a multiplicative model, the homozygote relative risk is the square of the heterozygote relative risk, i.e. $\phi_2 = \phi_1^2/\phi_0$ or $\gamma = \beta^2, \beta > 1$.

Affected Sibling Pair

When the disease status of the sibling is affected, the HWD coefficients are both zero

$$D_{1\mathcal{A}} = D_{2\mathcal{A}} = 0.$$

Discordant Sibling Pair

For a pair of discordant siblings, the HWD coefficients are

$$D_{1\bar{\mathcal{A}}} = \frac{-\alpha^3 p^2 q^2 \gamma}{4K_P^2(1-K_S)^2}(\sqrt{\gamma}-1)^2,$$

for the affected individual, which is always negative, and

$$\begin{aligned} D_{2\bar{\mathcal{A}}} = \frac{-\alpha^3 p^2 q^2}{16K_P^2(1-K_S)^2}\{& q^2(1+q)^2\gamma^3 + 2pq(3q+2)(1+q)\gamma^{5/2} \\ & + (15q^4 - 10q^3 - 21q^2 + 4q + 4)\gamma^2 + 4pq(5pq-6)\gamma^{3/2} \\ & + (15q^4 - 50q^3 + 39q^2 + 8q - 8)\gamma + 2pq(5-3q)(2-q)\sqrt{\gamma} \\ & + p^2(2-q)^2\}. \end{aligned}$$

for the unaffected sibling.

Wittke-Thompson et al. (2005) showed that the HWD coefficient is zero for cases and non-zero for controls under the multiplicative model (see below). The results above show that when the sibling is unaffected, the HWD coefficient for the affected person is non-zero and is always negative.

Unrelated Affected Individuals

The HWD coefficient for unrelated affected individuals is zero (Wittke-Thompson et al., 2005)

$$D_P = 0.$$

Unrelated Unaffected Individuals

The HWD coefficient for unrelated unaffected individuals (Wittke-Thompson et al., 2005)

$$D_C = \frac{-\alpha p^2 q^2}{(1-K_P)^2}(\sqrt{\gamma}-1)^2,$$

which is always negative.

Figure 2.4 illustrates the direction and magnitude of the HWD coefficient in the multiplicative model for affected individuals and their sibling. Also shown are the coefficients for unrelated affected and unaffected individuals. The vertical scale is the same as in the previous figures, and the values, when not zero, are small.

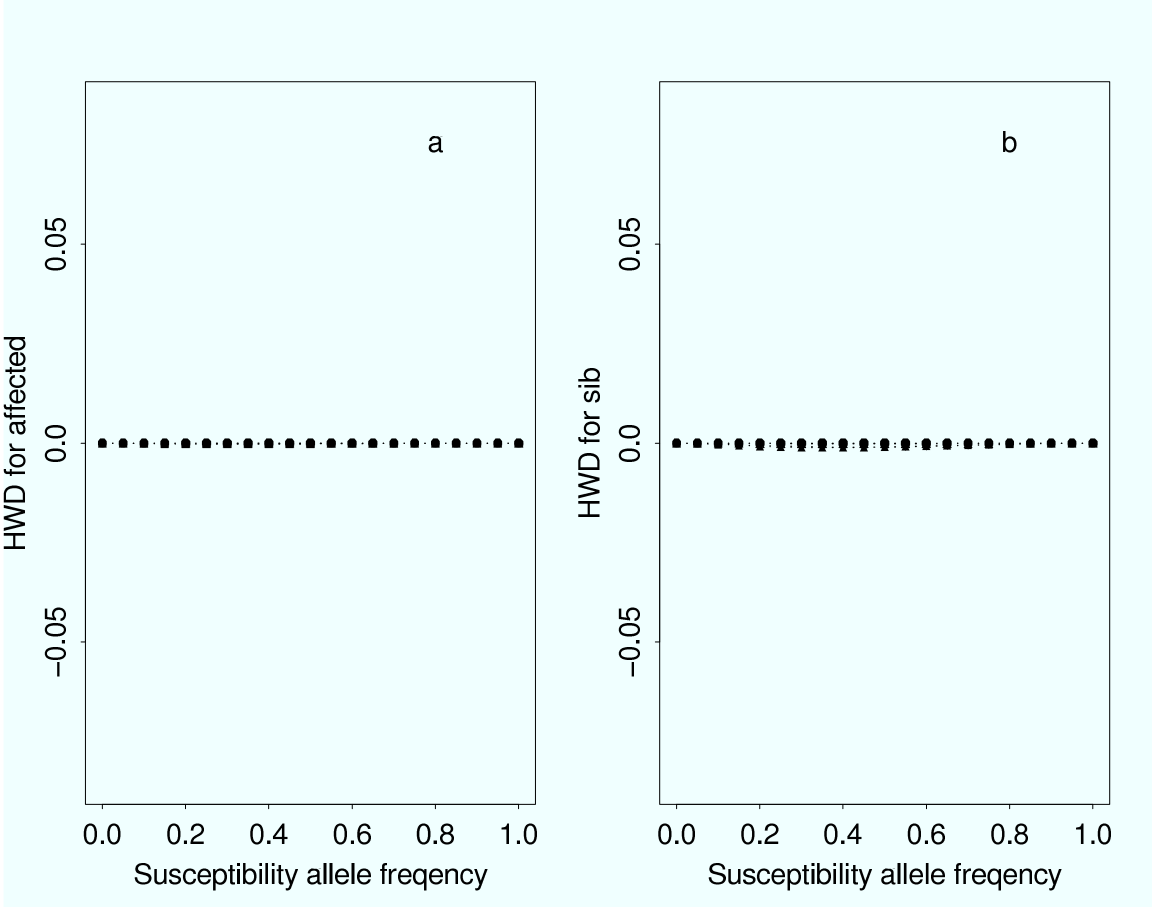


Figure 2.4: HWD coefficients for the multiplicative genetic model for an affected person (a) and their sibling or unrelated unaffected person (b).
$K_P = 0.05$, disease status of sibling (affected, unaffected) = ($\circ$, $\triangle$), unrelated affected/unaffected $\Diamond$, open/filled symbol for $\gamma = 1.5/3$.

2.4 Testing For Hardy-Weinberg Disequilibrium

The test of departure from Hardy-Weinberg equilibrium involves testing the hypotheses

$$\begin{aligned} & H_0: \quad P_{AA} = (1-q)^2, P_{Aa} = 2q(1-q) \text{ and } P_{aa} = q^2, \\ \text{versus} \quad & H_a: \quad P_{AA}, P_{Aa}, P_{aa} \text{ differ from above.} \end{aligned}$$

One test statistic for this test is

$$X^2 = \frac{N\hat{D}^2}{\hat{q}^2(1-\hat{q}^2)},$$

where N is the number of individuals, $\hat{q}$ is the estimated minor allele frequency and $\hat{D}$ is the HWD coefficient. Under H_0, the test statistic, X^2 has a χ^2 distribution in large samples with one degree of freedom for a biallelic locus (Weir, 1996).

The power of the test is approximately

$$Pr(\chi_1^2(\nu) \geq \chi_{1,1-\alpha}^2),$$

where the non-centrality parameter of the non-central χ_1^2 distribution is (Weir, 1996)

$$\nu = \frac{ND^2}{q^2(1-q)^2}.$$

The ability to detect deviations from HWE depends on the magnitude of the HWD coefficient, D, the sample sizes, N, the minor allele frequency, q and the level of significance, α.

For example, suppose $\hat{q} = 0.3$, $\hat{D} = 0.084$ (which corresponds to the largest HWD coefficient in Figure 2.2) and $n = 100$, then the noncentrality parameter is

$$\begin{aligned} \nu &= \frac{100 \times 0.084^2}{.3^2(1-.3)^2} \\ &= \frac{.7056}{.0441} \\ &= 16 \end{aligned}$$

and the power to detect HWD is

$$\begin{aligned} Power &= Pr(\chi_1^2(16) \geq \chi_{1,1-.05}^2) \\ &= Pr(\chi_1^2(16) \geq 3.8415) \\ &= 1 - 0.0207 \\ &= 0.9793. \end{aligned}$$

If other possible sources of HWD can be eliminated, rejection of the hypothesis of HWE can indicate genetic association.

For the multiplicative model in case-control studies Wittke-Thompson et al. (2005) showed that the HWD coefficient is zero in cases, so a test for HWD cannot be used to reveal association. The analysis of HWD in a pair of siblings shows that the coefficient is non zero for both siblings of a discordant pair. Figure 2.4 shows, however, the magnitude of the coefficient is small, so that the power to detect HWD and genetic association is small. For example, the largest deviation from HWD is -0.0011 that occurs for $q = 0.35$ (Figure 2.4). For 100 sibling pairs, the noncentrality parameter is, $\nu = 0.0024$ and the power to detect HWD is 0.0503. For 1000 sibling pairs, the noncentrality parameter is, $\nu = 0.0234$ and the power to detect HWD is 0.0527.

2.5 Comparison Among Different Relative Pairs

The genotypic probabilities and HWD coefficients have been obtained similarly for an affected person and its parent (see Appendix A) or grand parent (see Appendix B) when the relative (parent or grandparent) is affected or unaffected. The expressions for the HWD coefficients are quite complicated even for specific genetic models. Figures 2.5, 2.6, 2.7 and 2.8 give the HWD coefficients and power to detect HWD for the affected individuals and their relatives for dominant, recessive, additive and multiplicative genetic models respectively. They also illustrate the coefficients and power to detect HWD for unrelated affected and unaffected individuals. The power is calculated for the sample size of 1000 and level of significance, $\alpha = 0.05$.

2.5.1 Dominant Model

In the dominant model, the homozygote and heterozygote relative risks are equal implying the penetrances for the genotype Aa and aa are the same i.e., $\phi_2 = \phi_1$ or $\beta = \gamma, \gamma > 1$.

Figure 2.5 illustrates the direction and magnitude of the HWD coefficients and corresponding power for affected individual and their relatives when the disease status of the relative is affected or unaffected. The HWD coefficient is negative for all cases except for the unaffected parents and unrelated individuals. The affected individual shows the largest deviation from HWE when its sibling is affected. The power to detect HWD increases with the minor allele frequency to $q = 0.5$ and then decreases. The power to detect HWD among the affected individuals is least when its sibling is unaffected. Power to detect HWD is least for affected parents and unrelated unaffected individuals.

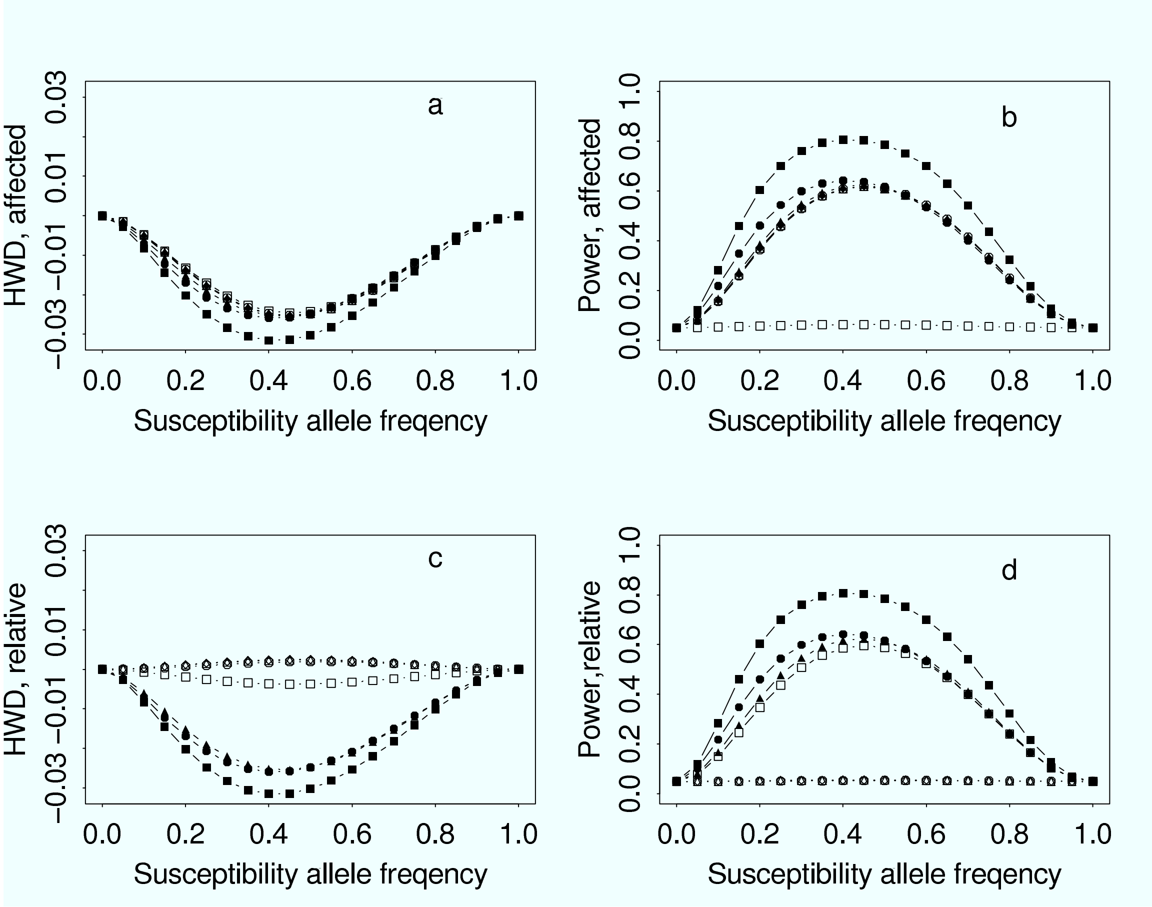


Figure 2.5: HWD coefficients (panels (a) and (c)) and power to detect HWD (panels (b) and (d)) for the dominant model as a function of the susceptibility-allele frequency for an affected individual and its relative.
$K_P = 0.1$, $\gamma = 1.5$, $n = 500$, relative pairs (sibling, parent, grandparent) = ($\Box$, $\circ$, $\triangle$), unrelated = $\Diamond$. Open/filled symbols for disease status of relative or unrelated individual, unaffected/affected.

2.5.2 Recessive Model

Under a recessive model, having two copies of the variant allele leads to an increased risk of disease susceptibility and the heterozygote relative risk is one, i.e., $\phi_0 = \phi_1$, $\phi_2 > \phi_1$ or $\beta = 1, \gamma > 1$.

Figure 2.6 and illustrates the direction and magnitude of the HWD coefficients for the affected individual and their relatives.

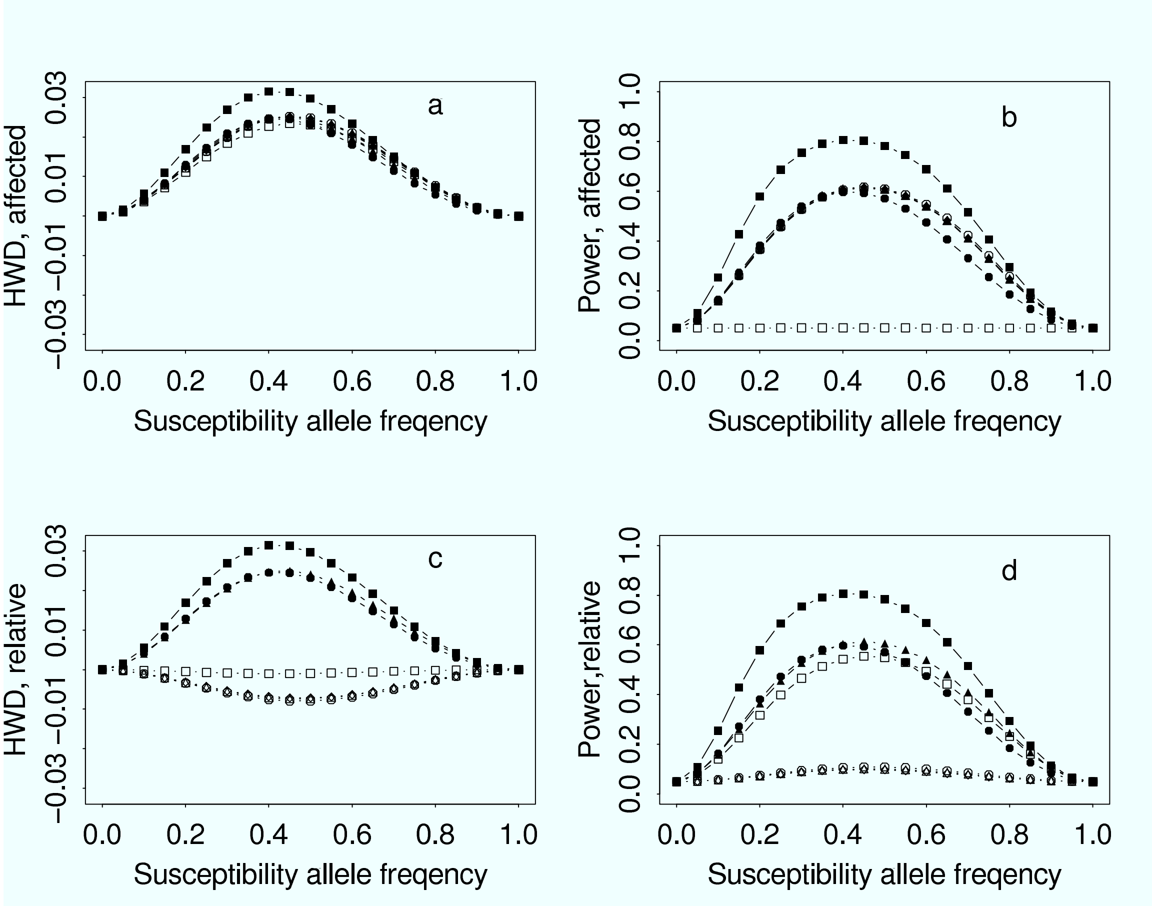


Figure 2.6: HWD coefficients (panels (a) and (c)) and power to detect HWD (panels (b) and (d)) for the recessive model.
$K_P = 0.2$, $\gamma = 1.5$, $n = 500$, relative pairs (sibling, parent, grandparent) = ($\Box$, $\circ$, $\triangle$), case-control = $\Diamond$. Open/filled symbols for disease status of relative or unrelated individual, unaffected/affected.

The HWD coefficient is always positive for the affected individuals. The coefficient is positive for the affected relatives but negative for the unaffected relatives and unrelated unaffected individuals. The power to detect HWD increases with the minor allele frequency to $q = 0.5$ and then decreases. The power to detect HWD for affected individuals is large for all cases except when its sibling is not affected. The power to detect HWD for the unaffected grandparents is similar to that of the unrelated unaffected individuals which is the smallest. The power is largest for the affected individuals as compared to its relative except for the affected individuals when its sibling is unaffected.

2.5.3 Additive Model

In an additive genetic model, the homozygote relative risk is $\gamma = 2\beta - 1, \beta > 1$, so that $\phi_2 - \phi_1 = \phi_1 - \phi_0$.

Figure 2.7 illustrates the direction and magnitude of the HWD coefficients for affected individual and their relatives. The HWD coefficient is small for the additive model for all relative pairs as compared to the recessive or dominant models. As a result, the power to detect HWD is also small.

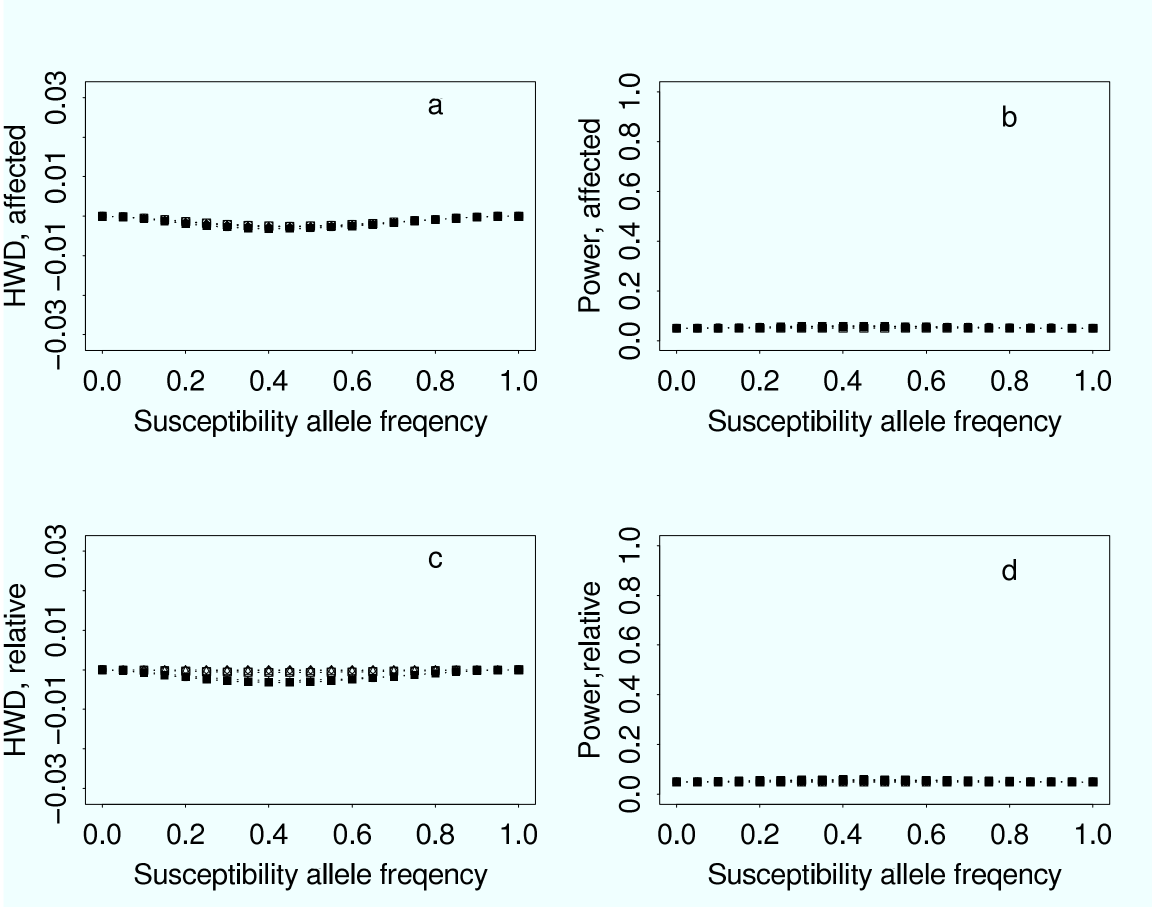


Figure 2.7: HWD coefficients (panels (a) and (c)) and power to detect HWD (panels (b) and (d)) for the additive model.
$K_P = 0.01$, $\gamma = 1.5$, $n = 500$, relative pairs (sibling, parent, grandparent) = (□, ○, △), case-control = ◊. Open/filled symbols for disease status of relative or unrelated individual, unaffected/affected.

2.5.4 Multiplicative Model

In a multiplicative model, the homozygote relative risk is the square of the heterozygote relative risk, i.e. $\phi_2 = \phi_1^2/\phi_0$ or $\gamma = \beta^2, \beta > 1$.

Figure 2.8 illustrates the direction and magnitude of the HWD coefficients for affected individual and their relatives.

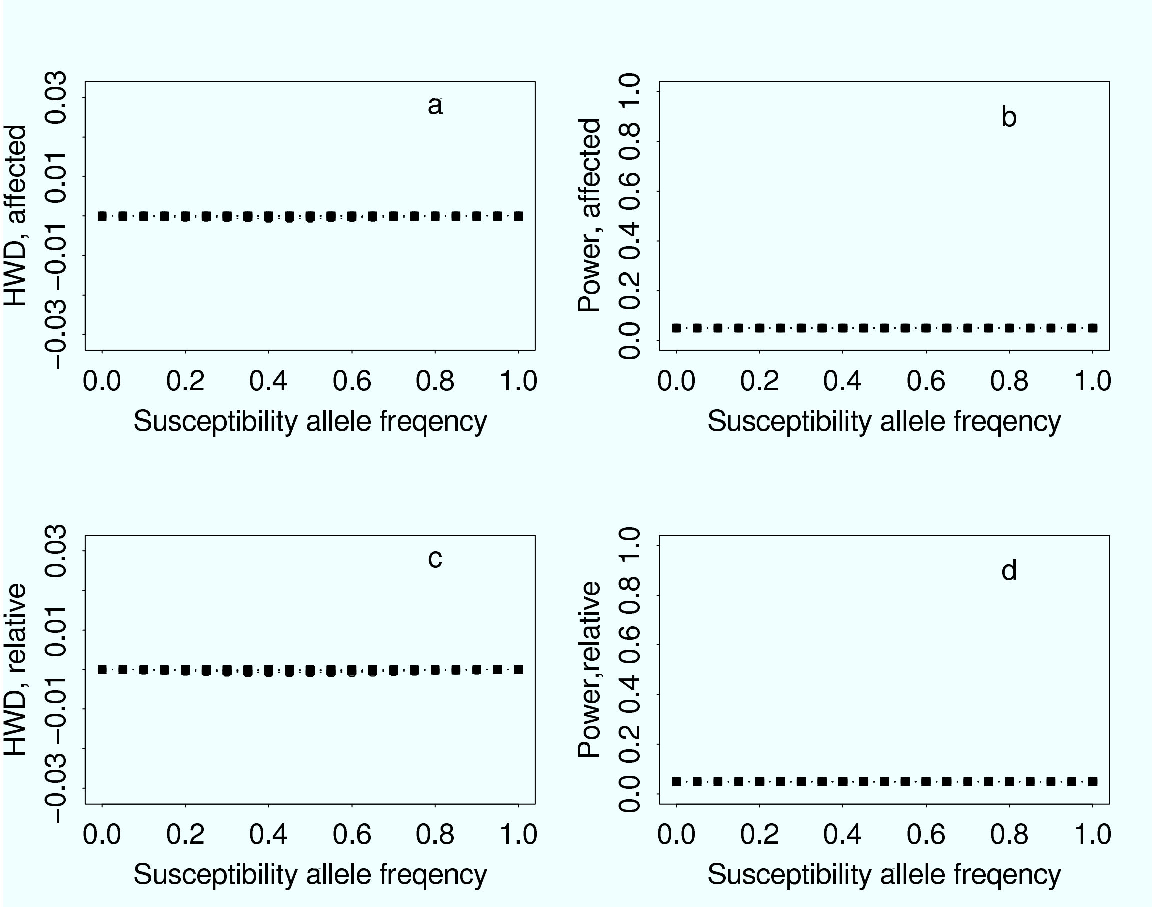


Figure 2.8: HWD coefficients (panels (a) and (c)) and power to detect HWD (panels (b) and (d)) for the multiplicative model.
$K_P = 0.05$, $\gamma = 1.5$, $n = 500$, relative pairs (sibling, parent, grandparent) = (□, ∘, △), case-control = ◊. Open/filled symbols for disease status of relative or unrelated individual, unaffected/affected.

Recall that Wittke-Thompson et al. (2005) showed that the HWD coefficient to be zero for affected individuals for this model. In all cases depicted in Figure 2.8, the HWD coefficient for the affected individual is zero or slightly negative. Its magnitude is largest with an affected parent. For the relative, the largest coefficient occurs for parents who are unaffected. The power to detect HWD is zero when the HWD is zero, and small when the HWD is slightly negative.

2.5.5 Summary Of Results On Relative Pairs

The figures above show that the greatest departure from HWD, and the greatest power to detect departure from HWE is for the dominant and recessive models. These results hold for both the affected individual and its relative, with largest HWD and power occurring when the relative is also affected. Sib pairs gave the largest power, while the affected-parent and affected grandparent pairs give similar but slightly less power. Unaffected sibs give the highest power to detect HWD among the unaffected relatives.

Table 2.5 gives some results regarding the sign of the HWD coefficients and relationships among them for the specific genetic models. For an affected relative pair, the HWD coefficient for both relatives is the same. For the dominant and additive models and the HWD coefficient is always negative for both individuals in an affected child-parent pair. For the additive model, the coefficient is negative for both siblings in an affected sibling pair. For the recessive model, the HWD coefficient is always positive for both relatives in an affected relative pair and is negative for both parent and grandparent when they are unaffected. For the multiplicative model, the coefficient is negative for both siblings in a discordant sibling pair. Other relationships between the HWD coefficients are shown for the additive and multiplicative models. For example, the HWD coefficient for the affected individual when its parent or grandparent is affected is $(K_R)^2(1-K_R)^2$ times its value when the parent or grandparent is unaffected. This multiplier is less than (greater than) one when K_R is less than (greater than) half.

Table 2.5: Some results regarding the HWD coefficients

Affected individual (1) and its relative (2)		
Sibling (S)	Parent (P)	Grandparent (GP)
Dominant, $\beta = \gamma, \gamma > 1$		
$D_{1\mathcal{A}} = D_{2\mathcal{A}}$	$D_{1\mathcal{A}} = D_{2\mathcal{A}} < 0$	$D_{1\mathcal{A}} = D_{2\mathcal{A}}$
Recessive, $\beta = 1$, $\gamma > 1$		
$D_{1\mathcal{A}} = D_{2\mathcal{A}} > 0$	$D_{1\mathcal{A}} = D_{2\mathcal{A}} > 0$ $D_{2\bar{\mathcal{A}}} < 0$	$D_{1\mathcal{A}} = D_{2\mathcal{A}} > 0$ $D_{2\bar{\mathcal{A}}} < 0$
Additive, $\gamma = 2\beta - 1, \beta > 1$		
$D_{1\mathcal{A}} = D_{2\mathcal{A}} < 0$	$D_{1\mathcal{A}} = D_{2\mathcal{A}} < 0$	$D_{1\mathcal{A}} = D_{2\mathcal{A}}$
D_{1A} (S) $= D_{1A}$ (P) $\dfrac{K_R^2}{(1-K_R)^2}$		
Multiplicative, $\gamma = \beta^2, \beta > 1$		
$D_{1\mathcal{A}} = D_{2\mathcal{A}} = 0$ $D_{1\bar{\mathcal{A}}}, D_{2\bar{\mathcal{A}}} < 0$	$D_{1\mathcal{A}} = D_{2\mathcal{A}}$ $D_{1\bar{\mathcal{A}}} = D_{1\mathcal{A}} \dfrac{K_R^2}{(1-K_R)^2}$ $D_{2\bar{\mathcal{A}}} = D_{2\mathcal{A}} \dfrac{K_R^2}{(1-K_R)^2} \Big(1 + \dfrac{1}{\alpha\sqrt{\gamma}}\Big)^2$ $D_{2\bar{\mathcal{A}}} = D_{1\bar{\mathcal{A}}} \Big(1 + \dfrac{1}{\alpha\sqrt{\gamma}}\Big)^2$	$D_{1\mathcal{A}} = D_{2\mathcal{A}}$ $D_{1\bar{\mathcal{A}}} = D_{1\mathcal{A}} \dfrac{K_R^2}{(1-K_R)^2}$
$D_{1\mathcal{A}}$ (GP) $= \dfrac{1}{4} D_{1\mathcal{A}}$ (P) $D_{1\bar{\mathcal{A}}}$ (GP) $= \dfrac{1}{4} D_{1\bar{\mathcal{A}}}$ (P)		

2.6 A Comparison With Li and Leal (2008)

Li and Leal (2008) derived expressions for genotypic probabilities and HWD coefficients at a functional SNP, and at a marker in LD, for parents and for unaffected siblings of affected individuals. They also examined the power of the test for deviation from HWE.

They obtained the genotypic proportions, $P^D_{g_j}$, $g_j = 0, 1, 2$, for the affected individual (pg 105) as

$$P^D_{g_j=0} = \frac{\phi_0 p^2}{K_P},$$

$$P^D_{g_j=1} = \frac{2\phi_1 pq}{K_P}$$

and

$$P^D_{g_j=2} = \frac{\phi_2 q^2}{K_P}.$$

These probabilities correspond to those of an affected person in the general population. However, the appropriate population is not a pool of affected individuals but a subset that consists of affected individuals considered jointly with an unaffected sibling. From the right margin of Table 2.4 the probabilities for the affected individual in discordant sibling pairs are

$$\frac{\phi_0 p^2}{4K_P(1-K_S)}[(1-\phi_0)(1+p)^2 + 2(1-\phi_1)q(1+p) + (1-\phi_2)q^2],$$

$$\frac{\phi_1 pq}{2K_P(1-K_S)}[(1-\phi_0)p(1+p) + 2(1-\phi_1)(1+pq) + (1-\phi_2)q(1+q)]$$

and

$$\frac{\phi_2 q^2}{4K_P(1-K_S)}[(1-\phi_0)p^2 + 2(1-\phi_1)p(1+q) + (1-\phi_2)(1+q)^2];$$

where K_S is given by (2.5). For small values of the disease prevalence and the penetrances, there is not much difference between the expressions in Table 2.4 and those in Li and Leal (2008).

For the unaffected sibling, Li and Leal (2008) obtained the genotypic probabilities

$$P_U(0) = \frac{(1-\phi_0)p^2}{4K_P(1-K_P)}[\phi_0(p+1)^2 + 2\phi_1(1+p)q + \phi_2 q^2],$$

$$P_U(1) = \frac{(1-\phi_1)pq}{2K_P(1-K_P)}[\phi_0 p(p+1) + 2\phi_1(1+pq) + \phi_2 q(1+q)]$$

and

$$P_U(2) = \frac{(1-\phi_2)q^2}{4K_P(1-K_P)}[\phi_0 p^2 + 2\phi_1 p(1+q) + \phi_2(1+q)^2].$$

Once again these probabilities differ from the results in the bottom margin of Table 2.4 which are

$$P_{0,2\bar{A}} = \frac{(1-\phi_0)p^2}{4K_P(1-K_S)}[\phi_0(p+1)^2 + 2\phi_1(1+p)q + \phi_2 q^2],$$

$$P_{1,2\bar{A}} = \frac{(1-\phi_1)pq}{2K_P(1-K_S)}[\phi_0 p(p+1) + 2\phi_1(1+pq) + \phi_2 q(1+q)]$$

and

$$P_{2,2\bar{A}} = \frac{(1-\phi_2)q^2}{4K_P(1-K_S)}[\phi_0 p^2 + 2\phi_1 p(1+q) + \phi_2(1+q)^2].$$

They differ because Li and Leal (2008) do not properly restrict their calculations to the subset of affected individuals with an unaffected sibling.

Note that

$$P_{g_j,2\bar{A}} = \frac{(1-K_P)}{(1-K_S)}P_U(g_j),$$

for $g_j = 0, 1, 2$. For small values of the disease prevalence, K_P, and the recurrence risk, K_S, the ratio in the above expression is almost one. Therefore, not much difference is noticed in the values of the HWD coefficients for the unaffected sibling (Figure 2.10).

Figures 2.9 and 2.10 depict the HWD coefficients (top panels) and the power of rejecting HWE (bottom panels) using the formulae derived by Li and Leal (2008) (left panels) and the expressions obtained earlier in this chapter (right panels). These figures use parameters from Li and Leal (2008) and are for 5000 discordant sibling pairs with relative risks, $\beta = 1$ and $\gamma = 1.5$, and level of significance, $\alpha = 10^{-7}$ for the four specific genetic models. The figures show that there is not much difference in the values of the HWD coefficients or power between the results of Li and Leal (2008) and the correct results.

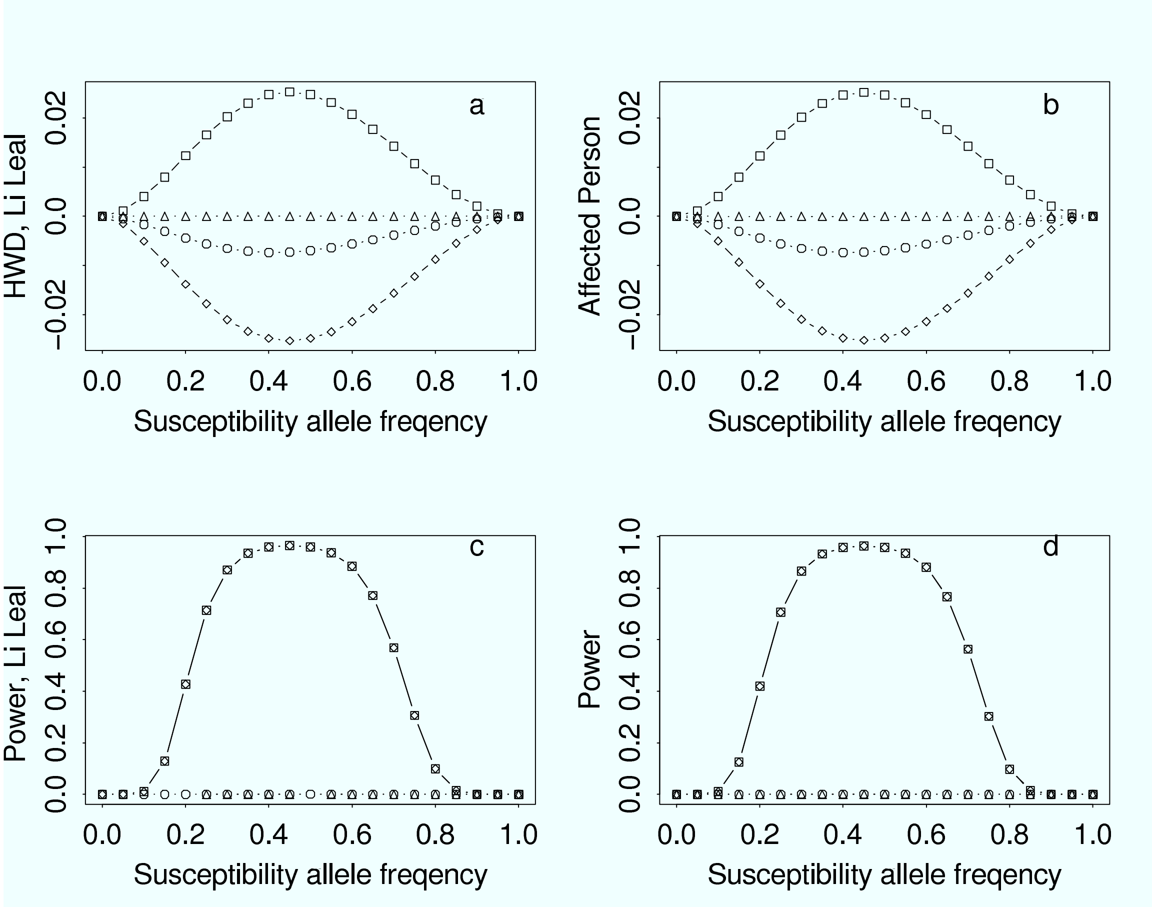


Figure 2.9: Comparing the HWD coefficient and the power of detecting HWD for the affected individual of a discordant sib pair using the result of Li and Leal (2008) (panels (a) and (c)) and the correct values (panels (b) and (d)).
$\alpha = 0.01$, $\gamma = 1.5$. Genetic models: (dominant, recessive, additive, multiplicative) = $(\Diamond, \Box, \circ, \triangle)$

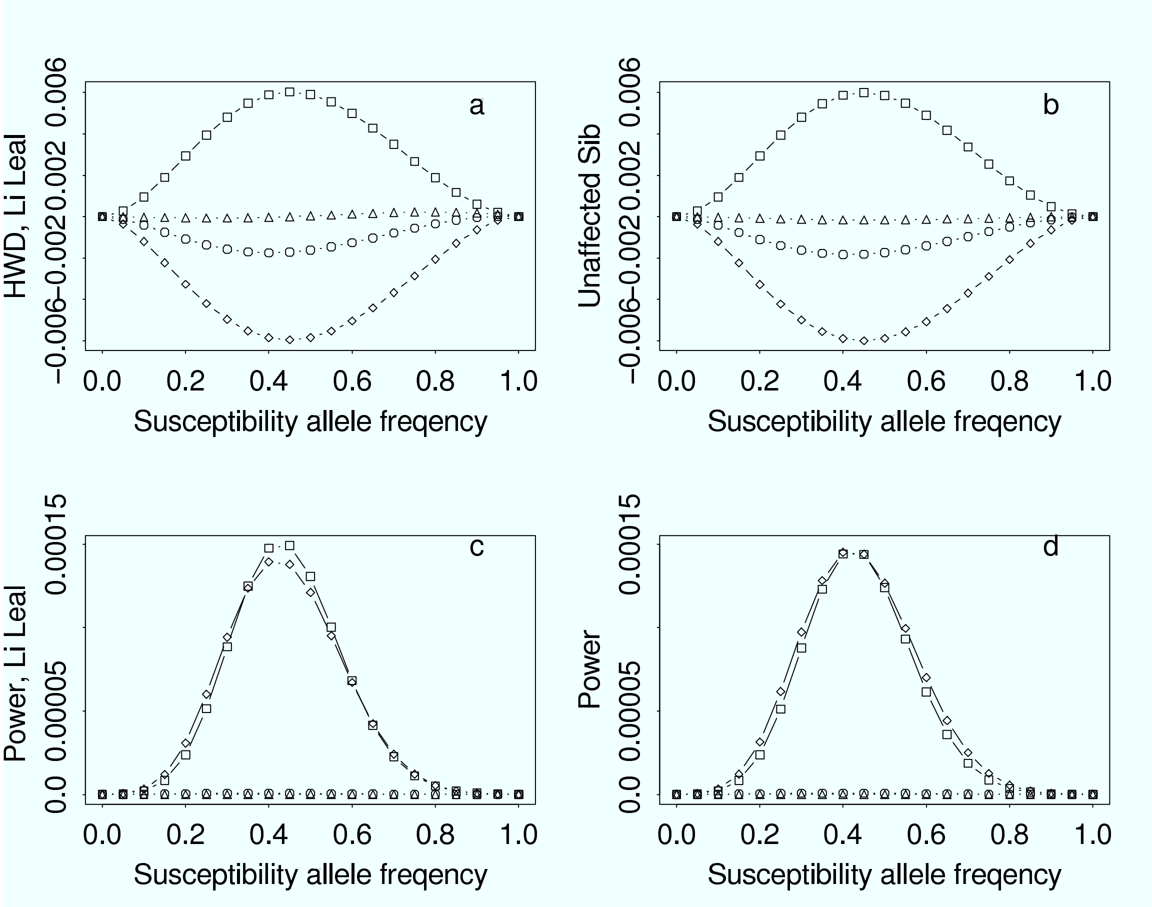


Figure 2.10: Comparing the HWD coefficient and the power of detecting HWD for the unaffected sibling of a discordant sib pair using the result of Li and Leal (2008) (panels (a) and (c)) and the correct values (panels (b) and (d)).
$\alpha = 0.01$, $\gamma = 1.5$. Genetic models: (dominant, recessive, additive, multiplicative) = $(\Diamond, \Box, \circ, \triangle)$

2.7 Discussion

In this chapter the genotypic frequencies and HWD coefficients for an affected individual and its relative were derived. The dominant genetic model shows excessive homozygosity and recessive genetic models shows excessive heterozygosity and these models give the largest power to detect departure from HWE. For the multiplicative model, non-zero HWD coefficients were found in some cases but these values were very small and the power to detect departure from HWE is small even with large samples. The recessive model for the affected individual and its parent displays the greatest deviation from HWE followed by the dominant model for sibling pairs. The power to detect HWD is largest for the recessive and dominant models and is slightly nonzero for multiplicative model. The unrelated affected and unaffected individuals have the smallest deviation from HWE and hence give the least power to detect HWD.

The genotypic frequencies derived in this Chapter are used in Chapter 3 to fit models to data from relative pairs, which can be used to test for association between the locus and the disease.

Chapter 3

Testing Association Using The Heterogeneous Model And Data From Affected Individuals And Their Relatives

3.1 Introduction

In the previous chapter, genotypic probabilities and HWD coefficients of relative pairs were derived under a general model of genetic association. The HWD coefficients of both individuals in a pair were also examined under specific genetic models. In this chapter, the general model is used in testing for association between the disease and a disease susceptibility locus for genotypic data from relative pairs.

The case-control design used often in association studies suffers from several shortcomings. If the case and control groups are not well matched for age, ethnicity, or sex, a case-control study could lead to false positive association. To overcome these problems, Spielman et al. (1993) proposed a transmission disequilibrium test (TDT) which is a family-based test to detect genetic linkage only in the presence of genetic association. The test uses the genetic information on a case-parents trios, an affected child and their parents and it measures the over-transmission of an allele from heterozygous parents to an affected offspring. In essence, TDT is an application of McNemar's test.

For late-onset complex diseases, parental data are not usually available. Some modifications and extensions of the TDT have been proposed involving siblings instead of parents, considering only affected or unaffected siblings or a combination thereof. One of the modifications proposed to overcome this problem is the Sib-TDT (Spielman and Ewens, 1998) that uses discordant sibships. Fengzhu et al. (1999) propose a test, TDT1, to detect linkage and association between a candidate marker locus and a disease locus by using genotypes of case subjects and only one parent. They also propose a method to combine the genotypic information from one or both parents and/or affected or unaffected siblings. Methods have also been proposed to use the genotypic information on case-parents trios and discordant siblings or unrelated control-parents trios combined with case-parents trios (Deng et al., 2002). Another modification is a paired Hotelling's T^2 test statistic that uses unaffected siblings as controls for affected siblings (Fan and Knapp, 2005). The test takes into account the correlation among the markers as well as the correlation within each sibling pair and is based on haplotype/allele coding and genotype coding. A generalization of the TDT, based on a score statistic (Schaid and Jacobsen, 1999) and an extension for multi-allele marker loci (Sham and Curtis, 1995) are given.

More recently Yan et al. (2009) compared several tests for testing disease-candidate gene association for discordant relative pairs based on genotype counts: McNemar's test, the Cochran-Armitage trend test, the maximum efficient robust test, and Bhapkar's test. They found that the power of the tests increase with the distance in relatedness between the affected individual and its relative.

The heterogeneous model for relative pairs is described in Section 3.2 and used for association testing. A partially conditional model for testing of association for discordant relative pairs is discussed in Section 3.3 and a fully conditional model for testing of association for discordant relative pairs is presented in Section 3.4. The three tests of association based on the heterogeneous model are compared to McNemar's test in Section 3.5. An application to an Alzheimer's data set is given in Section 3.6. The results of the chapter are discussed in Section 3.7.

3.2 Testing Association Using The Heterogeneous Model

Suppose that there are $n_{ij,Sd}$ sibling pairs with genotype i for the affected person and genotype j for its sibling, $i, j = 0, 1, 2$, and that their joint genotypic probability is denoted by $P_{ij,Sd}$, where $d = \mathcal{A}$ for an affected sibling pair and $\bar{\mathcal{A}}$ for a discordant sibling pair. The genotypic frequencies, $P_{ij,Sd}$, given in the Chapter 2, with the observed counts, $n_{ij,Sd}$, for the sibling pairs form a multinomial likelihood function

$$L(q, \mathbf{\Phi}) = \prod_{i=0}^{2}\prod_{j=0}^{2} P_{ij,Sd}^{n_{ij,Sd}},$$

where $\mathbf{\Phi} = (\phi_0, \phi_1, \phi_2)$ is the vector containing the penetrances.

The model can be extended to the situation where there are $n_{ij,rd}$ relative pairs with joint genotypic probability denoted by $P_{ij,rd}$ where $r = S$ for sibling pairs, P

for child-parent pairs and G for grandchild-grandparent pairs. Then for a mixture of relative pairs, the genotypic frequencies given in Chapter 2, Appendix A and Appendix B with the observed counts form a multinomial likelihood function

$$L(q, \mathbf{\Phi}) = \prod_{r=\{S,P,G\}} \prod_{d=\{\mathcal{A},\bar{\mathcal{A}}\}} \prod_{i=0}^{2} \prod_{j=0}^{2} P_{ij,rd}^{n_{ij,rd}}.$$

If, in addition, information is available on $n_{ij,Ud}$ unrelated discordant pairs (cases with matched controls) with genotype i, j their data can be included in the likelihood using their joint genotypic probabilities $P_{ij,Ud}$, giving

$$L(q, \mathbf{\Phi}) = \prod_{d=\{\mathcal{A},\bar{\mathcal{A}}\}} \prod_{i=0}^{2} \prod_{j=0}^{2} \prod_{r=\{S,P,G,U\}} P_{ij,rd}^{n_{ij,rd}}.$$

The disease prevalence K_P is assumed to be known from other sources, and the constraint (2.1),

$$K_P = p^2\phi_0 + 2pq\phi_1 + q^2\phi_2$$

allows one of the parameters to be evaluated from the others. For example, the baseline prevalence, ϕ_0 can be written as a function of K_P, ϕ_1, ϕ_2 and q as

$$\phi_0 = \frac{K_P - 2q(1-q)\phi_1 - q^2\phi_2}{(1-q)^2}, \tag{3.1}$$

The remaining parameters in the model are q, ϕ_1, ϕ_2. Once the other parameters have been estimated, ϕ_0 can be obtained from equation (3.1). The parameters can be estimated by maximizing the likelihood function numerically and approximate standard errors can be obtained by evaluating the inverse of the information matrix or by using the nonparametric bootstrap. A large number of random samples of relative pairs are selected without replacement, and the standard deviations are obtained from the distribution of parameter estimates.

The model was coded in S-Plus software and the built-in function "nlmin" was used to obtain the maximum likelihood estimates.

A lack of fit (LOF) test can be used to assess whether the model is appropriate for the data using the goodness of fit statistic. For sibling pairs this statistic is

$$X^2 = \sum_{i=0}^{2} \sum_{j=0}^{2} \frac{(n_{ij,Sd} - n_{sd}\hat{P}_{ij,Sd})^2}{n_{Sd}\hat{P}_{ij,Sd}},$$

where n_{Sd} are the total number of sibling pairs in the dataset. In general, for a mix of affected and unaffected relative pairs, the test statistic is

$$X^2 = \sum_{r=\{S,P,G\}} \sum_{d=\{\mathcal{A},\bar{\mathcal{A}}\}} \sum_{i=0}^{2} \sum_{j=0}^{2} \frac{(n_{ij,rd} - n_{rd}\hat{P}_{ij,rd})^2}{n_{rd}\hat{P}_{ij,rd}},$$

where n_{rd} are the total number of relative pairs in the dataset. The genotypic frequencies here are evaluated at the maximum likelihood estimates from the fitted model. In large samples the X^2 statistic is distributed as a χ^2 with degree of freedom depending on the number of parameters estimated. For example, for the general heterogeneous model, the parameters to be estimated are β, γ and q. The degree of freedom in this case is five. When the sample size and minor allele frequencies are small, the χ^2 approximation to the distribution of X^2 may not be valid. In this case, the parametric bootstrap can be used, where X^2 is compared to the distribution of the values obtained from samples generated from the fitted model.

The test of association between the disease and the disease susceptibility locus for the genotypic data uses the hypotheses

$$\begin{aligned} &H_0\text{:} \quad \beta = \gamma = 1 \\ &H_a\text{:} \quad \beta, \gamma \text{ both not 1.} \end{aligned}$$

The alternative hypothesis requires estimation of three parameters, the two relative risks, β, γ and the minor allele frequency q, and null hypothesis requires estimation of only q. Therefore, the likelihood ratio (LRT) test statistic, Λ_U, is asymptotically distributed as χ^2 with two degrees of freedom. The subscript U indicates that this is an unconditional test in contrast with two tests described later in the chapter.

Insight into the form of genetic effects: recessive, dominant, additive, or multiplicative, can be obtained by imposing constraints on the pentrances and comparing the fit to the general model. In these cases the likelihood ratio test has a χ^2 distribution with one degree of freedom in large samples.

Level and power of the proposed hypothesis tests were investigated by simulation based on an assumed disease prevalence of $K_P = 0.02$. For the simulations, 1,000 replicated data sets were used. To assess the level of the test, three values of the minor allele frequencies q (0.05, 0.1, 0.3) and two different samples sizes n (300, 1000) were used. To assess the power of the test, the samples of size, $n = 300$ and 1000 were generated for two different values of the minor allele frequencies, q (0.1, 0.3), for a recessive model with heterozygote relative risk, $\beta = 1$ and four different values of the homozygous relative risk, γ (1.5, 2, 2.5, 3). The level and power of the test is approximated using the proportion of hypotheses rejected when the data is generated under the null and alternative hypotheses, respectively. The data were generated for different affected and discordant relative pairs. In some cases mixtures involving different relatives and/or different disease status were used. Unrelated cases and controls were also generated for comparison.

To assess the level of the unconditional test for association, data were generated under the hypothesis of no genetic effect (H_0) for various choices of q and sample size and the likelihood ratio test (H_0 *vs.* H_a) was carried out at the 0.05 level of significance. Table 3.1 shows that the proportion of times H_0 is correctly rejected is close to the nominal level. With 1000 simulations, the standard error of the estimated level is 0.0069 when $\alpha = 0.05$. All but two of the estimated levels are within two standard errors.

Table 3.1: Type I errors obtained by simulation for the unconditional test of association for different relative pairs and disease status using $\alpha = 0.05$

Affection Status of Relative	$q = 0.05$		$q = 0.1$		$q = 0.3$	
	$n = 300$	$n = 1000$	$n = 300$	$n = 1000$	$n = 300$	$n = 1000$
	Sibling pair					
$\mathcal{A}$	0.050	0.056	0.058	0.045	0.036	0.042
$\bar{\mathcal{A}}$	0.046	0.054	0.055	0.048	0.055	0.051
	Child-Parent pair					
$\mathcal{A}$	0.040	0.057	0.050	0.045	0.044	0.054
$\bar{\mathcal{A}}$	0.058	0.051	0.063	0.059	0.061	0.046
	Grandchild-Grandparent pair					
$\mathcal{A}$	0.034	0.054	0.049	0.040	0.056	0.054
$\bar{\mathcal{A}}$	0.050	0.058	0.063	0.050	0.041	0.041
	50:50 mix of sibling pair					
$\mathcal{A}, \bar{\mathcal{A}}$	0.044	0.068	0.062	0.052	0.056	0.051
	50:50 mix of child-parent pair					
$\mathcal{A}, \bar{\mathcal{A}}$	0.043	0.045	0.059	0.047	0.053	0.054
	50:50 mix of sibling pair and child-parent pair					
$\mathcal{A}, \mathcal{A}$	0.045	0.062	0.059	0.047	0.051	0.046
$\mathcal{A}, \bar{\mathcal{A}}$	0.047	0.061	0.058	0.051	0.050	0.049
$\bar{\mathcal{A}}, \mathcal{A}$	0.042	0.055	0.058	0.054	0.053	0.050
$\bar{\mathcal{A}}, \bar{\mathcal{A}}$	0.041	0.055	0.059	0.043	0.051	0.057
	Unrelated discordant pair					
	0.039	0.060	0.055	0.044	0.056	0.041

To assess the power of the unconditional test for genetic association under different relative pairs and disease status, data were generated under the alternative hypothesis (H_a) and LRTs were carried out. The simulated power of the tests are summarized in Figure 3.1 for 300 and 1000 relative pairs. Also shown is the power for the unrelated discordant pair of the same size.

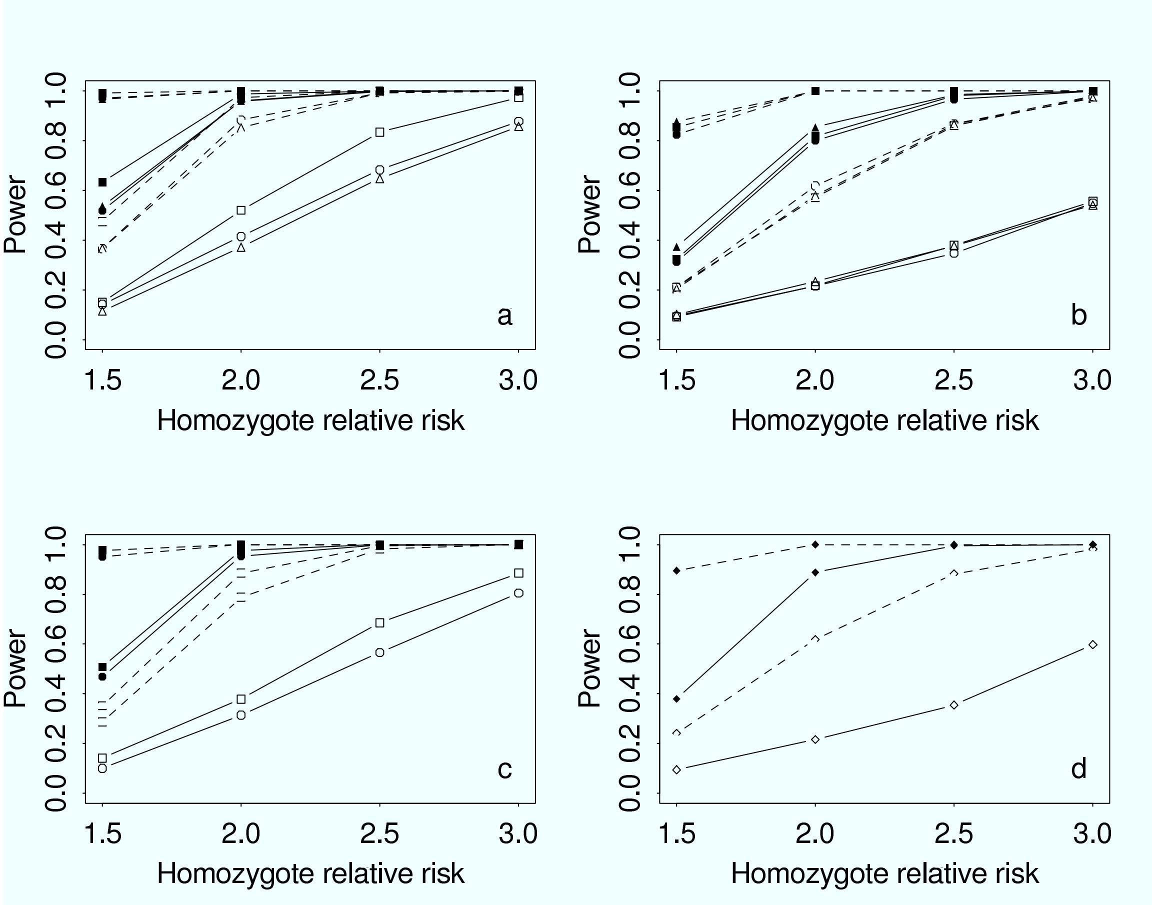


Figure 3.1: Power of the unconditional test for genetic association.
(a) affected relative pairs (b) discordant relative pairs (c) 50:50 mix of affected and discordant relative pairs (d) unrelated discordant pair for $K_P = 0.02$, $\beta = 1$. Relative pairs (sib, parent, grandparent) = ($\square$, $\circ$, $\triangle$), unrelated discordant pair = $\diamondsuit$. Open/filled symbols for $q = 0.1/0.3$. Solid/dashed lines for $n = 300, 1000$.

The power of the unconditional test of association increases with the sample size, n, the minor allele frequency, q and the homozygote relative risk, γ (Figure 3.1). Among the affected relative pairs (Panel (a)), the power of the test for sibling pairs is the largest followed by child-parent pairs and grandchild-grandparent pairs. For the larger allele frequency, the power of the test for all three relative pairs is almost the same. The power of the test is very similar for all three discordant relative pairs (Panel (b)) and for the unrelated discordant pair design (Panel (d)). There is not much difference in the power of the test for the 50:50 mix of affected and discordant sibling pairs and child-parent pairs (Panel (c)). The power for a 50:50 mix of affected and discordant relative pairs lies between that of the affected relative pairs and discordant relative pairs. Among the four panels, the power of the test for the unrelated discordant pair (Panel (d)) is the smallest but is very similar to that for the discordant relative pairs (Panel (b)).

Figure 3.2 illustrates the power for a mix of sibling pairs and child-parent pairs for different combinations of the affection status. Power increases with the sample size, n, the allele frequency, q and the homozygote relative risk, γ. Power of the mix of affected sibling pairs and affected child-parent pairs is the largest and the power of the discordant sibling pair and the discordant child-parent pairs is the smallest. The power of the unconditional test for a 50:50 mix of discordant child-parent pairs and discordant sibling pairs is larger than that of the discordant child-parent pairs and discordant sibling pairs only. The power of the test for a 50:50 mix of affected or discordant child-parent pairs and affected sibling pair lies between the power of the test for affected or discordant child-parent pairs and affected sibling pairs only. For large values of the minor allele frequency, q, the power of the test for a 50:50 mix of affected child-parent pairs and discordant sibling pairs is larger than that of affected child-parent pairs and discordant sibling pairs. However, for small values of q, the power of the test for a 50:50 mix of affected child-parent pairs and discordant sibling pairs lies between the power of the test for affected child-parent pairs and discordant sibling pairs only.

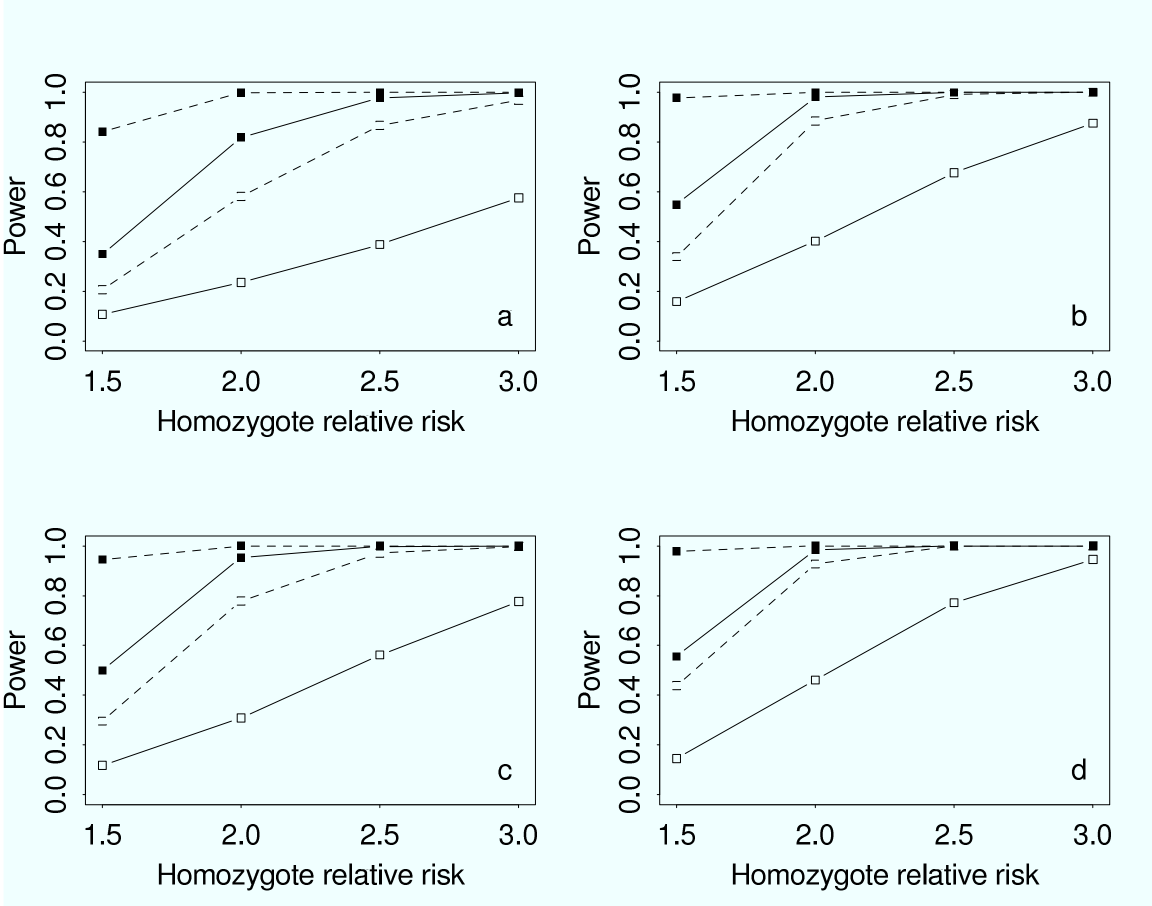


Figure 3.2: Power of the unconditional test for a 50:50 mix of different relative pairs and affection status.
(a) affected sibling pairs and affected child-parent pairs (b) affected sibling pair and discordant child-parent pair (c) discordant sibling pairs and affected child-parent pair (d) discordant sibling pairs and discordant child-parent pairs. For $K_P = 0.02$, $\beta = 1$. Open/filled symbols for $q = 0.1/0.3$. Solid/dashed lines for $n = 300, 1000$.

3.3 Testing Association Using A Partially Conditional Model For Discordant Pairs

Some of the tests of association for discordant pairs, like the McNemar's test, consider only the off-diagonal entries in the 3×3 data matrix for the relative pairs (Table 3.2).

Table 3.2: Data for a relative pair

Affected	**Unaffected Relative**		
Individual	0	1	2
0	n_{00}	n_{01}	n_{02}
1	n_{10}	n_{11}	n_{12}
2	n_{20}	n_{21}	n_{22}

McNemar's test (discussed later in Section 3.5) considers the difference in the sum of the frequencies in the upper triangle (n_{01}, n_{02} and n_{12}) and the sum of frequencies in the lower triangle (n_{10}, n_{20} and n_{21}). If there is no association between the disease and the allele, switching the affected individual and unaffected relative, i.e, switching the rows and columns in the Table 3.2 should not make a difference. Therefore, one can test for association by considering the differences $n_{01} - n_{10}$, $n_{02} - n_{20}$ and $n_{12} - n_{21}$. The entries on the diagonal do not contribute any useful information for assessing association between the disease and the locus because both relatives have the same genotype. The tests of association described above in Section 3.2 involve fitting the unconditional model to the complete table. It is also possible to use this model conditional on the two relatives having different genotypes. Therefore, the joint probabilities for the affected person and its unaffected relative (sibling, parent and grand parent) are derived conditional on their having different genotypes (Tables 3.3, 3.4 and 3.5, respectively). This approach is called partially conditional (PC) to contrast with another approach described later in Section 3.4.

Table 3.3: Joint probability of genotypes for a partially conditional model for a discordant sibling pair

Affected	**Unaffected Sibling**		
Sibling	AA	Aa	aa
AA	-	$\frac{\phi_0(1-\phi_1)p(1+p)}{PCS}$	$\frac{\phi_0(1-\phi_2)pq}{PCS}$
Aa	$\frac{\phi_1(1-\phi_0)p(1+p)}{PCS}$	-	$\frac{\phi_1(1-\phi_2)q(1+q)}{PCS}$
aa	$\frac{\phi_2(1-\phi_0)pq}{PCS}$	$\frac{\phi_2(1-\phi_1)q(1+q)}{PCS}$	-

where

$$PCS = 2[\phi_0 p + \phi_1(1 - pq) + \phi_2 q - \phi_0\phi_1 p(1 + p) - \phi_0\phi_2 pq - \phi_1\phi_2 q(1 + q)].$$

Table 3.4: Joint probability of genotypes for a partially conditional model for a discordant child-parent pair

Affected	**Unaffected Parent**		
Child	AA	Aa	aa
AA	-	$\frac{\phi_0(1-\phi_1)p}{PCP}$	0
Aa	$\frac{\phi_1(1-\phi_0)p}{PCP}$	-	$\frac{\phi_1(1-\phi_2)q}{PCP}$
aa	0	$\frac{\phi_2(1-\phi_1)q}{PCP}$	-

where

$$PCP = 2[\phi_0 p + (1-\phi_1)\phi_2 q].$$

Table 3.5: Joint probability of genotypes for a partially conditional model for a discordant grandchild-grandparent pair

Affected	**Unaffected Grandparent**		
Grandchild	AA	Aa	aa
AA	-	$\frac{\phi_0(1-\phi_1)p(1+2p)}{PCG}$	$\frac{\phi_0(1-\phi_2)pq}{PCG}$
Aa	$\frac{(1-\phi_0)\phi_1 p(1+2p)}{PCG}$	-	$\frac{\phi_1(1-\phi_2)q(1+2q)}{PCG}$
aa	$\frac{(1-\phi_0)\phi_2 pq}{PCG}$	$\frac{(1-\phi_1)\phi_2 q(1+2q)}{PCG}$	-

where

$$PCG = p(2+p)\phi_0 + (3-4pq)\phi_1 + q(2+q)\phi_2 - 2p(1+2p) - 2pq\phi_0\phi_2 - 2q(1+2q)\phi_1\phi_2.$$

Table 3.6: Joint probability of genotypes for a partially conditional model for unrelated discordant pairs

Affected	**Unaffected Individual**		
Individual	AA	Aa	aa
AA	-	$\frac{2\phi_0(1-\phi_1)p^2}{PCU}$	$\frac{\phi_0(1-\phi_2)pq}{PCU}$
Aa	$\frac{2(1-\phi_0)\phi_1 p^2}{PCU}$	-	$\frac{2\phi_1(1-\phi_2)q^2}{PCU}$
aa	$\frac{(1-\phi_0)\phi_2 pq}{PCU}$	$\frac{2(1-\phi_1)\phi_2 q^2}{PCU}$	-

where

$$PCU = \phi_0 p[2(1-\phi_1)p+(1-\phi_2)q]+2\phi_1[(1-\phi_0)p^2+(1-\phi_2)q^2]+\phi_2 q[(1-\phi_0)p+2(1-\phi_1)q].$$

Note that under the null hypothesis of no association, $\phi_0 = \phi_1 = \phi_2$ and the off-diagonal entries, ijth and jith probabilities are equal, for $i \neq j$. The conditional probabilities do not depend explicitly on the disease prevalence, K_P. However, estimation of q, ϕ_0, ϕ_1 and ϕ_2 would imply estimation of K_P which is not possible with a random sample of discordant relative pairs. As before, a value for K_P is assumed known and one parameter is eliminated from estimation.

The LRT is carried out by comparing the maximized likelihoods under the null and alternative hypotheses ($H_0 : \beta = \gamma = 1$, $H_a : \beta, \gamma$ both not 1). The test statistic, Λ_{PC}, is asymptotically distributed as χ^2 with two degrees of freedom in large samples (Appendix C). Simulations were carried out for discordant pairs (sibling, parent, grandparent and unrelated) to investigate the level and power of the partially conditional model.

To assess the level of the partially conditional model for testing association, 1000 data sets were generated under the hypothesis of no genetic effect (H_0) for various choices of q, and sample size, n and the likelihood ratio test was carried out at the 0.05 level of significance. Table 3.7 shows that the proportion of times H_0 is correctly rejected is close to the nominal level. The data were generated using the probabilities from the unconditional model and only off-diagonal cases with differing genotypes were used in estimation of parameters and testing.

Table 3.7: Type I errors obtained by simulation to test association under different discordant pairs for a partially conditional test at $\alpha = 0.05$

	$q = 0.05$		$q = 0.1$		$q = 0.3$	
	$n = 300$	$n = 1000$	$n = 300$	$n = 1000$	$n = 300$	$n = 1000$
sibling	.0320	.0560	.0490	.0550	.0420	.0545
Parent	.0440	.0522	.0515	.0583	.0443	.0460
Grandparent	.0370	.0460	.0650	.0530	.0460	.0510
Unrelated	0.046	0.060	0.073	0.057	0.058	0.057

Note that the test is slightly conservative for small n and q when some of the expected counts are small. This could be because the effective sample size, the number of pairs with different genotypes, used in the partially conditional model is small (Table 3.8).

To assess the power of the partially conditional test for genetic effects for discordant pairs, data were generated under the alternative hypothesis (H_a) and LRTs were carried out. The simulated power of the tests are summarized in Figures 3.3 for sample sizes of 300 and 1000 relative pairs. Also shown is the power for unrelated discordant pair studies of the same size.

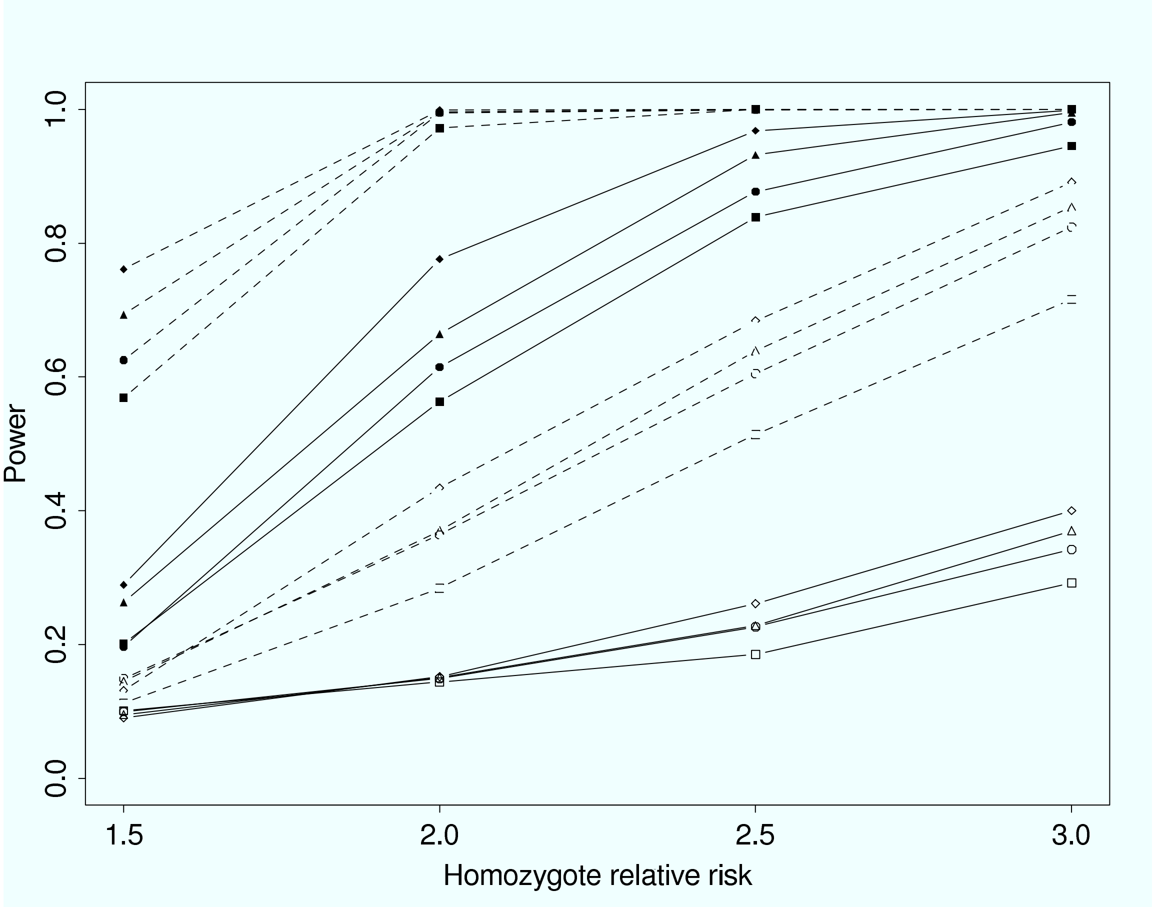


Figure 3.3: Power of the partially conditional test for association using discordant pairs.
For $K_P = 0.02$, $\beta = 1$. Relative pairs (sibling, parent, grandparent) = (□, ∘, △), unrelated discordant pair = ◊. Open/filled symbols for $q = 0.1/0.3$. Solid/dashed lines for $n = 300, 1000$.

Figure 3.3 shows that the power increases with the minor allele frequency, q, homozygote relative risk, γ and the sample size, n. For the larger sample size and minor allele frequency, the power for all discordant relative pairs is similar as is that of unrelated discordant pair. The power of the test increases with the distance in relatedness, i.e., the power is the largest for the unrelated discordant pair followed by the discordant grandchild-grandparent pair. The power is the smallest for discordant sibling pairs. These results agree with those obtained by Yan et al. (2009).

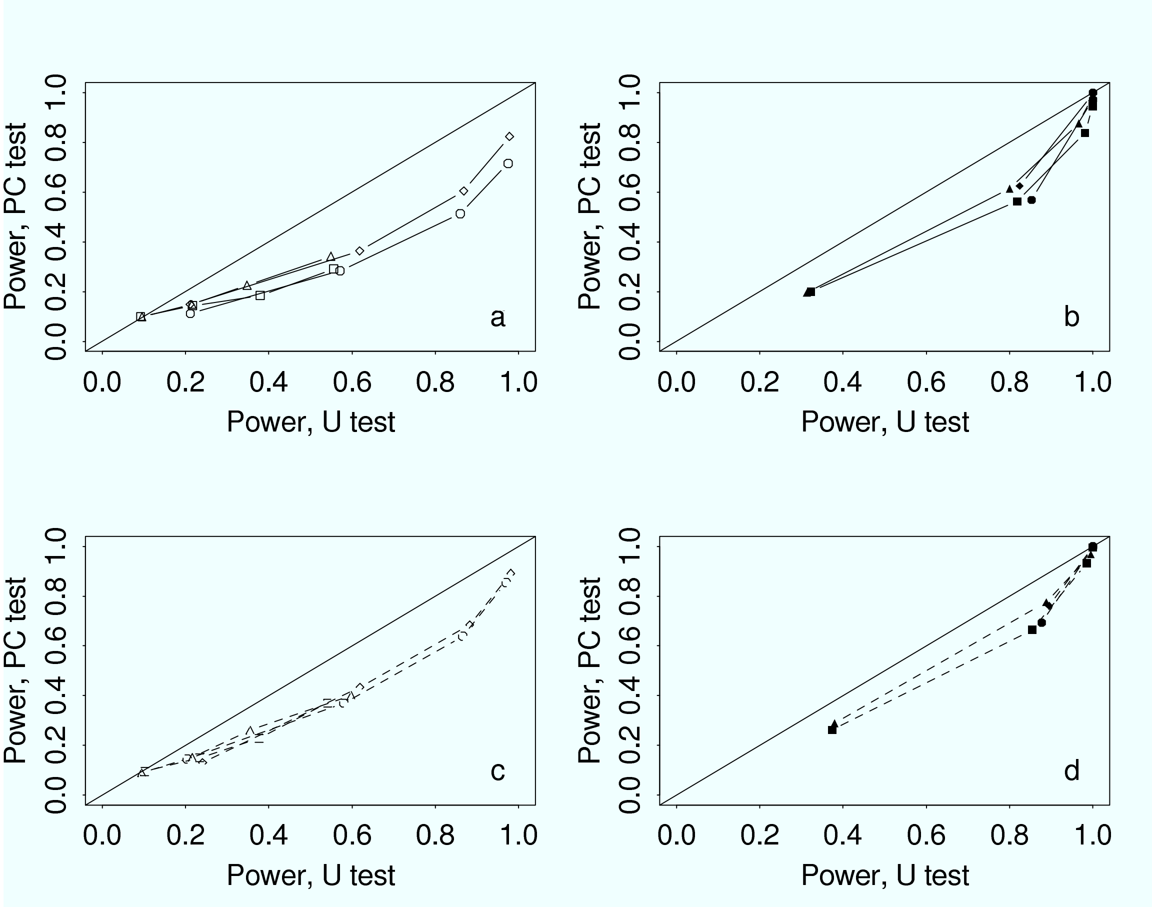


Figure 3.4: Comparison of the power of the unconditional (U) and partially conditional (PC) tests for genetic association for discordant pairs.
For $K_P = 0.02$, $\beta = 1$. Relative pairs (sibling, parent, grandparent) = (□, ∘, △), unrelated discordant pair = ◊, $q = 0.1$, $n = 300$ (a), $n = 1000$ (c); $q = 0.3$, $n = 300$ (b), $n = 1000$ (d).

The results of the unconditional and partially conditional test of association for discordant pairs are similar (Figure 3.4), but the power of the unconditional test is slightly larger than that of the partially conditional test. This may be because the effective sample size for the partially conditional test is smaller than that of the unconditional test (Table 3.8).

Table 3.8: Effective sample size for the partially conditional test for association

	$q = 0.1$		$q = 0.3$	
γ	$n = 300$	$n = 1000$	$n = 300$	$n = 1000$
sibling pair				
1	49	165	105	351
1.5	51	170	109	364
2	52	172	111	372
2.5	53	176	114	380
3	54	178	117	388
Child-parent pair				
1	51	172	115	384
1.5	55	183	130	432
2	56	187	133	443
2.5	57	190	136	452
3	59	193	139	463
Grandchild-grandparent pair				
1	72	241	139	464
1.5	75	249	153	510
2	76	252	157	522
2.5	77	256	160	534
3	78	259	164	544
Unrelated discordant pair				
1	92	308	163	546
1.5	93	312	163	546
2	94	314	164	548
2.5	95	317	164	547
3	95	320	165	550

The effective sample size increases with the overall sample size, n, the minor allele frequency, q and the distance in relatedness. The power of the test in both cases increases with the distance in relatedness.

3.4 Testing Association Using A Fully Conditional Model For Discordant Pairs

If there is no association between the disease and the locus, then the genotypes of each member of a relative pair are irrelevant and could be switched. This suggests conditioning not only on the relatives having different genotypes, but also on the number of such pairs of each type. That is, condition on $n_{01} + n_{10}$, $n_{02} + n_{20}$, and $n_{12} + n_{21}$, the number of pairs with genotypes (AA, Aa), (AA, aa) and (Aa, aa) respectively. The conditional distribution of the counts are binomial for each type, with probabilities as shown in Tables 3.9, 3.10, 3.11 and 3.12.

Table 3.9: Joint probability of genotypes for a fully conditional model for a discordant sibling pair

Affected	**Unaffected Sibling**		
Sibling	AA	Aa	aa
AA	-	$\frac{\phi_0(1-\phi_1)}{\phi_0(1-\phi_1)+\phi_1(1-\phi_0)}$	$\frac{\phi_0(1-\phi_2)}{\phi_0(1-\phi_2)+\phi_2(1-\phi_0)}$
Aa	$\frac{\phi_1(1-\phi_0)}{\phi_0(1-\phi_1)+\phi_1(1-\phi_0)}$	-	$\frac{\phi_1(1-\phi_2)}{\phi_1(1-\phi_2)+\phi_2(1-\phi_1)}$
aa	$\frac{\phi_2(1-\phi_0)}{\phi_0(1-\phi_2)+\phi_2(1-\phi_0)}$	$\frac{\phi_2(1-\phi_1)}{\phi_1(1-\phi_2)+\phi_2(1-\phi_1)}$	-

Table 3.10: Joint probability of genotypes for a fully conditional model for a discordant child-parent pair

Affected	**Unaffected Parent**		
Child	AA	Aa	aa
AA	-	$\frac{\phi_0(1-\phi_1)}{\phi_0(1-\phi_1)+\phi_1(1-\phi_0)}$	0
Aa	$\frac{\phi_1(1-\phi_0)}{\phi_0(1-\phi_1)+\phi_1(1-\phi_0)}$	-	$\frac{\phi_1(1-\phi_2)}{\phi_1(1-\phi_2)+\phi_2(1-\phi_1)}$
aa	0	$\frac{\phi_2(1-\phi_1)}{\phi_1(1-\phi_2)+\phi_2(1-\phi_1)}$	-

Table 3.11: Joint probability of genotypes for a fully conditional model for a discordant grandchild-grandparent pair

Affected	**Unaffected Grandparent**		
Grandchild	AA	Aa	aa
AA	-	$\frac{\phi_0(1-\phi_1)}{\phi_0(1-\phi_1)+\phi_1(1-\phi_0)}$	$\frac{\phi_0(1-\phi_2)}{\phi_0(1-\phi_2)+\phi_2(1-\phi_0)}$
Aa	$\frac{\phi_1(1-\phi_0)}{\phi_0(1-\phi_1)+\phi_1(1-\phi_0)}$	-	$\frac{\phi_1(1-\phi_2)}{\phi_1(1-\phi_2)+\phi_2(1-\phi_1)}$
aa	$\frac{\phi_2(1-\phi_0)}{\phi_0(1-\phi_2)+\phi_2(1-\phi_0)}$	$\frac{\phi_2(1-\phi_1)}{\phi_1(1-\phi_2)+\phi_2(1-\phi_1)}$	-

Table 3.12: Joint probability of genotypes for a fully conditional model for an unrelated discordant pair

Affected	**Unaffected Individual**		
Individual	AA	Aa	aa
AA	-	$\frac{\phi_0(1-\phi_1)}{\phi_0(1-\phi_1)+\phi_1(1-\phi_0)}$	$\frac{\phi_0(1-\phi_2)}{\phi_0(1-\phi_2)+\phi_2(1-\phi_0)}$
Aa	$\frac{\phi_1(1-\phi_0)}{\phi_0(1-\phi_1)+\phi_1(1-\phi_0)}$	-	$\frac{\phi_1(1-\phi_2)}{\phi_1(1-\phi_2)+\phi_2(1-\phi_1)}$
aa	$\frac{\phi_2(1-\phi_0)}{\phi_0(1-\phi_2)+\phi_2(1-\phi_0)}$	$\frac{\phi_2(1-\phi_1)}{\phi_1(1-\phi_2)+\phi_2(1-\phi_1)}$	-

Note that these probabilities depend only on the three penetrances and not on the disease prevalence or allele frequency. In addition, the entries in the four tables are the same except for the $(AA,\ aa)$ genotype for child-parent pairs, which is impossible.

Under the null hypothesis of no association, the genotypic probabilities given in Chapter 2 depend on the allele frequency, q. The sufficient statistics for q under H_0 are n_{00}, n_{11}, n_{22}, $n_{01}+n_{10}$, $n_{02}+n_{20}$ and $n_{12}+n_{21}$ and conditioning on these statistics gives the tables above, which depend only on the penetrances. This approach to estimating the nuisance parameter q is analogous to Fisher's exact test in 2×2 contingency tables.

Although the conditional probabilities depend on the three penetrances, it is not possible to estimate them uniquely. A sample of discordant pairs is not a random sample from a population and so does not allow estimation of the three probabilities of disease given the genotypes. It is possible to write the probabilities in terms of two odds ratios and these can be estimated using a sample of discordant pairs. Denote by P_{10} the conditional probability that the genotype of the affected individual is Aa given that the two genotypes of the discordant individuals are AA and Aa. Then

$$P_{10} = \frac{\phi_1(1-\phi_0)}{\phi_0(1-\phi_1)+\phi_1(1-\phi_0)},$$

and dividing the numerator and denominator by $\phi_0(1-\phi_1)$ gives

$$P_{10} = \frac{\theta_1}{\theta_0+\theta_1} = \frac{\rho_1}{1+\rho_1}$$

where $\theta_i = \dfrac{\phi_i}{1-\phi_i}$ is the odds and $\rho_i = \dfrac{\theta_i}{\theta_0}$ is the odds ratio. Similarly,

$$P_{21} = \frac{\theta_2}{\theta_1+\theta_2} = \frac{\rho_2}{\rho_1+\rho_2}$$

and for sibling pairs and grandchild-grandparent pairs,

$$P_{20} = \frac{\theta_2}{\theta_0+\theta_2} = \frac{\rho_2}{1+\rho_2}.$$

In these expressions the subscripts on the conditional probability, P, refer to the number of variant alleles in the genotype of the affected individual and their relative. The odds ratios can be estimated using the likelihood

$$L(\rho_1,\rho_2) = \prod_{i=0}^{2}\prod_{j=0}^{2} P_{ij}^{n_{ij}},$$

where $P_{ij} = 1 - P_{ji}$. To enforce the constraints that $\rho_i \geq 0$, it is convenient to reparametrize the likelihood in terms of the log odds ratios, $\delta_i = log(\rho_i)$. Under the null hypothesis of no association, the odds ratios are 1, $\delta_i = 0$, and the conditional probabilities are $P_{ij} = \dfrac{1}{2}$. The fully conditional LRT statistic for association, Λ_{FC}, is twice the difference in log likelihood under the null and alternative hypotheses, and is asymptotically distributed as χ^2 with two degrees of freedom (Appendix D).

For the child-parent pairs the genotype pair $(AA,\ aa)$ is impossible and the log likelihood function is

$$\ell = log(L(\rho_1, \rho_2)) = (n_{10} + n_{12})log(\rho_1) + n_{21}log(\rho_2) - (n_{10} + n_{01})log(1 + \rho_1) \\ - (n_{12} + n_{21})log(\rho_1 + \rho_2),$$

and the maximum likelihood estimates are

$$\hat{\rho}_1 = \frac{n_{01}n_{12} - n_{10}n_{21}}{n_{10}n_{12} - n_{01}n_{21}}$$

and

$$\hat{\rho}_2 = \frac{n_{21}}{n_{12}}\hat{\rho}_1.$$

Insight into the form of genetic effects: dominant, recessive, additive or multiplicative can be obtained by imposing constraints on the pentrances and comparing the fit to the general model. In these cases the likelihood ratio test has a χ^2 distribution with one degree of freedom.

Simulations were carried out for discordant pairs to investigate the level and power of the fully conditional test. To assess the level, 1000 replicated data sets were generated under the hypothesis of no genetic effect (H_0) for various choices of allele frequency q, and sample size n, and the likelihood ratio test was carried out at the 0.05 level of significance. Table 3.13 shows that the proportion of times H_0 is correctly rejected is close to the nominal level. The data were generated using the probabilities from the unconditional model and only off-diagonal cases with differing genotypes were used in estimation of parameters.

Table 3.13: Type I errors obtained by simulation for the fully conditional test and different discordant pairs at $\alpha = 0.05$

	$q = 0.05$		$q = 0.1$		$q = 0.3$	
	$n = 300$	$n = 1000$	$n = 300$	$n = 1000$	$n = 300$	$n = 1000$
Sibling	0.0288	0.0583	0.0585	0.0402	0.0522	0.0541
Parent	0.0420	0.0590	0.0660	0.0550	0.0510	0.0610
Grandparent	0.0418	0.0626	0.0734	0.0660	0.0522	0.0500
Unrelated	0.036	0.066	0.057	0.050	0.053	0.048

To assess the power of the fully conditional test, data were generated under the alternative hypothesis (H_a) and the LRT was carried out. The simulated power of the tests are summarized in Figures 3.5 for sample sizes of 300 and 1000 relative pairs. Also shown is the power for a study of the same size involving unrelated discordant pairs.

Figure 3.5 shows that the power increases with the minor allele frequency, q, homozygote relative risk, γ, and the sample size, n.

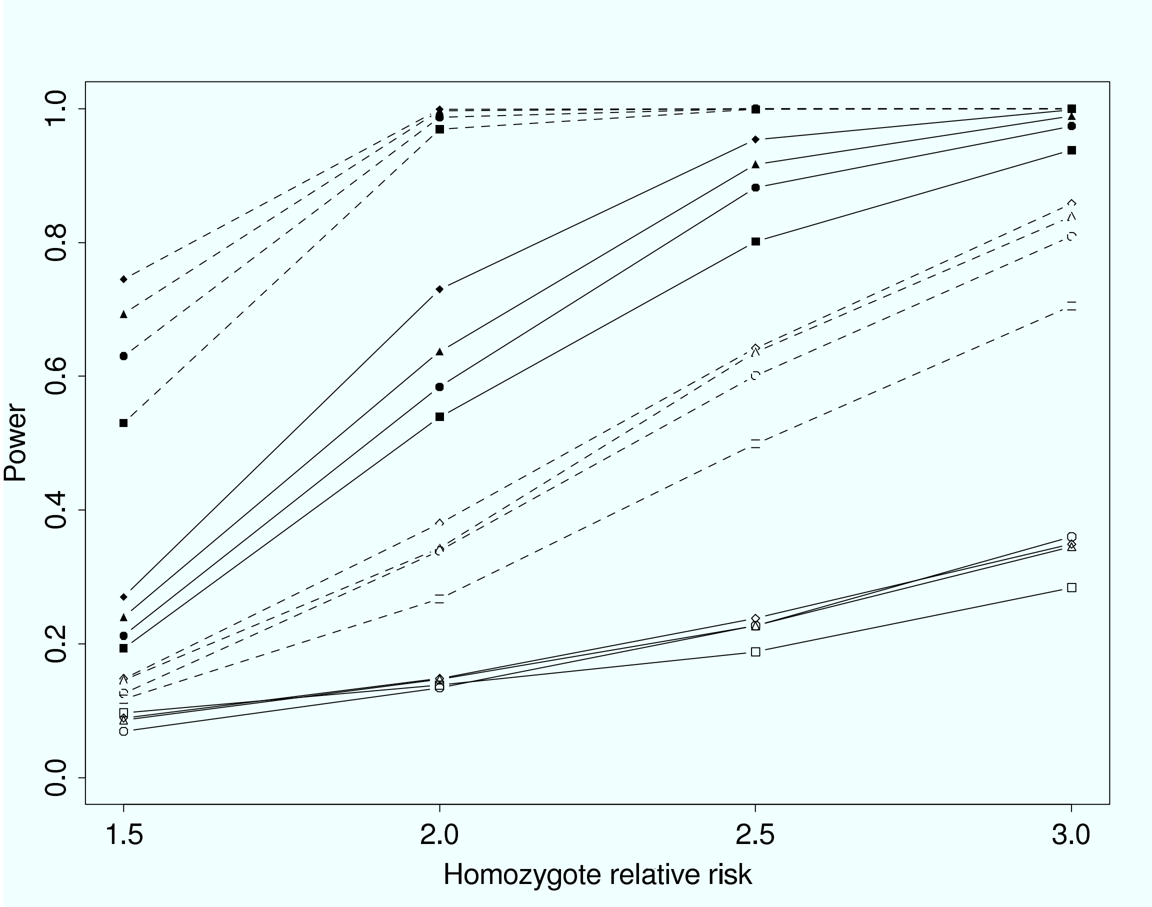


Figure 3.5: Power of the fully conditional test for discordant relative pairs. For $K_P = 0.02$, $\beta = 1$. Relative pairs (sibling, parent, grandparent) = (□, ∘, △), unrelated discordant pairs = ◊. Open/filled symbols for $q = 0.1/0.3$. Solid/dashed lines for $n = 300, 1000$.

For the larger sample size and minor allele frequency, the power for all discordant relative pairs is similar as is that of the unrelated discordant pairs. The power of the test increases with the distance in relatedness, i.e., the power is the largest for the unrelated discordant pairs followed by the discordant grandchild-grandparent pairs. The power is the smallest for discordant sibling pairs. These results agree with those obtained by Yan et al. (2009).

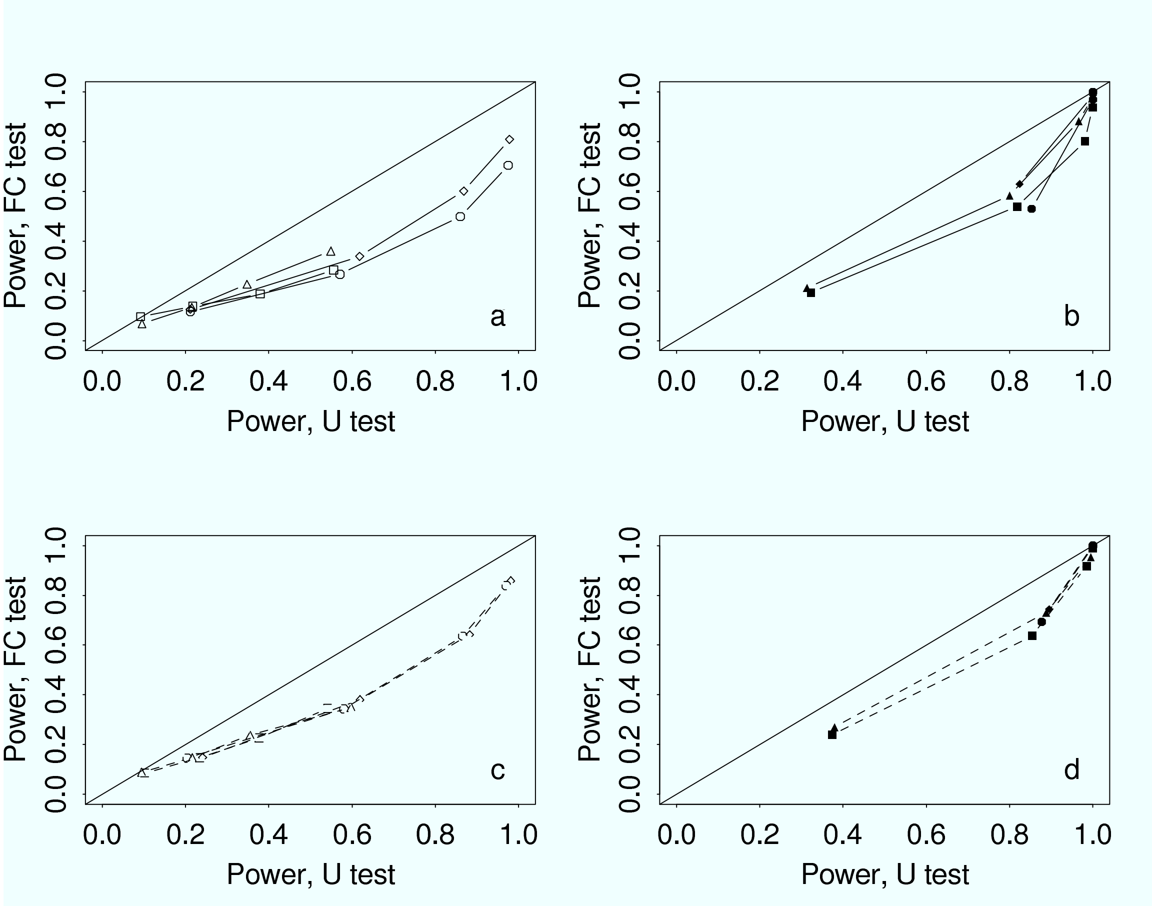


Figure 3.6: Comparison of the power of the unconditional (U) and fully conditional (FC) LR tests for genetic association for discordant pairs.
For $K_P = 0.02$, $\beta = 1$. Relative pairs (sibling, parent, grandparent) = ($\Box$, $\circ$, $\triangle$), unrelated discordant pairs = $\Diamond$, $q = 0.1$, $n = 300$ (a), $n = 1000$ (c); $q = 0.3$, $n = 300$ (b), $n = 1000$ (d).

The results of the unconditional and fully conditional test for testing of association for discordant pairs are similar (Figure 3.6), but the power of the unconditional test is slightly larger than that of the fully conditional test. This may be due to the larger effective sample size, and the extra assumption that the prevalence of disease, K_P, is known. The power of the partially and fully conditional tests are the same (Figure 3.7).

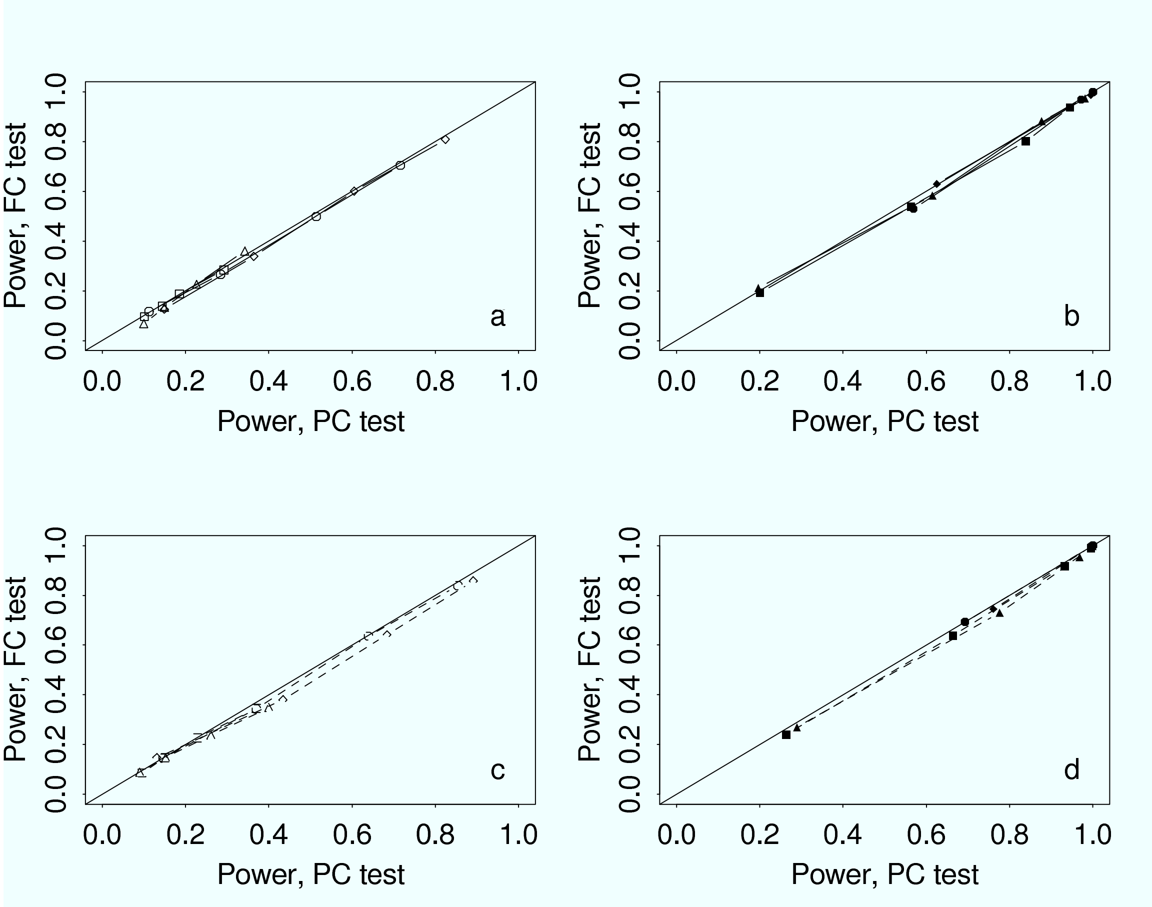


Figure 3.7: Comparison of the power of the partially (PC) and fully conditional (FC) tests for genetic association for discordant pairs.
For $K_P = 0.02$, $\beta = 1$. Relative pairs (sibling, parent, grandparent) = (□, ○, △), unrelated discordant pairs = ◇, $q = 0.1$, $n = 300$ (a), $n = 1000$ (c); $q = 0.3$, $n = 300$ (b), $n = 1000$ (d).

The effective sample sizes for the fully conditional test are shown in Table 3.14, and are similar to those for the partially conditional test (Table 3.8).

Table 3.14: Effective sample sizes for the fully conditional test of association

	$q = 0.1$		$q = 0.3$	
γ	$n = 300$	$n = 1000$	$n = 300$	$n = 1000$
sibling pair				
1	50	167	105	353
1.5	51	170	108	363
2	51	173	111	372
2.5	52	175	114	379
3	53	178	116	388
Child-parent pair				
1	53	180	125	419
1.5	55	183	129	431
2	55	187	132	443
2.5	57	190	136	453
3	58	194	138	462
Grandchild-grandparent pair				
1	74	245	149	497
1.5	74	248	153	510
2	75	252	156	522
2.5	76	255	159	533
3	77	258	162	544
Unrelated discordant pair				
1	93	311	172	575
1.5	94	314	176	589
2	95	318	180	602
2.5	96	321	184	614
3	96	324	187	625

3.5 Comparison Of Results With Those Of Yan et al. (2009)

Yan et al. (2009) discussed different tests of association using data on discordant relative pairs for full-sib, half-sib and first-cousin pairs: McNemar's test, the matched Cochran-Armitage trend tests, the matched maximum efficient robust test and Bhapkar's test. They obtained the joint genotypic frequencies for relative pairs and the expressions for the full sibling pair are identical to those given in Table 2.2.

The genotypic data for relatives can be summarized in a 3 × 3 table as in Table 3.2. McNemar's test statistic is

$$T_M = \frac{(n_U - n_L)^2}{n_U + n_L}$$

where

$$n_U = n_{12} + n_{02} + n_{01}$$

and

$$n_L = n_{21} + n_{20} + n_{10}.$$

In order to compare the tests proposed in this chapter with McNemar's test, simulations were carried out assuming two different values of the minor allele frequency, $q = 0.2, 0.4$, sample size, $n = 200$, the prevalence of disease $K_P = 0.1$ under the null model ($\beta = 1, \gamma = 1$) for the level and under the dominant model ($\beta = 2, \gamma = 2$), recessive model ($\beta = 1, \gamma = 2$), and additive model ($\beta = 1.5, \gamma = 2$) for the power. These were the values used in Yan et al. (2009). For the simulations, 1,000 replicated data sets were used. Data were generated for different relative pairs and for unrelated discordant pairs. The results are summarized in Table 3.15. The table shows that the level of all the tests are comparable and close to the nominal value of 0.05.

Table 3.15: Level and power of tests for association for discordant pairs using the unconditional (U), partially (PC) and fully (FC) conditional LRTs and McNemar's test (M), $K_P = 0.1$, $n = 200$

	Level		Power					
			Dominant model		Recessive model		Additive model	
	$\beta = \gamma = 1$		$\beta = \gamma = 2$		$\beta = 1$, $\gamma = 2$		$\beta = 1.5$, $\gamma = 2$	
q	0.2	0.4	0.2	0.4	0.2	0.4	0.2	0.4
Sibling pair								
U	0.056	0.041	0.806	0.813	0.386	0.764	0.336	0.416
PC	0.055	0.060	0.685	0.638	0.246	0.597	0.371	0.400
FC	0.060	0.058	0.236	0.597	0.364	0.396	0.697	0.645
M	0.050	0.050	0.702	0.501	0.110	0.409	0.411	0.463
Child-Parent pair								
U	0.048	0.055	0.829	0.812	0.426	0.813	0.366	0.444
PC	0.074	0.046	0.749	0.682	0.306	0.665	0.344	0.367
FC	0.068	0.055	0.303	0.677	0.360	0.404	0.716	0.660
M	0.050	0.050	0.684	0.500	0.132	0.456	0.402	0.483
Grandchild-Grandparent pair								
U	0.055	0.045	0.898	0.894	0.434	0.841	0.501	0.584
PC	0.051	0.058	0.861	0.824	0.336	0.801	0.495	0.554
FC	0.050	0.043	0.334	0.756	0.489	0.543	0.869	0.804
M	0.050	0.050	0.871	0.676	0.130	0.544	0.563	0.623
Unrelated discordant pair								
U	0.052	0.054	0.956	0.920	0.494	0.896	0.630	0.673
PC	0.055	0.056	0.382	0.846	0.608	0.720	0.940	0.884
FC	0.050	0.043	0.333	0.822	0.604	0.692	0.942	0.904
M	0.050	0.050	0.956	0.810	0.129	0.623	0.683	0.739

In general the power is higher for $q = 0.4$ than for $q = 0.2$. Power is largest for the dominant model, and smaller for the recessive and additive models, for the unconditional, partially conditional and McNemar's tests. The fully conditional test, however, has highest power for the additive model. For each test, power increases with the distance in relatedness as noted by Yan et al. (2009). McNemar's test has low power for the recessive model, in particular when $q = 0.2$. The fully conditional test has low power for the dominant model when $q = 0.2$.

In making these comparisons, one should keep in mind that the unconditional and partially conditional tests require the assumption that the prevalence of disease and minor allele frequency are known and that the effective sample size for the partially conditional, fully conditional and McNemar's test is substantially smaller than the overall sample size.

3.6 Application To A Data Set

The proposed methods of testing for association based on the heterogeneous models were applied to a data set involving 112 discordant sibling pairs from 100 unrelated families ascertained for the presence of one or more individuals with Alzheimer's disease and typed for the ApoE polymorphism (Boehnke and Langefeld, 1998). Table 5 in Boehnke and Langefeld (1998) gives the joint genotypic counts for the affected and unaffected sibling at three alleles, ε_2, ε_3, and ε_4, at the ApoE gene. The allele ε_4 is the high risk allele and ε_2 is the low risk allele for Alzheimer's disease. The data for the alleles ε_2 and ε_3 were combined to give the joint genotypic data in Table 3.16.

Table 3.16: ApoE genotypes for Alzheimer's disease

	Unaffected Sibling			
Affected Sibling	$\bar{\varepsilon}_4\bar{\varepsilon}_4$	$\varepsilon_4\bar{\varepsilon}_4$	$\varepsilon_4\varepsilon_4$	Total
$\bar{\varepsilon}_4\bar{\varepsilon}_4$	23	4	0	27
$\varepsilon_4\bar{\varepsilon}_4$	25	36	2	63
$\varepsilon_4\varepsilon_4$	8	8	6	22
Total	56	48	8	112

The unconditional, partially and fully conditional tests and the McNemar's test were applied to the data set. The prevalence of Alzheimer's disease, K_P is assumed to be 0.06 (Wittke-Thompson et al., 2005) for the unconditional and partially conditional tests. The results are summarized in Table 3.17. Exact p-values for the McNemar's and fully conditional test were obtained by enumerating all the possible tables with the totals $n_{01}+n_{10} = 29$, $n_{02}+n_{20} = 8$ and $n_{12}+n_{21} = 10$, resulting in 2970 different possibilities. Approximate p-values for the unconditional and partially conditional tests were obtained from the χ^2_2 distribution. Confidence intervals were obtained using the parametric bootstrap. The estimate of the recurrence risk for a siblings is $\hat{K}_S = 0.11$, the heterozygote relative risk is $\hat{\beta} = 6$ and homozygote relative risk is $\hat{\gamma}$ $= 24$, using either the unconditional or partially conditional model.

Table 3.17: Parameter estimates and tests for the Alzheimer's data set using the unconditional (U), partially (PC) and fully (FC) conditional LRTs and McNemar's test (M) for the ApoE data, $K_P = 0.06$

	Estimates, (95% quantile interval)	Test
U	$\hat{q} = 0.1720$ (0.1288,0.2348), $\hat{\phi}_1 = 0.1179$ (0.0863,0.1589), $\hat{\phi}_2 = 0.4407$ (0.2018,0.7774). $\hat{\phi}_0 = 0.0195$ (0.013,0.0274), $\hat{\beta} = 6.0319$ (3.4252,10.6457), $\hat{\gamma} = 22.5535$ (8.2075,47.6059)	$\Lambda_U = 36.6129$, $p = 1.12 \times 10^{-8}$
PC	$\hat{q} = 0.1572$ (0.0871, 0.2435), $\hat{\phi}_1 = 0.1239$ (0.0882, 0.1839), $\hat{\phi}_2 = 0.5066$ (0.2170, 1). $\hat{\phi}_0 = 0.0206$ (0.0049, 0.0382), $\hat{\beta} = 6.0046$ (2.7600, 27.3502) $\hat{\gamma} = 24.5541$ (9.1893, 135.4449)	$\Lambda_{PC} = 34.3583$, $p = 3.46 \times 10^{-8}$
FC	$\hat{\delta}_1 = 1.9057$ (1.0475, 3.4444), $\hat{\delta}_2 = 3.4531$ (2.1030, 26.2612). $\hat{\rho}_1 = 6.7239$ (2.8507, 31.3240), $\hat{\rho}_2 = 31.5979$ (8.1910, 2.54 $\times 10^{11}$)	$\Lambda_{FC} = 31.3219$ $p = 2.23 \times 10^{-7}$
M	-	$T_M = 26.0638$ $p = 1.77 \times 10^{-7}$

The results suggest that the individuals with two ε_4 variant alleles have about a 24-fold increased risk and those with only one copy of the variant allele have a six-old increased risk of getting Alzheimer's disease compared to those with no copies of the variant allele. Recall that for a relatively uncommon disease, the odds ratios are nearly the same as relative risks, so the three tests based on the heterogeneous model give very similar results. While these tests give the same small p-value as the McNemar's test, they give added insight into the mode of inheritance. The estimated penetrances indicate that dominant and recessive models can be ruled out, but additive and multiplicative models are possible. The conclusion of a strong association of the allele ε_4 with Alzheimer's disease coincides with those of Boehnke and Langefeld (1998) and Yan et al. (2009).

For the exact test of the Alzheimer's data, out of the 2970 possible tables, there are 360 cases considered considered more extreme than the observed table using the FC test statistic, Λ_{FC}, whereas for the McNemar's test statistic, T_M, there are only 167 cases considered more extreme than the observed table. Of these, 108 tables were found to be extreme by both tests whereas there were 59 tables that were more extreme than the observed by T_M but not by Λ_{FC} and 252 were found to be more extreme by Λ_{FC} and not by T_M.

Figure 3.8 shows the relationship between the McNemar's and FC test statistics for each of the 2970 possible tables. The horizontal and vertical dotted lines on the graph correspond to the 0.95 quantiles of χ_2^2 (5.9915) and χ_1^2 (3.8415) respectively, indicating the asymptotic rejection regions for the tests. The points O, B, M, C and

N in Figure 3.8 are examples of tables that are found to be more extreme than the observed table (O) by both the FC and McNemar's tests (B), only by McNemar's test (M), only by the FC test (C), and a table that was not extreme by either test (N). These tables are presented in Table 3.18.

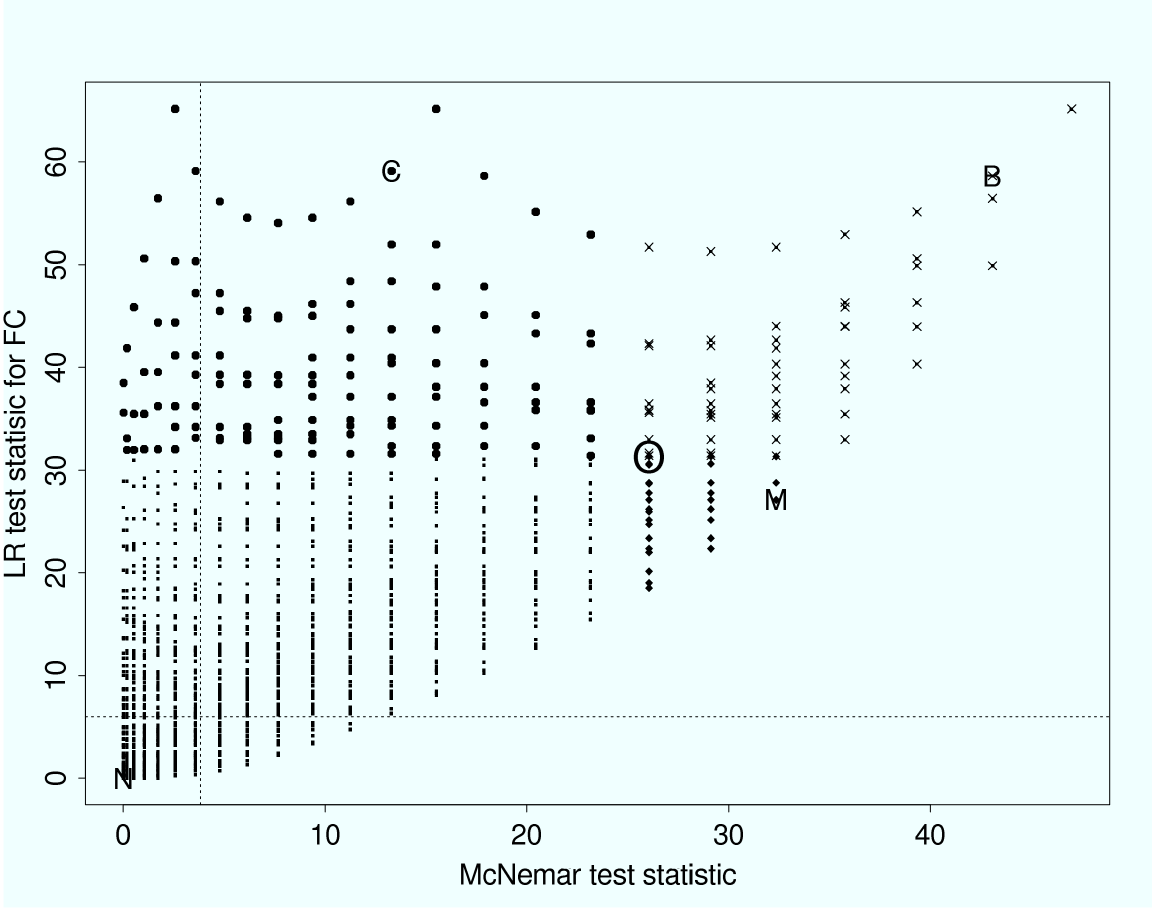


Figure 3.8: Comparison of the fully conditional and McNemar's test statistics. Tables more extreme or as extreme than the observed table (O) by both FC and McNemar's tests (×), by FC only (●), by McNemar's only (♦); tables less extreme than the observed table (.).

The vertical lines in Figure 3.8 shows that McNemar's test is more discrete than the LR test. There are only seven values for T_M as extreme or more extreme than the observed. The 0.05 exact critical value of T_M is 3.5957 and of Λ_{FC} is 6.2479.

Table 3.18: Five tables with their McNemar's (T_M) and fully conditional (Λ_{FC}) test statistics

Point	Table			T_M	Λ_{FC}
O	-	4	0	26.0638	31.3219
	25	-	2		
	8	8	-		
B	-	0	0	43.0851	58.6542
	29	-	1		
	8	9	-		
M	-	0	4	32.3617	27.1013
	29	-	0		
	4	10	-		
C	-	0	1	13.2979	59.1275
	29	-	10		
	7	0	-		
N	-	14	4	0.0213	0.0299
	15	-	5		
	4	5	-		

The tables, like N, that have fairly symmetric counts below and above the diagonal do not lead to the rejection of the null hypothesis of no association by either test. The tables that do lead to rejection by both tests, like B, have nearly all counts either below or above the diagonal. Table C shows strong association in two of the three genotypic categories, but the opposite in the third. McNemar's test statistic is relatively small for this table but the maximum likelihood estimates for the odds ratio are large as is the test statistic Λ_{FC}. There are some tables where Λ_{FC} is larger than the observed table but McNemar's test is not significant. All the tables, like M, which give more extreme values for T_M but not for Λ_{FC} give fairly large values for

Λ_{FC} which lead to rejection of the hypothesis of no association.

Profile log likelihood and contour plots for δ_1 and δ_2 in the fully conditional model for the Alzheimer's data are shown in Figure 3.9 and Figure 3.10. The profile likelihoods are fairly symmetric in the log odds ratios. The approximate 95% confidence interval for δ_1 is (0.960, 3.140) and δ_2 is (1.89, 5.529). The interval for δ_1 is similar to the 95% parametric bootstrap interval obtained in Table 3.17, however the interval for δ_2 is much narrower than the bootstrap interval.

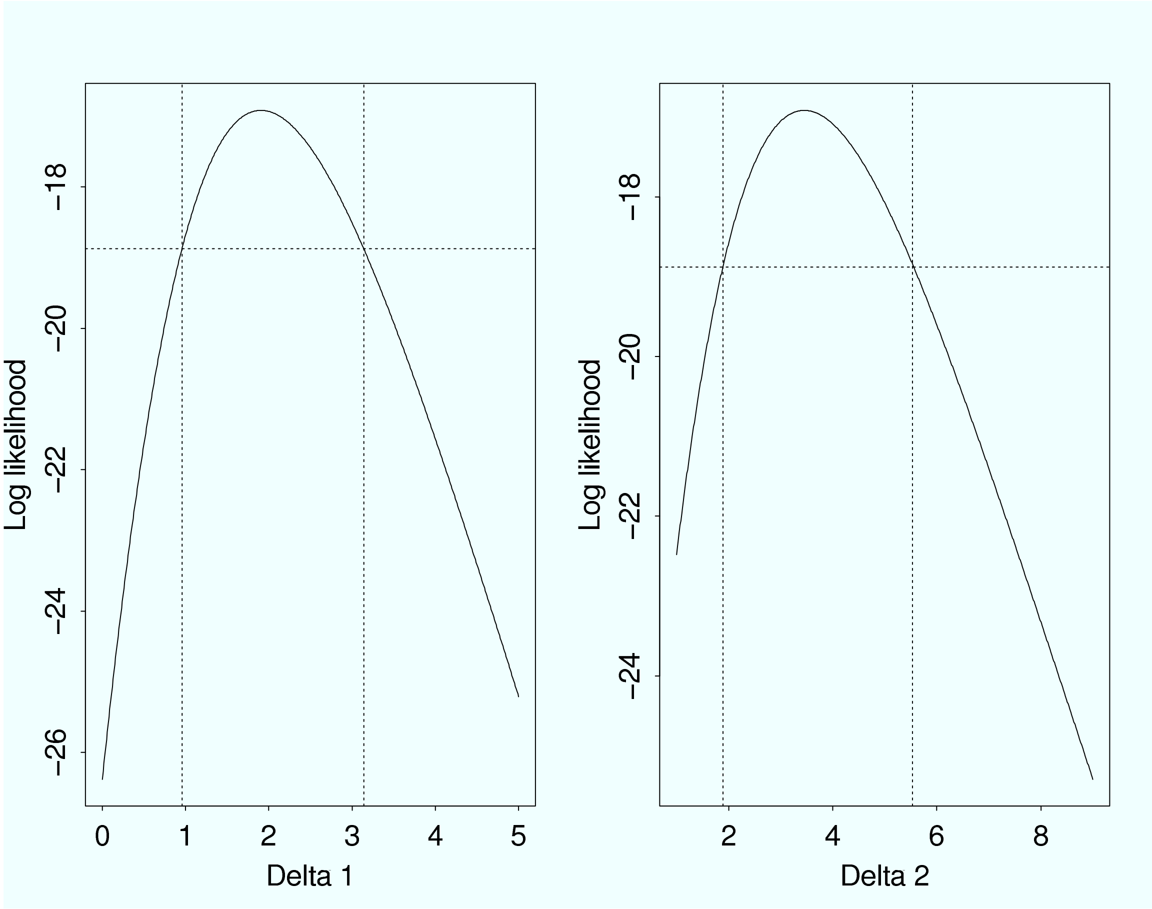


Figure 3.9: Profile log likelihood functions for δ_1 and δ_2.

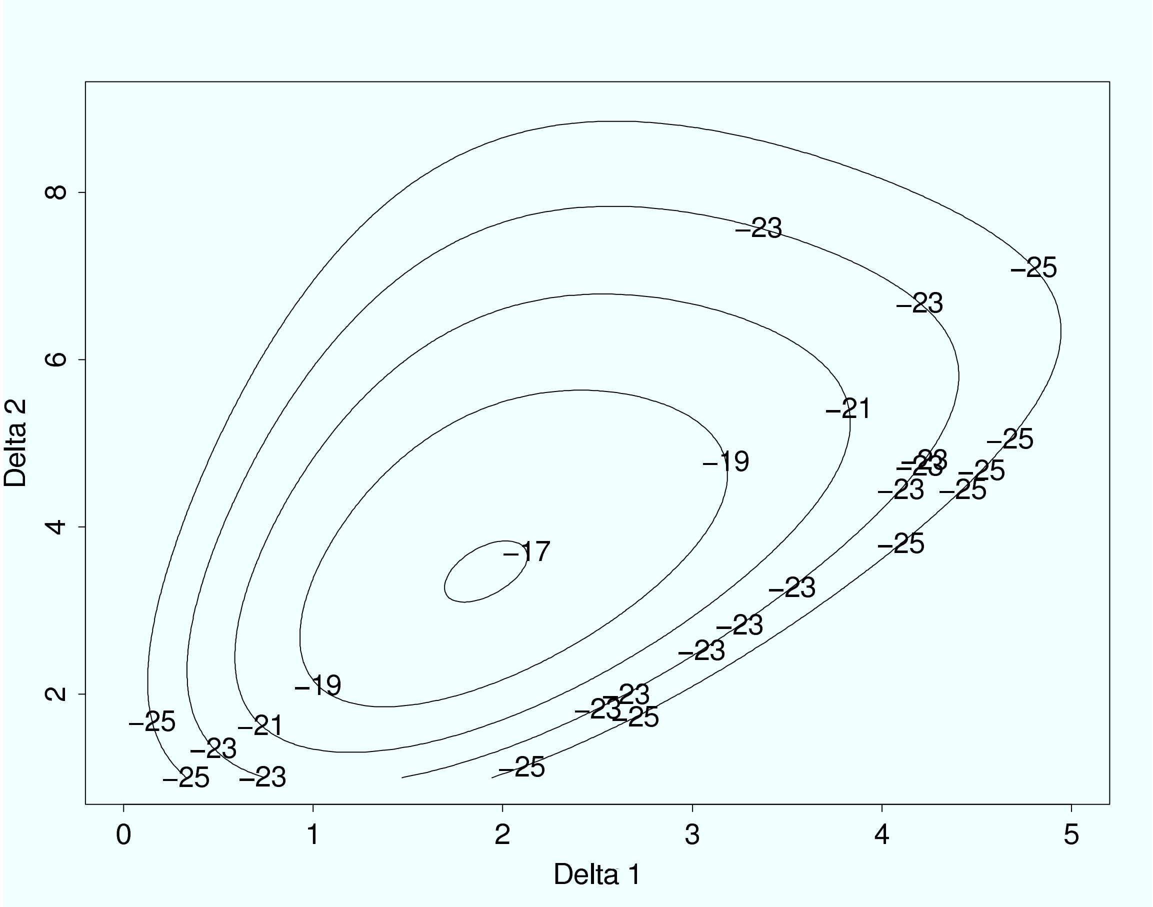


Figure 3.10: Contour plot of the log likelihood as a function of δ_1 and δ_2.

The joint likelihood contours are quite elliptical and show a positive association between the two log odds ratio estimates.

3.7 Discussion

Tests of association between a locus and a disease using the data from an affected individual and a relative using likelihood ratio tests based on the heterogeneous disease model are described in this chapter. For affected relative pairs, the power of the test increases with the nearness in the relatedness, however, for the discordant relative pairs, the power of the test increases with the distance in the relatedness.

The results of the unconditional, partially and fully conditional tests for discordant relative pairs are similar, but the power of the unconditional test is slightly larger than that of the partially and fully conditional tests. This may be because the effective sample size for the conditional tests is smaller than for the unconditional model (Table 3.14) because the data from pairs with the same genotypes are ignored. However, there is a very slight difference between the power of the test under the fully conditional model (Figure 3.5) and under the partially conditional model (Figure 3.3). The effective sample size under the fully conditional model (Table 3.14) and the partially conditional model (Table 3.8) are very similar. The effective sample size increases with the overall sample size, n, the minor allele frequency, q and the distance in relatedness. The power of the test in all three cases increases with the distance in relatedness.

One of the advantages of the fully conditional test is that it does not require any assumptions regarding the prevalence of disease, K_P, or allele frequency, q. The power of the fully conditional model is larger than that of McNemar's test in most cases considered and it also gives information regarding the mode of inheritance of the disease in addition to the degree of association. In the example, the fully conditional test considers a broader range of tables to be as extreme or more extreme than the observed table because of an extra degree of freedom than the McNemar's test.

Another advantage of using family data is that it is robust to population stratification, which can lead to spurious association. The next chapter, Chapter 4 discusses testing for association accounting for possible stratification in the population.

Chapter 4

Accounting For Stratification And Genetic Association In Case-Control Studies

4.1 Introduction

A stratified population is one that contains sub-populations with different allele frequencies at the locus of interest. One of the reasons for stratification is migration of individuals from one population into another; for example, migration of Asians or Africans to Europe and North America. If a certain disease is more prevalent in one sub-population than the other, taking a random sample of cases without regard to the sub-populations, is likely to contain more subjects from that sub-population. Thus a case-control design will show association with any locus with different allele frequencies in the two sub-populations leading to spurious association.

There are several reasons for observing Hardy-Weinberg Disequilibrium (HWD), including stratification, genotyping error and selection bias. This chapter gives an extension of the heterogeneous disease model for case-control studies introduced in Chapter 1 to include both association and population stratification. The extended model can be used to determine if the observed HWD can be explained by stratification and/or association with the disease.

The following section, Section 4.2, summarizes the heterogeneous disease model described in Chapter 1. In Section 4.3 genotypic frequencies for stratified population are derived. Section 4.4 discusses the model fitting and assessment followed by the inferences for the extended model in Section 4.5. Simulations were carried out to assess the level and power of the test of stratification and association. Section 4.6 discusses the details and results of the simulations. The Hardy-Weinberg disequilibrium coefficients for cases and controls in a stratified population are obtained in Section 4.7. Also discussed are the direction and magnitude of the HWD coefficient in specific genetic models. The method was applied to an augmented heterogeneous stone former cohort and a control sample (Cole et al., 1998) to demonstrate how this method can be used to apportion HWD to genetic association in the presence of stratification in Section 4.8. The results are discussed in Section 4.9. This chapter contains material from the paper "Attributing Hardy-Weinberg disequilibrium to population stratification and genetic association in case-control studies", published in Annals of Human Genetics, 2010, 74(1), 77-87.

4.2 Heterogeneous Disease Model

In Chapter 1 a heterogeneous disease model was described for case-control studies for a biallelic locus, with wild type and variant alleles A and a, with frequencies

$1-q$ and q, assuming Hardy-Weinberg equilibrium (HWE) in the population. For the genotypes, $0(AA), 1(Aa)$ and $2(aa)$, with frequencies $P_0 = (1-q)^2$, $P_1 = 2q(1-q)$ and $P_2 = q^2$ in the population, Wittke-Thompson et al. (2005) obtained case ($d = D$) and control ($d = C$) frequencies

$$P_{id} = \frac{P(d|i)P_i}{P(d)}, \tag{4.1}$$

where $P(D|i) = \phi_i$ is the penetrance for genotype i, i= $0, 1$ or 2 and $P(D) = K_P$ is the prevalence of disease

$$K_P = P_0\phi_0 + P_1\phi_1 + P_2\phi_2.$$

Writing the penetrances in terms of the baseline risk, $\alpha = \phi_0$, the heterozygote relative risk, $\beta = \phi_1/\phi_0$ and homozygote relative risk $\gamma = \phi_2/\phi_0$, Wittke-Thompson et al. (2005) also obtained expressions for the Hardy-Weinberg coefficients (Weir (1996))

$$D = P_{AA} - P_A^2 = P_0 - (1-q)^2 \tag{4.2}$$

for cases and controls as

$$D_D = \frac{q^2(1-q)^2\alpha^2(\gamma-\beta^2)}{K_P^2} \tag{4.3}$$

and

$$D_C = \frac{q^2(1-q)^2\alpha(2\beta-1-\gamma-\alpha\beta^2+\alpha\gamma)}{(1-K_P)^2}, \tag{4.4}$$

respectively. After fitting the model to the data, a lack of fit (LOF) test is used to indicate whether any observed HWD is consistent with the genetic association.

4.3 Genotypic Frequencies For A Stratified Population

Consider a population that has two strata, labelled as 1 and 2, with proportions $\varepsilon, 1-\varepsilon$ in the population. The penetrances, ϕ_0, ϕ_1, ϕ_2, are assumed to be constant over both strata.

The genotypic frequency in the stratified population is a weighted sum of the respective stratum genotypic frequencies

$$P_i = \varepsilon P_{i1} + (1-\varepsilon)P_{i2},$$

where P_{ij} is the frequency of the genotype i, where $i = 0, 1, 2$ is the number of variant alleles in strata j, $j = 1, 2$. Assuming HWE in each sub-population the genotypic frequencies are given by

$$P_0 = \varepsilon(1-q_1)^2 + (1-\varepsilon)(1-q_2)^2 \tag{4.5}$$

$$P_1 = \varepsilon 2q_1(1-q_1) + (1-\varepsilon)2q_2(1-q_2) \tag{4.6}$$

and

$$P_2 = \varepsilon q_1^2 + (1-\varepsilon)q_2^2 \tag{4.7}$$

where q_j is the minor allele frequency in strata j, $j = 1, 2$. The prevalence of the disease in stratum $j, j = 1, 2$ is

$$\begin{aligned} K_{P_j} =& P_{0j}\phi_0 + P_{1j}\phi_1 + P_{2j}\phi_2 \\ =& (1-q_j)^2\phi_0 + 2(1-q_j)q_j\phi_1 + q_j^2\phi_2 \end{aligned}$$

and the prevalence of disease in the population is

$$\begin{aligned}P(D) = K_P &= P_0\phi_0 + P_1\phi_1 + P_2\phi_2 \\ &= [\varepsilon P_{01} + (1-\varepsilon)P_{02}]\phi_0 + [\varepsilon P_{11} + (1-\varepsilon)P_{12})\phi_1 + (\varepsilon P_{21} + (1-\varepsilon)P_{22}]\phi_2 \\ &= [\varepsilon(1-q_1)^2 + (1-\varepsilon)(1-q_2)^2]\phi_0 \\ &+ [\varepsilon 2q_1(1-q_1) + (1-\varepsilon)2q_2(1-q_2)]\phi_1 + [\varepsilon q_1^2 + (1-\varepsilon)q_2^2]\phi_2 \\ &= \varepsilon[(1-q_1)^2\phi_0 + 2(1-q_1)q_1\phi_1 + q_1^2\phi_2] \\ &+ (1-\varepsilon)[(1-q_2)^2\phi_0 + 2(1-q_2)q_2\phi_1 + q_2^2\phi_2]\end{aligned}$$

or

$$K_P = \varepsilon K_{P_1} + (1-\varepsilon)K_{P_2}, \tag{4.8}$$

and is a convex combination of the disease prevalences K_{P_j} in each stratum, $j = 1, 2$.

The probability of a genotype i, $i = 0, 1, 2$ conditional on disease status d, $d = D, C$ is

$$P_{id} = P(i|d) = \frac{P(i \bigcap d)}{P(d)}.$$

Applying Bayes' Rule allows the numerator to be expressed in terms of penetrances and population genotypic frequencies

$$P_{id} = \frac{P(d|i)P_i}{P(d)}. \tag{4.9}$$

Using (4.5), (4.6), (4.7) and (4.9), the genotypic frequencies for cases and controls can be summarized as in Table 4.1.

Table 4.1: The genotypic frequencies for cases and controls.

Genotype	Frequency
Cases (D)	
AA (0)	$\frac{\phi_0}{K_P}[\varepsilon(1-q_1)^2 + (1-\varepsilon)(1-q_2)^2]$
Aa (1)	$\frac{\phi_1}{K_P}[2\varepsilon(1-q_1)q_1 + 2(1-\varepsilon)(1-q_2)q_2]$
aa (2)	$\frac{\phi_2}{K_P}[\varepsilon q_1^2 + (1-\varepsilon)q_2^2]$
Controls (C)	
AA (0)	$\frac{1-\phi_0}{1-K_P}[\varepsilon(1-q_1)^2 + (1-\varepsilon)(1-q_2)^2]$
Aa (1)	$\frac{1-\phi_1}{1-K_P}[2\varepsilon(1-q_1)q_1 + 2(1-\varepsilon)(1-q_2)q_2]$
aa (2)	$\frac{1-\phi_2}{1-K_P}[\varepsilon q_1^2 + (1-\varepsilon)q_2^2]$

When the stratification proportion, ε is either zero or one or when the minor allele frequencies are the same in the two strata, *i.e.* $q_1 = q_2$, there is only one population and the results reduce to the heterogeneous model (4.1) as described in Chapter 1.

4.4 Model Fitting And Assessment

In case-control association studies the data consists of genotypic counts for cases and controls denoted by n_{iD} and n_{iC} respectively, $i = 0, 1, 2$; as shown in Table 1.1, Chapter 1. The genotypic frequencies (Table 4.1) with the observed counts for cases and controls form a multinomial likelihood function

$$L(\mathbf{p}, \mathbf{\Phi}) = \prod_{d=\{D,C\}} \prod_{i=0}^{2} P_{id}^{n_{id}},$$

where $\mathbf{p}$ and $\mathbf{\Phi}$ are vectors containing the genotypic probabilities and penetrances respectively. The disease prevalence K_P is assumed to be known from other sources because case-control studies give no information about the prevalence of disease. The constraint (4.8) allows one of the parameters to be evaluated from the others. For example, the baseline prevalence, ϕ_0 can be written as a function of K_P, ϕ_1, ϕ_2, q_1 and q_2 as

$$\phi_0 = \frac{K_P - P_1\phi_1 - P_2\phi_2}{P_0},$$

or

$$\phi_0 = \frac{K_P - \phi_1[2\varepsilon q_1(1-q_1) + 2(1-\varepsilon)q_2(1-q_2)] + \phi_2[\varepsilon q_1^2 + (1-\varepsilon)q_2^2]}{\varepsilon(1-q_1)^2 + \varepsilon(1-q_2)^2}. \qquad (4.10)$$

The remaining parameters in the model are $\varepsilon, q_1, q_2, \phi_1, \phi_2$. There are only four degrees of freedom in the data so at most four parameters can be estimated. To allow for LOF testing, at least two or more of ε, q_1 and q_2 are assumed to be known in the analyses. Once the other parameters have been estimated, ϕ_0 can be obtained from equation (4.10).

The parameters can be estimated by maximizing the likelihood function numerically, and approximate standard errors can be obtained by evaluating the inverse of information matrix or by using the nonparametric bootstrap.

A LOF test can be used to assess whether the model is appropriate for the data and whether any observed HWD can be explained by the genetic association with or without stratification in the population. A significant lack of fit would imply presence of other causes or violation of assumptions. Lack of fit can be tested using the goodness of fit statistic

$$X^2 = \sum_{d=\{D,C\}} \sum_{i=0}^{2} \frac{(n_{id} - n_d \hat{P}_{id})^2}{n_d \hat{P}_{id}},$$

where n_D and n_C are the total number of cases and controls in the dataset. The genotypic frequencies here are evaluated at the the maximum likelihood estimates from the fitted model.

In large samples the X^2 statistic is distributed as a χ^2 with degrees of freedom depending on the number of parameters estimated. If q_1 and q_2 are assumed fixed and ε, ϕ_1 and ϕ_2 are estimated, then there is one degree of freedom. If q_1, q_2 and ε are assumed to be fixed and ϕ_1 and ϕ_2 are estimated, then there are two degrees of freedom.

When the sample size and minor allele frequencies are small, the χ^2 approximation to the distribution of X^2 may not be valid. In this case, the parametric bootstrap can be used, where X^2 is compared to the distribution of the values obtained from samples generated from the fitted model.

4.5 Inferences For The Model With Stratification

It is possible to make inferences about the size and type of the genetic effect and the extent of stratification by comparing the fit of the general model with that of a reduced model with fewer effects, using a likelihood ratio test (LRT) described in

Chapter 1. The hypotheses of interest are

H_0:	Neither genetic nor stratification effects,	$\varepsilon = 0, \beta = \gamma = 1$
H_1:	Stratification effects only,	$\beta = \gamma = 1$
H_2:	Genetic effects only,	$\varepsilon = 0$
H_a:	Both genetic and stratification effects,	$0 < \varepsilon < 1, \beta, \gamma$

It is of most interest to test for genetic effects in the presence of stratification, which can be done by comparing the hypotheses H_1 and H_a. The alternative hypothesis has three parameters and the null hypothesis has one so the test statistic, Λ, is asymptotically distributed as χ^2 with two degrees of freedom.

To test for stratification in the presence of genetic effects, the hypotheses H_2 and H_a are compared. Under H_2 the stratification proportion, ε, lies on the boundary of the parameter space so the usual χ^2 approximation does not apply. In this case the test statistic Λ has a distribution which is a 50:50 mixture of a mass of probability at zero and the χ^2_1 distribution (Self and Liang, 1987).

Insight into the form of genetic effects: dominant, recessive, additive and multiplicative, or can be obtained by imposing constraints on the pentrances and comparing the fit to the general model. In these cases the likelihood ratio test has a χ^2 distribution with one degree of freedom.

4.6 Simulations To Assess Level And Power

Level and power of the proposed hypothesis tests were determined by simulation based on an assumed disease prevalence of $K_P = 0.02$. For the simulations, 1,000 replicated data sets were used for each combination of three different values of the stratification proportion ε (0.05, 0.10, 0.20), six different combinations of the minor allele frequencies q_1 and q_2 ((0.05, 0.1), (0.05, 0.3), (0.05, 0.5), (0.1, 0.3), (0.1, 0.5), $(0.3, 0.5)$), three values of the relative risks β and γ, (1, 2, 3), two different samples sizes for cases $(300, 1000)$ and two ratios of cases to controls r $(1, 4)$. The level and power of the tests described above are approximated using the proportion of hypotheses rejected when the data is generated under the appropriate null and alternative hypotheses, respectively.

4.6.1 Significance Level Of The Tests

To assess the level of the test for genetic effects in the presence of stratification, data were generated under the hypothesis of no genetic effect (H_1) for various choices of ε, q_1 and q_2 and the likelihood ratio test $(H_1$ *vs.* $H_a)$ was carried out at the 0.05 level of significance. Table 4.2 shows that the proportion of times H_1 is correctly rejected is close to the nominal level.

Table 4.2: Type I errors obtained by simulation for tests for genetic effects in the presence of stratification at $\alpha = 0.05$

Parameters			**Type 1 Errors**		
ε	q_1	q_2	$n_D = 300$ $r = 1$	$n_D = 300$ $r = 4$	$n_D = 1000$ $r = 1$
0.2	0.05	0.1	0.067	0.047	0.052
0.2	0.05	0.3	0.051	0.050	0.041
0.2	0.05	0.5	0.052	0.055	0.056
0.2	0.1	0.3	0.050	0.057	0.050
0.2	0.1	0.5	0.062	0.044	0.062
0.2	0.3	0.5	0.055	0.051	0.061
0.1	0.05	0.1	0.063	0.073	0.046
0.1	0.05	0.3	0.044	0.053	0.046
0.1	0.05	0.5	0.047	0.046	0.051
0.1	0.1	0.3	0.049	0.048	0.054
0.1	0.1	0.5	0.058	0.052	0.054
0.1	0.3	0.5	0.047	0.056	0.049
0.05	0.05	0.1	0.060	0.060	0.039
0.05	0.05	0.3	0.052	0.048	0.057
0.05	0.05	0.5	0.052	0.042	0.046
0.05	0.1	0.3	0.042	0.044	0.043
0.05	0.1	0.5	0.050	0.054	0.054
0.05	0.3	0.5	0.042	0.048	0.042

To assess the distribution of the likelihood ratio test statistic, the χ^2_2 Q-Q plot of the test statistic were plotted (Figure E.1, Figure E.2, Figure E.3 in Appendix E). The figures indicate that the LRT statistic follows a χ^2 distribution with two degrees of freedom.

In this simulation, the standard error of the estimate of ε was seen to depend on the actual value of ε, the difference in the values of the minor allele frequencies of the two strata, and the inverse of the square root of the sample size. Linear regression of $\log(SE(\hat{\varepsilon}))$ on these factors for the 18 cases in Table 4.2 gives

$$log(SE(\sigma_{\hat{\varepsilon}})) = -3.09 + 0.83\varepsilon - 3.67|q_2 - q_1| + 16.89(1/\sqrt{n})$$

with $R^2 = 0.93$ and all terms are significant at the $\alpha = 0.05$ level. This shows that the standard error increases with ε but decreases with the sample size and the difference in the allele frequencies.

To assess the level of the test for stratification in the presence of genetic effects, data were generated under the hypothesis of no stratification (H_2) and the LRT (H_2 *vs.* H_a) was carried out at the 0.05 level of significance. The simulated levels in Table 4.3 are close to the nominal level.

Table 4.3: Type I errors obtained by simulation for the test for stratification in the presence of genetic effects at $\alpha = 0.05$

Parameters			**Type 1 Errors**		
q	β	γ	$n_D = 300$ $r = 1$	$n_D = 300$ $r = 4$	$n_D = 1000$ $r = 1$
0.05	1	3	0.062	0.043	0.053
0.05	3	1	0.054	0.054	0.047
0.05	3	3	0.042	0.050	0.061
0.05	3	6	0.062	0.042	0.053
0.05	3	9	0.058	0.049	0.049
0.1	1	3	0.044	0.043	0.061
0.1	3	1	0.056	0.047	0.055
0.1	3	3	0.044	0.052	0.034
0.1	3	6	0.041	0.059	0.054
0.1	3	9	0.054	0.046	0.045
0.3	1	3	0.056	0.047	0.045
0.3	3	1	0.043	0.050	0.046
0.3	3	3	0.055	0.062	0.051
0.3	3	6	0.043	0.044	0.051
0.3	3	9	0.050	0.061	0.041

The distribution of this likelihood ratio test statistic should be distributed as a 50:50 mixture of a χ^2_0 (point mass at 0) and a χ^2_1 (Self and Liang, 1987). To verify this, the proportion of times the population proportion estimate, $\hat{\varepsilon}$, under H_a is zero was calculated (Table F.1 in Appendix F) and the Q-Q plots of the non-zero values of the test statistic were plotted for the χ^2_1 distribution (Figure F.1, Figure F.2 and Figure F.3 in Appendix F). Table F.1 illustrates that $\hat{\varepsilon}$ is zero approximately 50% of the time and Figures F.1, F.2 and F.3, confirm that the distribution of the non-zero values of the LRT statistic follow a χ^2 distribution with one degree of freedom.

4.6.2 Power Of The Tests

To assess the power of the tests for genetic effects in the presence of stratification and stratification in presence of genetic effects, data were generated under the alternative hypothesis (H_a) and LRTs (H_2 and H_1 *vs.* H_a) were carried out. The simulated power of the tests are summarized in Figures 4.1 and 4.2.

For both hypotheses the power increases with the sample size. The power for the test of genetic effects increases with the size of the genetic effect and with the increase in the difference between the minor allele frequencies of the two strata. The power of the test for stratification is not affected by the size of the genetic effect. For all tests the power increases with the increase in the difference between the minor allele frequency of the two strata and with the increase in the stratification proportion.

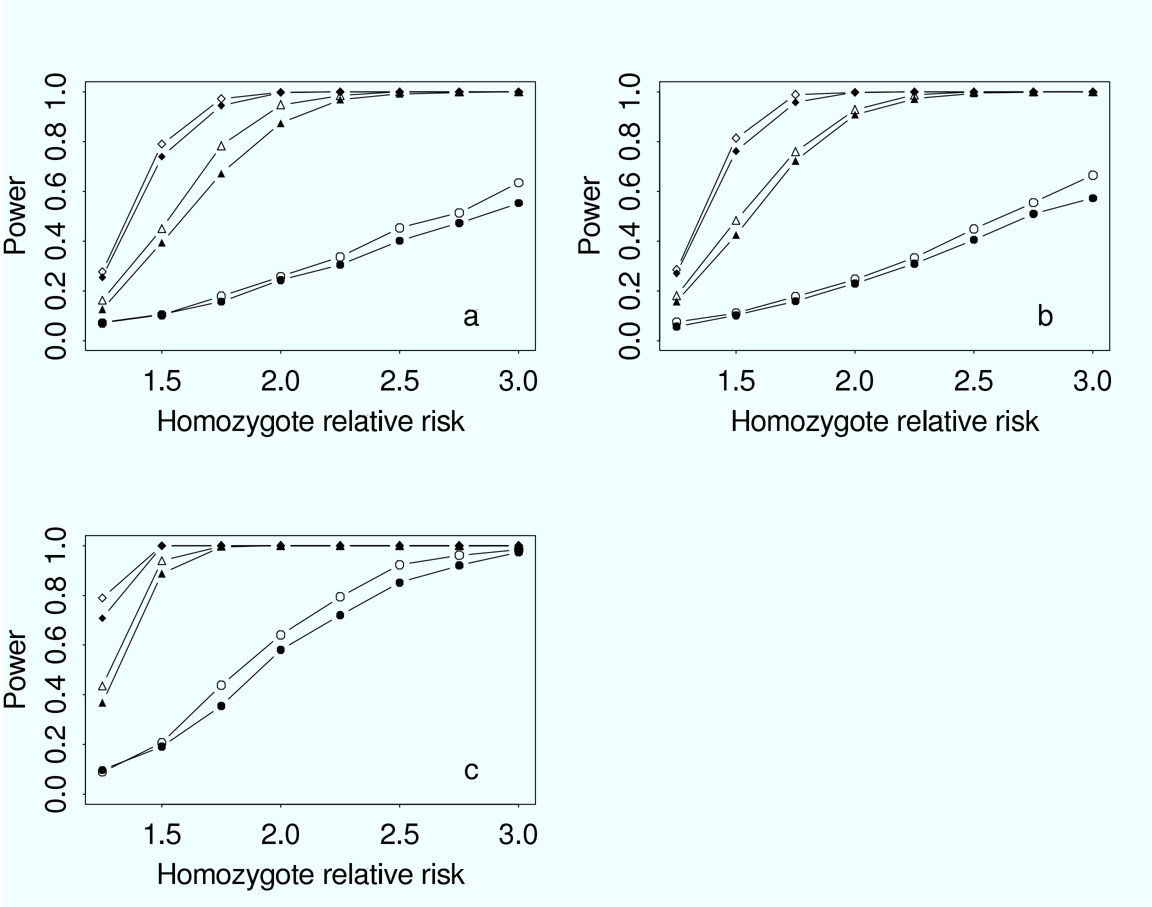


Figure 4.1: Power of the test for genetic association in the presence of stratification for $K_P = 0.02$, $q_1 = 0.05$, $\beta = 1$, $q_2 = (0.1, 0.3, 0.5) = (\circ, \triangle, \Diamond)$. Open/filled symbols for $\varepsilon = 0.05/0.2$. a) $n = 300$, $r = 1$ b) $n = 300$, $r = 4$ c) $n = 1000$, $r = 1$

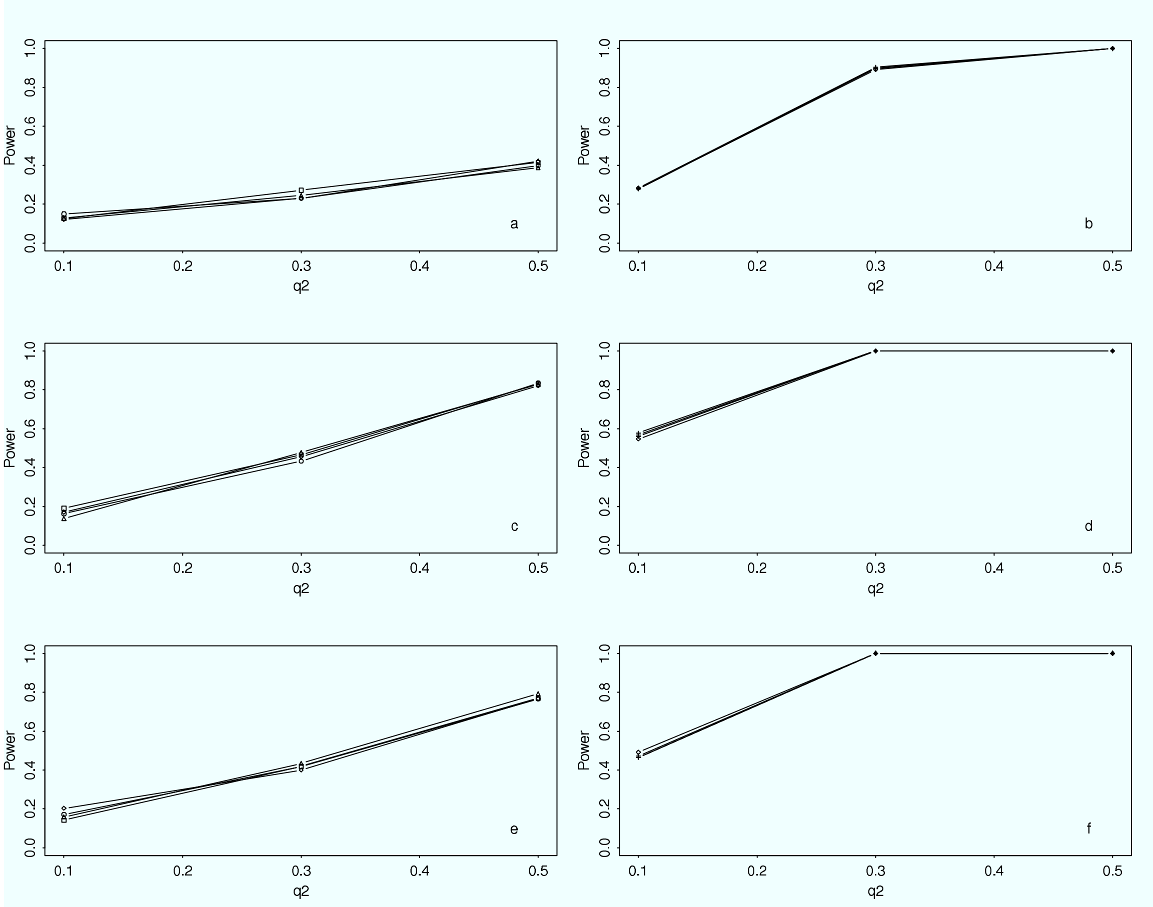


Figure 4.2: Power of the test for stratification in presence of genetic effects for $K_P = 0.02$, $\beta = 1$, $q_1 = 0.05$, $\gamma = (1.5, 2, 2.5, 3) = (\Diamond, \Box, \circ, \triangle)$. Left panel $\varepsilon = 0.05$, right panel $\varepsilon = 0.2$ a) and b) $n = 300$, $r = 1$ c) and d) $n = 300$, $r = 4$ e) and f) $n = 1000$, $r = 1$

4.7 The Hardy-Weinberg Disequilibrium Coefficient For The Stratified Population

The Hardy-Weinberg coefficient, D, measures the excess homozygosity and is given by

$$D = P_{aa} - q^2,$$

where the minor allele frequency can be obtained from the genotypic frequencies using the relationship

$$q = P_{aa} + \frac{1}{2}P_{Aa}.$$

Substituting for P_{aa} and P_{Aa} from (4.6) and (4.7), respectively gives

$$\begin{aligned} q &= [\varepsilon P_{21} + (1-\varepsilon)P_{22}] + \frac{1}{2}[\varepsilon P_{11} + (1-\varepsilon)P_{12}] \\ &= \varepsilon(P_{21} + \frac{1}{2}P_{11}) + (1-\varepsilon)(P_{22} + \frac{1}{2}P_{12}) \end{aligned}$$

or

$$q = \varepsilon q_1 + (1-\varepsilon)q_2, \tag{4.11}$$

which is a convex combination of the minor allele frequencies in the two strata.

The Hardy-Weinberg coefficient in the stratified population assuming each stratum is in HWE is

$$\begin{aligned} D_S =& P_2 - q^2 \\ =& [\varepsilon q_1^2 + (1-\varepsilon)q_2^2] - [\varepsilon q_1 + (1-\varepsilon)q_2]^2 \\ =& \varepsilon q_1^2 + (1-\varepsilon)q_2^2 - \varepsilon^2 q_1^2 - (1-\varepsilon)^2 q_2^2 - 2\varepsilon(1-\varepsilon)q_1 q_2 \\ =& \varepsilon(1-\varepsilon)(q_1^2 - 2q_1q_2 + q_2^2) \end{aligned}$$

which simplifies to

$$D_S = \varepsilon(1-\varepsilon)(q_2 - q_1)^2 \tag{4.12}$$

in both cases and controls (Deng et al., 2001). This expression reduces to zero if there is only one stratum, *i.e.* $\varepsilon = 0$ or 1, or $q_1 = q_2$.

The minor allelic frequency among patients, using the genotypic frequencies from Table 4.1 is

$$\begin{aligned} q_D &= \frac{\phi_2}{K_P}[\varepsilon q_1^2 + (1-\varepsilon)q_2^2] + \frac{1}{2}\frac{\phi_1}{K_P}[2\varepsilon(1-q_1)q_1 + 2(1-\varepsilon)(1-q_2)q_2] \\ &= \frac{\varepsilon}{K_P}[q_1^2\phi_2 + q_1(1-q_1)\phi_1] + \frac{1-\varepsilon}{K_P}[q_2^2\phi_2 + q_2(1-q_2)\phi_1] \end{aligned}$$

or

$$q_D = \varepsilon q_{1D} + (1-\varepsilon)q_{2D}, \tag{4.13}$$

which is a convex combination of the minor allele frequencies.

The Hardy-Weinberg coefficient for cases, using (4.13) and the genotypic frequencies in Table 4.1 is given by

$$\begin{aligned} D_D =& P_{2D} - q_D^2 \\ =& \frac{\phi_2}{K_P}[\varepsilon q_1^2 + (1-\varepsilon)q_2^2] - \{\frac{\varepsilon}{K_P}[q_1^2\phi_2 + q_1(1-q_1)\phi_1] + \frac{1-\varepsilon}{K_P}[q_2^2\phi_2 + q_2(1-q_2)\phi_1]\}^2 \\ =& \frac{\alpha^2}{K_P^2}\{(\gamma-\beta^2)[\varepsilon(1-q_1)q_1 + (1-\varepsilon)(1-q_2)q_2]^2 + \gamma\varepsilon(1-\varepsilon)(q_2-q_1)^2\} \end{aligned}$$

which simplifies to

$$D_D = \frac{\alpha^2}{K_P^2}\left[\gamma D_S + (\gamma-\beta^2)V^2\right] \tag{4.14}$$

where

$$V = \varepsilon q_1(1-q_1) + (1-\varepsilon)q_2(1-q_2).$$

Similarly, the minor allelic probability among controls, using the genotypic frequencies from Table 4.1 is

$$\begin{aligned} q_C = & \frac{\varepsilon}{1-K_P}[q_1^2(1-\phi_2)+q_1(1-q_1)(1-\phi_1)] \\ & + \frac{1-\varepsilon}{1-K_P}[q_2^2(1-\phi_2)+q_2(1-q_2)(1-\phi_1)] \end{aligned}$$

or

$$q_C = \varepsilon q_{1C} + (1-\varepsilon)q_{2C} \tag{4.15}$$

which is a convex combination of the minor allele frequencies. The HWD coefficient for controls is

$$\begin{aligned} D_C =& P_{2C} - q_C^2 \\ =& \frac{1-\phi_2}{1-K_P}[\varepsilon q_1^2 + (1-\varepsilon)q_2^2] - \frac{1}{(1-K_P)^2}\{\varepsilon[q_1^2(1-\phi_2)+q_1(1-q_1)(1-\phi_1)] \\ & + (1-\varepsilon)[q_2^2(1-\phi_2)+q_2(1-q_2)(1-\phi_1)]\}^2 \\ =& \frac{1}{(1-K_P)^2}\{(1-\alpha)(1-\alpha\gamma)\varepsilon(1-\varepsilon)(q_2-q_1)^2 \\ & + \alpha(2\beta-1-\gamma-\alpha\beta^2+\alpha\gamma)[\varepsilon q_1(1-q_1)+(1-\varepsilon)q_2(1-q_2)]^2\} \end{aligned}$$

which reduces to

$$D_C = \frac{1}{(1-K_P)^2}\left[(1-\alpha)(1-\alpha\gamma)D_S + \alpha(2\beta-1-\gamma-\alpha\beta^2+\alpha\gamma)V^2\right]. \tag{4.16}$$

The HWD coefficients for both cases and controls are combinations of D_S, the expression for a stratified population, and the coefficients for an unstratified population (Wittke-Thompson et al., 2005). If there is only one stratum, *i.e.* $\varepsilon = 0$ or 1, or $q_1 = q_2$, then $D_S = 0$ and $V = q(1-q)$ and the above expressions simplify to the results of Wittke-Thompson et al. (2005) in equations (4.3) and (4.4), for cases and controls, respectively. When there is no association, *i.e.* $\beta = \gamma = 1$, both expressions simplify to the value for stratified populations, D_S, in equation (4.12).

The HWD coefficients also simplify for specific genetic models.

4.7.1 Specific Genetic Models

The dependence of the HWD coefficients for specific genetic models on the model parameters is illustrated for the specific genetic models by plotting them as a function of q_2 for q_1 =0.25, 0.5, 0.75, ε =0, 0.1, 0.2 and $\gamma = 1.5$, 5.

Dominant Model $\boldsymbol{\gamma = \beta, \beta > 1}$

The HWD coefficient for cases is

$$D_D = \frac{\alpha^2\gamma}{K_P^2}[D_S - (\gamma - 1)V^2]$$

and the HWD coefficient for controls is

$$D_C = \frac{(1-\alpha\gamma)}{(1-K_P)^2}[(1-\alpha)D_S + \alpha(\gamma-1)V^2].$$

Figure 4.3 illustrates the direction and magnitude of HWD in the dominant genetic model for cases and controls as a function of q_2. The HWD coefficient for controls is always positive whereas in cases it is mostly negative for the large values of γ. In controls, the coefficient increases with the stratification proportion, ε, and the absolute difference in allele frequencies. There is little effect of the relative risk, γ. In cases, there is little effect of the stratification proportion except when the absolute difference in the allele frequencies is large.

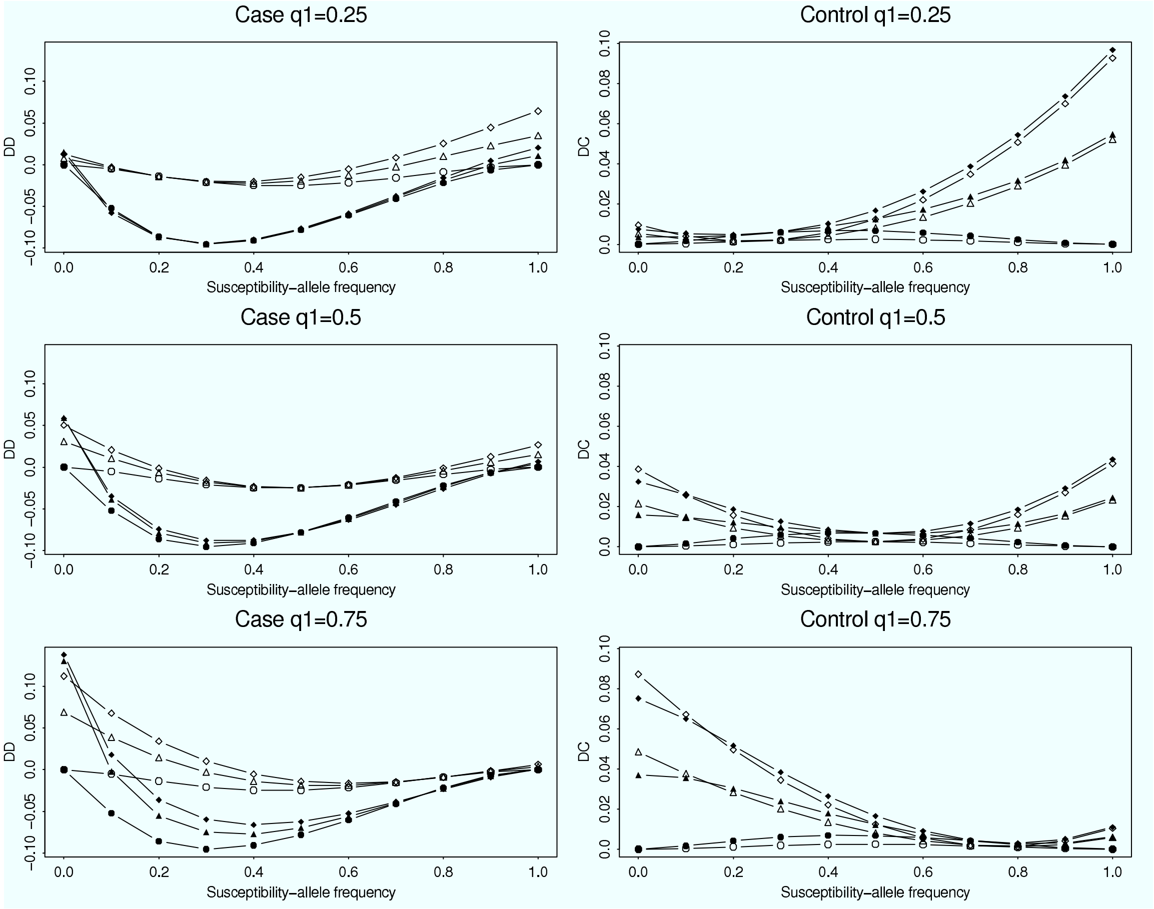


Figure 4.3: HWD as a function of the susceptibility-allele frequency for cases (left) and controls (right), dominant genetic model.
$K_P = 0.1$, $\varepsilon = (0.0, 0.1, 0.2) = (\circ, \triangle, \diamondsuit)$. Open/filled symbols for $\gamma = 1.5/5$.

***Recessive Model* $\boldsymbol{\beta = 1, \gamma > 1}$**

The HWE coefficient for cases is

$$D_D = \frac{\alpha^2}{K_P^2}[\gamma D_S + (\gamma - 1)V^2]$$

and the HWE coefficient for controls is

$$D_C = \frac{(1-\alpha)}{(1-K_P)^2}[(1-\alpha\gamma)D_S - \alpha(\gamma-1)V^2].$$

Figure 4.4 illustrates the direction and magnitude of HWD in the recessive genetic model for cases and controls. The HWD coefficient is always positive for cases. In both cases and controls, the effect of the stratification proportion increases with the absolute difference between the two minor allele frequencies. When the allele frequencies are similar in the two strata, the HWD coefficient increases with γ for cases, but decreases with γ for controls.

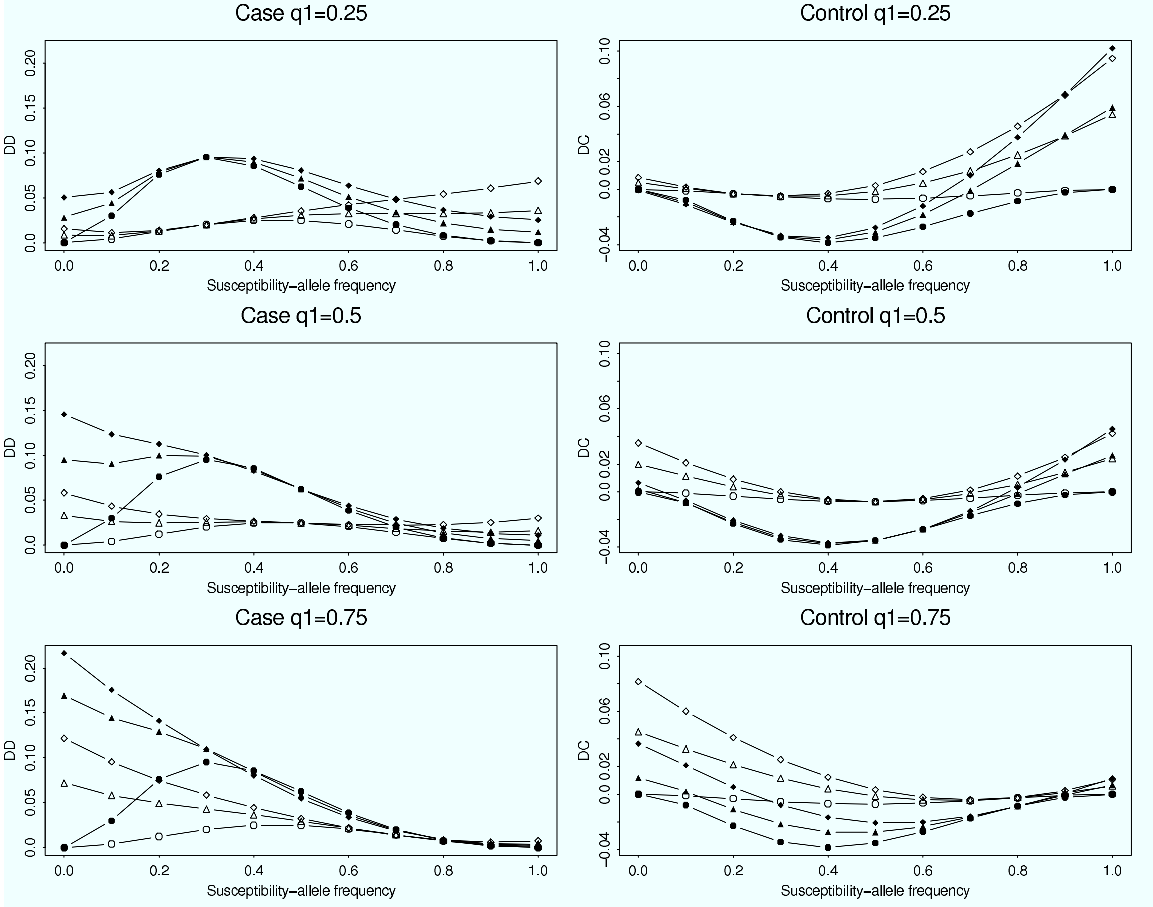


Figure 4.4: HWD as a function of the susceptibility-allele frequency for cases (left) and controls (right), recessive genetic model.
$K_P = 0.2$, $\varepsilon = (0.0, 0.1, 0.2) = (\circ, \triangle, \diamondsuit)$. Open/filled symbols for $\gamma = 1.5/5$.

Additive Model $\boldsymbol{\gamma = 2\beta - 1, \beta > 1}$

The HWD coefficient for cases is

$$D_D = \frac{\alpha^2}{4K_P^2}[4\gamma D_S - (\gamma - 1)^2 V^2]$$

and the HWD coefficient for controls is

$$D_C = \frac{1}{4(1 - K_P)^2}[4(1 - \alpha)(1 - \alpha\gamma)D_S - \alpha^2(\gamma - 1)^2 V^2].$$

When there is no stratification, *i.e.*, $\varepsilon = 0$ or 1 or $q_1 = q_2 = q$ then the HWD coefficients for cases and controls simplify to

$$D_D = -\frac{\alpha^2(\gamma - 1)^2}{4K_P^2} q^2(1-q)^2$$

and

$$D_C = -\frac{\alpha^2(\gamma - 1)^2}{4(1-K_P)^2} q^2(1-q)^2$$

which are always negative. These expressions for the coefficients are different from those obtained by Wittke-Thompson et al. (2005) because they define the additive model to have $\gamma = 2\beta$ rather than $\gamma = 2\beta - 1$, which corresponds to $\phi_2 - \phi_1 = \phi_1 - \phi_0$, an additive effect on the penetrances.

Figure 4.5 illustrates the direction and magnitude of HWD in the additive genetic model. For controls the coefficient increases with the stratification proportion, ε, and the absolute difference in the allele frequencies of the two strata. There is a very little effect of the homozygote relative risk, γ. When $\varepsilon = 0$, the HWD coefficient is very slightly negative. For cases, the effect of stratification is largest when the absolute difference in the allele frequencies is large, and when this difference is small the HWD coefficient decreases with γ.

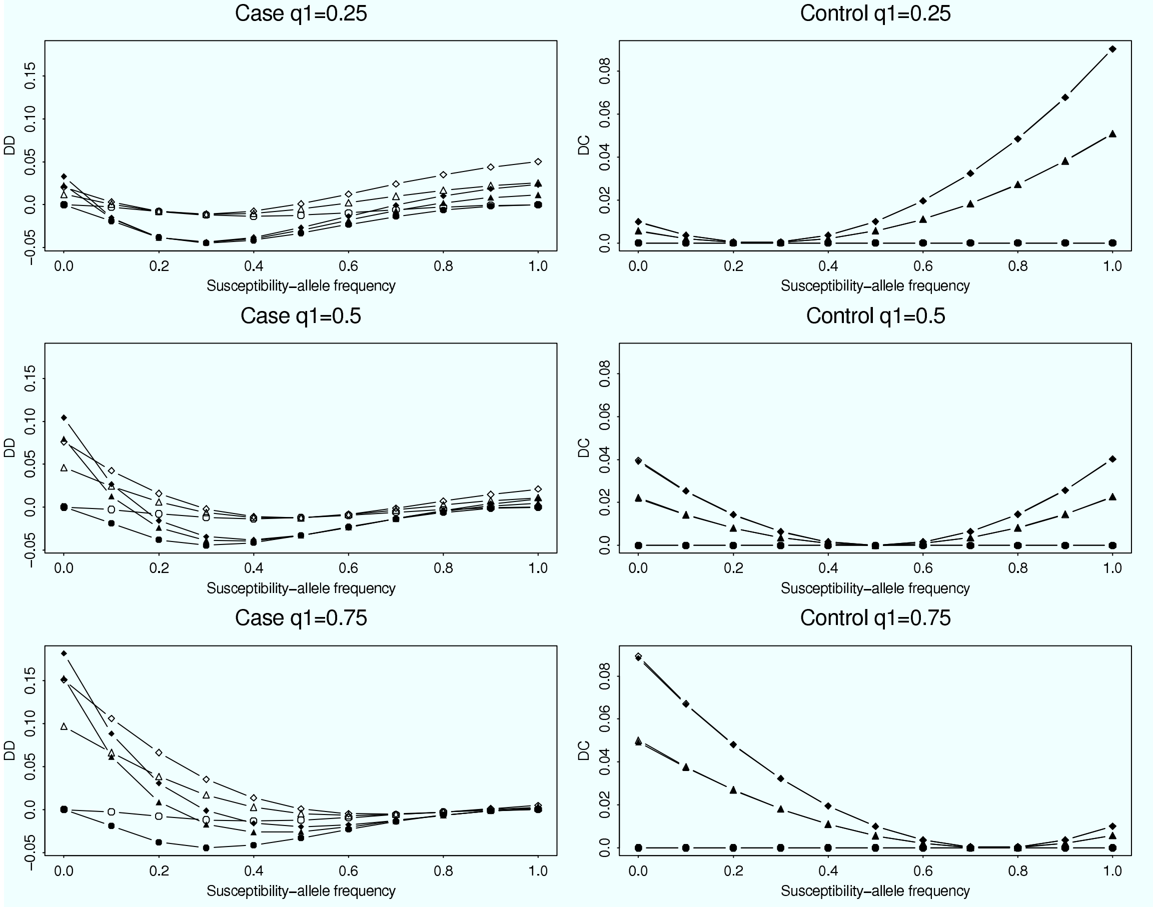


Figure 4.5: HWD as a function of the susceptibility-allele frequency for cases (left) and controls (right), additive genetic model.
$K_P = 0.01$, $\varepsilon = (0.0, 0.1, 0.2) = (\circ, \triangle, \diamondsuit)$. Open/filled symbols for $\gamma = 2.2/5$.

Multiplicative Model $\gamma = \beta^2, \beta > 1$

The HWE coefficient for cases is

$$D_D = \frac{\alpha^2 \gamma D_S}{K_P^2}$$

and the coefficient for controls is

$$D_C = \frac{1}{(1-K_P)^2}[(1-\alpha)(1-\alpha\gamma)D_S - \alpha(\sqrt{\gamma}-1)^2V^2].$$

Figure 4.6 illustrates the direction and magnitude of HWD in the multiplicative genetic model for cases and controls as a function of the minor allele frequency q_2 for different parameter values.

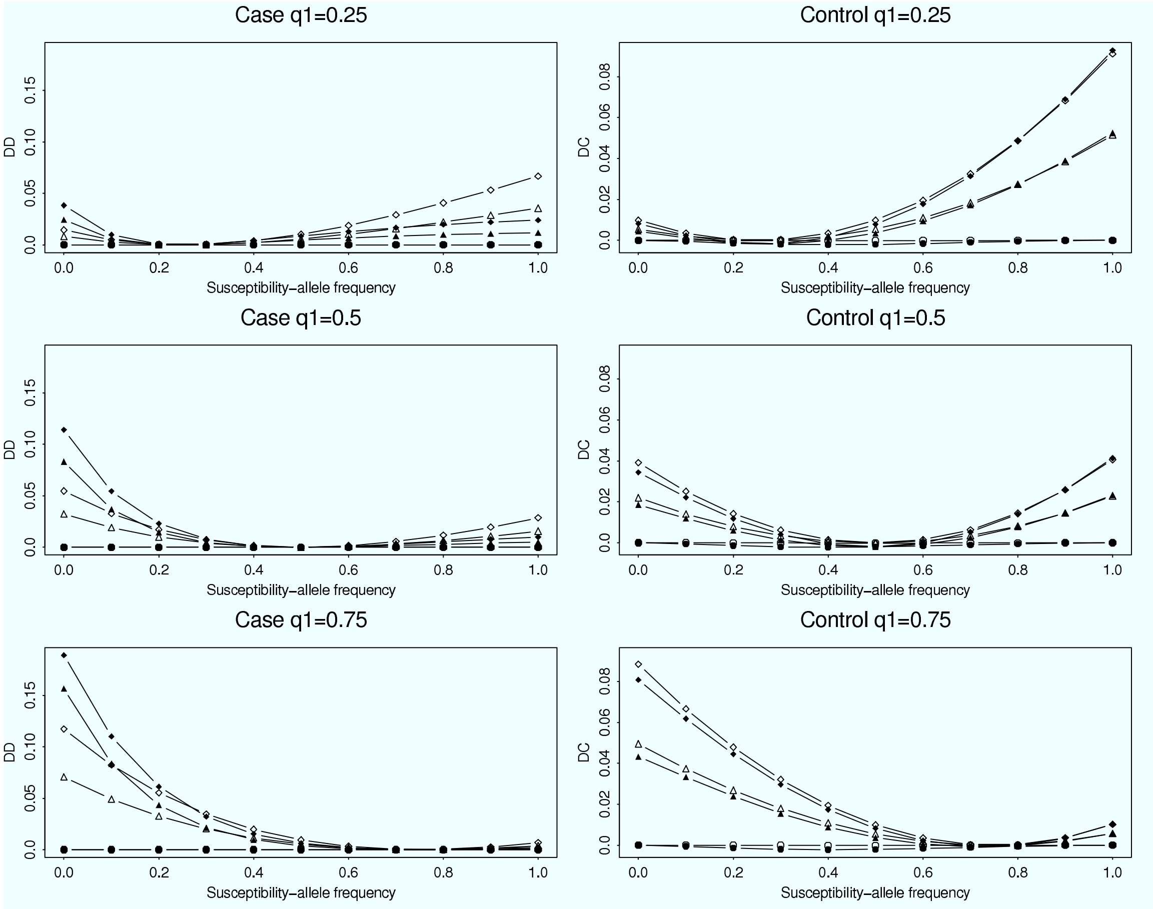


Figure 4.6: HWD as a function of the susceptibility-allele frequency for cases (left) and controls (right), multiplicative genetic model.
$K_P = 0.05$, $\varepsilon = (0.0, 0.1, 0.2) = (\circ, \triangle, \diamondsuit)$. Open/filled symbols for $\gamma = 1.5/5$.

The HWD coefficient is always positive for cases unless there is no stratification when it is zero. The coefficient is slightly negative in controls for some parameter values. For both cases and controls the coefficient increases with the absolute difference between the minor allele frequencies in the two strata and with the amount of stratification for both cases and controls. The homozygote relative risk, γ, has little effect on HWD for controls, but a greater effect for cases, especially when the absolute difference between the allele frequencies is large.

It is possible to obtain the conditions under which the HWD coefficient is positive or negative for the general and the specific disease models (Table 4.4) under the extended model. A positive (negative) value of the HWD represents an excess (deficit) of homozygotes. For most, the HWD coefficient can be positive or negative in both cases and controls, depending on the relative risks, allele frequencies and stratification proportion. Exceptions are the controls for the dominant model and the cases for the recessive model, which have positive HWD coefficients. For the multiplicative model note that the HWD coefficient for cases is positive when there is stratification, rather than zero when there is not, as shown by Wittke-Thompson et al. (2005).

Table 4.4: Sign of the HWD coefficient

Cases		Controls	
Condition	Direction	Condition	Direction
General Genetic Model			
$\gamma > \dfrac{\beta^2}{E^1+1}$	$+$	$\gamma < \dfrac{1}{1+E}\left(\dfrac{E}{\alpha} + \dfrac{\beta(2-\alpha\beta)-1}{1-\alpha}\right)$	$+$
$\gamma < \dfrac{\beta^2}{E+1}$	$-$	$\gamma > \dfrac{1}{1+E}\left(\dfrac{E}{\alpha} + \dfrac{\beta(2-\alpha\beta)-1}{1-\alpha}\right)$	$-$
Dominant, $\beta = \gamma$, $\gamma > 1$			
$\gamma < E+1$	$+$	Always	$+$
$\gamma > E+1$	$-$		
Recessive, $\beta = 1$, $\gamma > 1$			
Always	$+$	$\gamma < \dfrac{1}{1+E}\left(\dfrac{E}{\alpha} + 1\right)$	$+$
		$\gamma > \dfrac{1}{1+E}\left(\dfrac{E}{\alpha} + 1\right)$	$-$
Additive, $\gamma = 2\beta - 1$, $\gamma > 1$			
$\dfrac{(\gamma-1)^2}{\gamma} < 4E$	$+$	$\dfrac{(\gamma-1)^2}{4(1-\alpha\gamma)} < \dfrac{E(1-\alpha)}{\alpha^2}$	$+$
$\dfrac{(\gamma-1)^2}{\gamma} > 4E$	$-$	$\dfrac{(\gamma-1)^2}{4(1-\alpha\gamma)} > \dfrac{E(1-\alpha)}{\alpha^2}$	$-$
Multiplicative, $\gamma = \beta^2$, $\gamma > 1$			
Always	$+$	$\dfrac{(\sqrt{\gamma}-1)^2}{1-\alpha\gamma} < \dfrac{E}{\alpha} - 1$	$+$
		$\dfrac{(\sqrt{\gamma}-1)^2}{1-\alpha\gamma} > \dfrac{E}{\alpha} - 1$	$-$

[1]$E = D_S/V^2$

4.8 Application To A Data Set

The data consist of genotypic counts at the R990G SNP of the calcium-sensing receptor (CASR) gene on 223 calcium stone forming patients (159 men and 64 women, mean age 52.5 ± 12.6 (SD) years) and 718 healthy young adults. The cases were recruited from the Lithotripsy Clinic at The Wellesley-Central Hospital (Toronto ON, Canada) after obtaining written informed consent. The controls were from the same urban population and are mixture of self-reporting Caucasian (n = 673) and Asian Canadians (n = 45) (Rubin et al., 1999; Patel et al., 2000). The cases with uric acid or cysteine stones, co-morbid conditions including hyperparathyroidism, hypercalcemia, and with drug-induced stones were excluded (Cole et al., 1998). The minor allele frequency is known to be higher in the Asian population.

4.8.1 Preliminary Statistical Analysis For Heterogeneous Model

HWD coefficients were calculated for cases and controls (by strata and pooled) and significance was assessed using an exact test (Weir, 1996). Confidence intervals (CI) were obtained using non-parametric bootstrap. Genetic association was assessed by a contingency table analysis of the genotypic counts. The heterogeneous disease model of Wittke-Thompson et al. (2005) was fitted to the data and lack of fit was evaluated.

The observed frequencies and HWD coefficients for cases and controls are summarized in Table 4.5. There is significant HWD in both cases and pooled controls even though neither stratum in the controls has significant HWD. The genotypic test of association shows a strong genetic effect at the locus ($X^2 = 39.85$, df $= 2$, $p < 0.0001$).

Table 4.5: Observed frequencies and HWD coefficient for cases and controls.

	n	RR	RG	GG	$D(CI)$	p
Cases	223	171	38	14	0.04 (0.003, 0.041)	< 0.0001
Controls						
Pooled	718	576	122	20	0.02 (0.006, 0.026)	0.0002
Caucasian	676	568	102	6	0.002 (-0.007, 0.006)	0.46
Asian	42	8	20	14	0.007 (-0.020, 0.030)	0.99

Applying the model of Wittke-Thompson et al. (2005) yielded estimates (95% CI) of the minor allele frequency $\hat{q} = 0.12$ $(0.085, 0.116)$, and penetrances $\hat{\phi}_0 = 0.02$ $(0.018, 0.021)$, $\hat{\phi}_1 = 0.02$ $(0.014, 0.026), \hat{\phi}_2 = 0.08$ $(0.029, 0.140)$. This gives relative risks $\hat{\beta} = \hat{\phi}_1/\hat{\phi}_0 = 0.81$ $(0.654, 1.423)$ and $\hat{\gamma} = \hat{\phi}_2/\hat{\phi}_0 = 3.88$ $(1.453, 7.444)$. The LOF is significant ($X^2 = 17.65$, df $= 1$, $p < 0.0001$,) indicating that this model with only genetic effects does not adequately explain the HWD.

One of the possible reasons for HWD is genotyping error (Xu et al., 2002), another possibility is that of population stratification that is explored in next section.

The heterogeneous disease model as well as the model that accounts for stratification were coded in S-Plus software and the built-in function “nlmin” was used to obtain the maximum likelihood estimates.

4.8.2 Model Fitting That Includes Stratification

The extended model was fitted to the kidney stones data at the R990G locus assuming the population to be a mix of Caucasians and Asians. The minor allele frequencies (MAFs) for the two strata were assumed to be $q_1 = 0.429$ (Asian), $q_2 = 0.084$ (Caucasian) (Yun et al., 2007) for all analyses. The MAF for Caucasians and Asians obtained from Yun et al. (2007) are not significantly different from those obtained by the HAPMAP project (www.hapmap.org).

First, the penetrances, ϕ_0, ϕ_1 and ϕ_2 were estimated for different fixed values of the stratification proportion ε and the LOF test was carried out in each case (Table 4.6).

Table 4.6: Estimates of ϕ_0, ϕ_1 and ϕ_2, X^2 statistic and p-value for a range of fixed ε.

	Estimates			LOF	
ε	$\hat{\phi}_0$	$\hat{\phi}_1$	$\hat{\phi}_2$	$X^2, 2df$	p
0.00	0.0187	0.0224	0.1209	57.2662	< 0.0001
0.01	0.0187	0.0219	0.1092	37.934	< 0.0001
0.05	0.0189	0.02	0.0735	8.05	0.02
0.06	0.019	0.0197	0.0674	5.3532	0.0688
0.07	0.0191	0.0193	0.0622	3.5581	0.1688
0.08	0.0192	0.019	0.0576	2.463	0.2919
0.09	0.0193	0.0186	0.0537	1.9282	0.3813
0.10	0.0195	0.0183	0.0502	1.8543	0.3957
0.11	0.0196	0.018	0.0471	2.1691	0.3381
0.12	0.0197	0.0177	0.0444	2.8183	0.2444
0.13	0.0198	0.0174	0.0419	3.7605	0.1526
0.14	0.0199	0.0171	0.0397	4.9635	0.0836
0.15	0.0201	0.0169	0.0377	6.4019	0.0407
0.16	0.0202	0.0166	0.0359	8.0553	0.0178

There is no LOF for the stratification proportion between 6 and 14% which suggests that the extended model with both stratification and genetic effects explains the data adequately. The minimum X^2 occurs at 10% Asians and 90% Caucasians and the penetrance estimates and relative risks, with confidence intervals in parenthesis obtained using nonparametric bootstrap, are shown in Table 4.7. Note that the LOF results for no stratification $\varepsilon = 0$ in Table 4.6 differ from the basic model described above because in this case the allele frequencies are assumed to be known for each strata, whereas in the basic case the minor allele frequency was estimated and there was only one allele frequency. There was also one lesser degree of freedom for testing lack of fit in the basic model.

Table 4.7: Penetrances estimates under general and recessive disease model

	Assumed	Estimated
	$\varepsilon = 0.10$	ε
	General model	
$\hat{\varepsilon}$	-	0.10 (0.045,0.140)
$\hat{\phi}_0$	0.02 (0.018,0.021)	0.02 (0.018,0.021))
$\hat{\phi}_1$	0.02 (0.014,0.025)	0.02 (0.014,0.025)
$\hat{\phi}_2$	0.05 (0.033,0.085)	0.05 (0.030,0.100)
$\hat{\beta}$	0.94 (0.667,1.369)	0.95 (0.694,1.420)
$\hat{\gamma}$	2.58 (1.684,4.677)	2.64 (1.474,5.554)
LOF	$X^2 = 1.854$, $p = 0.396$	$X^2 = 1.833$, $p = 0.176$
	Recessive disease model	
$\hat{\varepsilon}$	-	0.10 (0.050,0.137)
$\hat{\phi}_0$	0.02 (0.018,0.020)	0.02 (0.018,0.020)
$\hat{\phi}_1$	-	-
$\hat{\phi}_2$	0.05 (0.033,0.085)	0.05 (0.029,0.099)
$\hat{\beta}$	-	-
$\hat{\gamma}$	2.58 (1.663,4.705)	2.67 (1.440,5.425)
LOF	$X^2 = 1.931$, $p = 0.587$	$X^2 = 1.896$, $p = 0.388$

The estimates are essentially unchanged when the stratification proportion is estimated in addition to the penetrances (Table 4.7). This model indicates a relative risk of 2.6 for those with two copies of the variant allele. The fact that $\hat{\beta}$ is approximately one indicates the genetic effect could be recessive.

Table 4.7 also shows the results when the recessive model was also fitted to the data with stratification proportion fixed or estimated. Likelihood ratio tests indicate no significant difference between the recessive and general models (ε fixed, p = 0.22; ε estimated, p = 0.19).

4.8.3 Sensitivity To K_P

To investigate the sensitivity of the fit of the model to the assumption that the overall prevalence of kidney stones is $K_P = 0.02$, this value was allowed to vary (Table 4.8). The penetrance estimates increase appropriately with K_P but the estimate of the stratification proportion, ε, is unaffected. There is no significant LOF for any of the choices of the prevalence.

Table 4.8: The effect of changing K_P on the estimates obtained from the R990G data.

	Estimates (SD)				LOF	
K_P	$\hat{\varepsilon}$	$\hat{\phi}_0$	$\hat{\phi}_1$	$\hat{\phi}_2$	χ^2	p
0.005	0.0945 (0.0300)	0.0049 (0.0002)	0.0046 (0.0008)	0.0132 (0.0029)	1.6411	0.2002
0.01	0.0953 (0.0299)	0.0097 (0.0005)	0.0092 (0.0016)	0.0261 (0.0057)	1.7036	0.1918
0.02	0.0967 (0.0295)	0.0194 (0.0009)	0.0184 (0.0032)	0.0513 (0.0113)	1.8326	0.1758
0.05	0.101 (0.0284)	0.0487 (0.0024)	0.0462 (0.0078)	0.1215 (0.0272)	2.2493	0.1337
0.1	0.1076 (0.0272)	0.0975 (0.0045)	0.0937 (0.015)	0.2215 (0.0499)	3.024	0.082

4.9 Discussion

Departure from Hardy-Weinberg equilibrium may indicate genotyping error, population stratification, selection bias, or some combination thereof. HWD could also indicate association with a disease in affected patients (Nielsen et al., 1998; Lee, 2003) or genetic association (Wittke-Thompson et al., 2005). It is therefore, important to investigate the reason for departure before excluding the loci exhibiting HWD from association studies.

This chapter extends the heterogeneous model described by Wittke-Thompson et al. (2005) to determine if the HWD observed in a data set from a stratified population can be explained by genetic association and stratification. Applying the extended stratification model in Section 4.8 to the R990G SNP of the CASR gene, in a cohort of ethnically and clinically heterogeneous kidney stone formers and a cohort of self-reporting Caucasian and Asian Canadians, it was found that the HWD in the data was adequately explained by a recessive genetic association and a stratification proportion of 10%, consistent with the population of Toronto.

The HWD coefficients for cases and controls under the extended model are in general combinations of D_S, the expression for a stratified population, and the coefficients for an unstratified population (Wittke-Thompson et al., 2005). In all cases, the magnitude of the coefficient increases with the stratification proportion as well as the difference between the minor allele frequencies and the genetic association. For most models, the HWD coefficient can be either positive or negative in both cases and controls, but is always non-negative for the recessive and multiplicative models in cases and for the dominant model in controls.

For the extended model, it is necessary to know the ethnicity and the minor allele frequency of at least one of the two strata. The allele frequencies can be estimated from previous studies or from the HapMap. If one has an idea about one stratum minor allele frequency, the frequency for the other stratum could be estimated from our method but then there would be no degrees of freedom left for the LOF test.

The proposed method can be extended easily to more than two strata or to more than two alleles. However, one would need to make some additional assumptions because the degrees of freedom in the data do not change from four.

Chapter 5

Conclusions

5.1 Summary

Genetic association studies involve determining whether a genetic variant is associated with a disease. If an allele is associated with disease, it will occur more often than expected by chance in affected individuals. Hardy-Weinberg equilibrium (HWE) holds in a population if the allele frequency remains constant from one generation to the next. HWE in a sample requires a large population, random sampling, no migration, mutation or selection and random mating. The Hardy-Weinberg disequilibrium coefficient, is defined as the difference in the observed and expected genotypic frequency for the genotype AA assuming HWE, or $D = P_{AA} - p_A^2$.

Departure from HWE (HWD) in a sample may indicate genotyping error, population stratification, selection bias, or some combination thereof. Therefore, loci exhibiting HWD are often excluded from association studies. However, it has also been shown that in case-control studies HWD can result from a genetic effect at the locus (Wittke-Thompson et al., 2005) and HWD at a marker locus can be interpreted as evidence for association with a disease (Nielsen et al., 1998; Lee, 2003). In an unpublished study in Toronto it was observed that cases were in HWE at a locus whereas their family members were in HWD. Wittke-Thompson et al. (2005) observed that the HWD coefficient for the multiplicative model is zero. It was therefore considered important to investigate HWD in relatives of affected individuals, and in particular to see whether the multiplicative model could be revealed.

In this thesis HWD coefficients were derived for affected individuals and their affected and unaffected relatives. A substantial HWD was found in dominant and recessive genetic models but HWD is only slightly nonzero for additive and multiplicative model. Methods (based on unconditional, partially conditional and fully conditional models) were also developed to test for association using data from affected individuals and their affected or unaffected relatives. Parameter estimates for these models can be obtained using maximum likelihood estimation methods, and estimates provide valuable information regarding the mode of inheritance of the disease. The methods were applied to 112 discordant sib pairs with Alzheimer's disease typed for the ApoE polymorphism (Boehnke and Langefeld, 1998). A significant association was observed between the ε_4 ApoE allele and Alzheimer's disease, which is consistent with the results obtained by Boehnke and Langefeld (1998) and Yan et al. (2009). The power of the fully conditional test was larger than the McNemar test and it also gives information regarding the mode of inheritance of the disease in addition to the degree of association.

Case-control studies may indicate spurious association with a marker locus in a stratified population. Methods were developed in Chapter 4 to determine if the HWD observed in a data set from a stratified population can be explained by both genetic association and stratification. Parameter estimates for these models can be obtained using maximum likelihood estimation methods, and provide valuable information regarding the mode of inheritance of the disease. Applying the model to the R990G SNP of the CASR gene, it was found that the HWD was adequately explained by a recessive genetic association and a stratification proportion of 10%, consistent with the population of Toronto.

5.2 Future Work

The methods developed in this thesis to deal with relative pairs and stratification are for a single biallelic disease susceptibility locus (DSL). In some complex diseases there might be a combined effect of two or more loci. It would, therefore, be of interest to extend the single locus model to two or more loci to model the joint effects. One of the ways to model multiple loci as a single locus with multiple alleles. Therefore extending the models to multi-allelic loci is also of interest.

Often association studies involve testing at a marker loci that is in linkage disequilibrium with the disease locus rather than at the locus itself. The models described in this thesis are for candidate genes, which are suspected of being associated with the disease. It would be useful to investigate model for genotypic counts at a marker locus, which would depend on the penetrances at the true DSL and the recombination fraction between the DSL and marker.

The tests for association and stratification described in the thesis are for two strata, and it would be of interest to extend the model to more than two strata. An alternative method, rather than an extension of heterogeneous disease model, may be required since the number of parameters increases with an increase in the number of strata. However, the degrees of freedom in the data do not change making it difficult or impossible to estimate all the parameters.

5.2.1 Modelling Penetrance As A Function Of A Continuous Variable

Another avenue for further research is to allow the penetrance, in the heterogeneous model to depend on a continuous covariate. Initial investigation of the topic is described below.

In association studies, genetic information is used to find association between genes and disease. Many diseases like obesity, coronary heart disease, diabetes, are related to age. The formation of kidney stones is related to a measure of kidney function like serum creatinine. Age dependent penetrance is also of interest for diseases like Alzheimer's, diabetes, cancer, Huntington's disease (Cupples et al., 1989), manic-depressive illness (Crowe and Smouse, 1977), motor neuron disease (Aggarwal and Nicholson, 2005), leprosy (Abel et al., 1989), facioscapulohumeral muscular dystrophy (Lunt et al., 1989).

Several methods have been proposed for the estimation of age-of-onset of disease from population data. Risch (1983) used a maximum likelihood method (MLE) that gives unbiased and efficient estimates of the morbidity risk when the prior age-of-onset distribution is known. Some other methods estimate probabilities of a cumulative age-of-onset distribution from the data (Heimbuch et al., 1980) or estimate penetrances within each age group (Debniak et al., 2005). Life-table and Cox proportional hazards regression or survival analysis are often used to find a relationship between various mutations and age-at-diagnosis methods have also been proposed (Chase et al., 1983; Chidambaram et al., 1988; Meyer and Eaves, 1988; Al-Mulla et al., 2009; Aggarwal and Nicholson, 2005; F-de Misa et al., 2008; Sturt, 1986; Crowe and Smouse, 1977; Strahan et al., 1983; Risch, 1983). Different survivorship functions, the Weibull, exponential, gamma, and log-normal distributions have been used to describe age-of-onset.

The survival methods usually assume that age-of-onset is independent of the genotype of the affected individual and uncorrelated between relatives. However, diseases like Huntington's disease, schizophrenia, and depression show a significant age-of-onset correlation between family members and early age-of-onset of breast cancer, alcoholism, affective disorders, and Alzheimer's dementia has been associated with an increased risk in relatives. The age-of-onset may not only be correlated among relatives, but it may also be correlated with an individual's inherited liability to an illness. Some modifications like using survival time models with nonproportional hazard functions, allowed for the effect of a proband's age-of-onset or the age-of-onset of a first degree relative (Wickramaratne et al., 1986) and pedigree analysis to include genotype-dependent ages of onset (Elston, 1973; Crowe and Smouse, 1977) have also been proposed. Meyer and Eaves (1988) developed a model to explain both age-of-onset correlations and distributions within the survival analysis framework. They specify genetic heterogeneity in one of the parameters of the gamma survival distribution by allowing it to be a function of liability and estimate the parameters using MLE.

Some of the methods for case-control studies, where both disease history and covariate status are available on relatives, were based on likelihood (Whittemore, 1995) and estimating equations (Zhao et al., 1998) methodology and involve estimation of disease covariate association and magnitude of familial aggregation. Different likelihood and pseudo-likelihood methodologies in survival analysis were also developed to account for censoring and age-at-onset information of disease (Li et al., 1998; Hsu et al., 1999; Shih and Chatterjee, 2002). These methods are often referred to as case-control family data designs.

Wacholder et al. (1998) proposed a kin-cohort design in which randomly sampled case and control groups are genotyped and penetrances are estimated based on the history of disease of their first degree relatives. The method uses survival analysis where the disease status and age-at-onset of the relatives are treated as outcome variables. They used the fact that the survival distribution for first-degree relatives of probands who carried (or did not carry) a mutation was a mixture of survival distributions for carriers and non-carriers, with mixing proportions about 50:50 (or 0:100) for rare mutations to estimate the disease survival distribution. The mixing proportions are functions of the allele frequency q. One of the limitations of the method is that for small samples the estimates of q are not necessarily monotone. Chatterjee and Wacholder (2001) developed a marginal-likelihood approach for analysing kin-cohort data, that allows for the possibility of obtaining non-monotone estimates of age-specific cumulative risk function.

Gail et al. (1999) and Gail et al. (1999) refer to the kin-cohort design as the genotyped-proband design. They assume that initial family member (cases or controls) is genotyped and is selected at random, conditional on disease status. They derived the likelihood of the genotypes and disease history data of the relatives conditional on the disease status of the probands which assume that all familial outcomes are conditionally independent given the individual's genotypes. They showed that this likelihood can be factored as the product of a case-control likelihood of the genotype of probands given their disease status and a kin-cohort likelihood for the relative's disease outcome data given the genotype of the proband. Any violation of the assumption of no residual familial aggregation can lead to biased parameter estimates. Some extensions of these methods have been proposed to estimate the parameters for survival models. Moore et al. (2001) proposed a pseudo-likelihood method since full maximum likelihood estimation using the true likelihood of the data can be computationally challenging. However, the pseudo-likelihood approach is inefficient as it is not possible to extract the relative risk information from case-control data. A parametric method, Proband's phenotype Exclusion Likelihood (PEL) (Alarcon et al., 2009) and a nonparametric method, Index Discarding EuclideAn Likelihood (IDEAL) (Alarcon et al., 2009) have also been proposed to estimate the penetrance functions based on survival analysis. These methods also correct for the ascertainment bias.

Langbehn et al. (2004), in the context of Huntington's disease (HD), developed a parametric survival model that incorporates information from those with onset and those still at risk, predicting risk of onset for any person at risk of the disease at any age. The method involves finding a distribution family that gives a close fit to all of the observed (non-parametric) survival distributions from the individual CAG trinucleotide repeats followed by finding classes of mathematical functions that adequately described the relationship between CAG (the mutation associated with clinical manifestations of HD) and both the mean and dispersion of the age-of-onset. These functions combined with the parametric distribution from the final parametric model to predict the age-specific probability of onset. They found that the logistic distribution had the best average fit to the non-parametric survival curves across CAG lengths and the exponential function provided excellent fits to both the mean and variance of the age-of-onset.

All the above mentioned methods are based on survival analysis and yield biased hazard function estimators due to the sampling bias, over-representation of cases, and the residual dependency among relatives. In order to overcome these issues, Chatterjee et al. (2006) proposed an extension and combination of the above mentioned methods, kin-cohort, genotype-proband and case-control family data. It models the joint distribution of failure times of family members. The method extends the data available in the kin-cohort design to include covariate information and genotypes of the first-degree relatives of case and control subjects. The methodology combines information on relative risk parameters from the kin-cohort data of relatives and case-control data of participants. It also estimates the baseline risk and familial aggregation parameters using the kin-cohort data of the relatives.

More recently, a frailty-model-based approach has been proposed to estimate the hazard function from two-phase case-control data with family history information (Chen et al., 2009). It accounts for the shared risk among family members that is not accounted for by observed risk factors. In the first phase, a random sample of cases and controls (probands) is obtained from a population. The cases and controls are stratified based on certain aspects of variables collected in the first phase. In the second phase, a random subset of cases and controls from each stratum are selected for genotyping. All strata have representative samples, to ensure a consistent estimation of the odds ratios from two-phase data. The dependent failure outcomes of family members was described by a shared gamma frailty with the conditional proportional hazards model. The censoring times are assumed to be independent of the failure time and noninformative of the frailty conditional on the frailty and the covariates. Frailty was also assumed to be independent of the observed covariates. The model is based on likelihood that conditions on the proband's survival time and allows for residual dependency via a frailty. The estimates of the regression coefficients, nonparametric baseline hazard function, and dependence parameter are obtained using the expectation conditional maximization (ECM) algorithm which is a variation of the expectation-maximization algorithm. Some of the strengths of the model are that the method is robust against ascertainment biases and the residual dependency estimates can shed light on whether one or more candidate genes or other shared environmental risk factors may contribute to diseases. Extension of the method accommodate missing genotypes in family members and a two-phase case-control sampling design were also described.

The heterogeneous model described in this thesis can be extended to include dependence on a continuous variable. The probability of a genotype $g, g = 0, 1, 2$ given the disease status, $d = D$ (case) or C (control), and a continuous variable, X, can be expressed in terms of the penetrance and minor allele frequency using Bayes rule

$$P_{gd}(X) = P(g|d; X) = \frac{P(g \cap d; X)}{P(d; X)} = \frac{P(d|g; X)P_g}{K_P(X)},$$

where, $g = 0, 1, 2$ is the number of variant alleles, $K_P(X)$ is the overall prevalence of the disease,

$$K_P(X) = \phi_0(X)(1-q)^2 + 2\phi_1(X)q(1-q) + \phi_2(X)q^2, \tag{5.1}$$

and q is the minor allele frequency. The penetrance for genotype g as a function of X, $\phi_g(X)$ can be modelled using a logistic function

$$\phi_g(X) = \frac{1}{1 + e^{-(\beta_{0g} + \beta_{1g}X)}}, \tag{5.2}$$

where $g = 0, 1, 2$. The genotypic probabilities for cases and controls can be summarized as in Table 5.1

Table 5.1: The genotypic frequencies for cases and controls

	Genotype 0	 1	 2
Cases	$\frac{\phi_0(X)(1-q)^2}{K_P(X)}$	$\frac{2\phi_1(X)q(1-q)}{K_P(X)}$	$\frac{\phi_2(X)q^2}{K_P(X)}$
Controls	$\frac{[1-\phi_0(X)](1-q)^2}{1-K_P(X)}$	$\frac{2[1-\phi_1(X)]q(1-q)}{1-K_P(X)}$	$\frac{[1-\phi_2(X)]q^2}{1-K_P(X)}$

where the penetrances are given in (5.2) and the disease prevelance by (5.1).

The genotypic frequencies (Table 5.1) for cases and controls form a likelihood function

$$L = \prod_{i=1}^{n} P_{gd,i}(X_i), \tag{5.3}$$

where n is the total number of cases and controls and $P_{gd,i}(X_i)$ is the genotypic probability for subject i, who has genotype g, disease status d and variable X_i.

There are seven parameters in the model - the minor allele frequency, q, the three intercepts, β_{00}, β_{01}, β_{02}, and three slopes β_{10}, β_{11} and β_{12}. The minor allele frequency could be considered known from previous studies or from the HapMap. The βs can be estimated by maximizing the likelihood function numerically and approximate standard errors can be obtained by using the nonparametric bootstrap.

The disease prevalence K_P is assumed to be known at some value of X such as the mean $\bar{X}$ from other sources. This allows one of the parameters to be evaluated from the others using

$$K_P(\bar{X}) = (1-q)^2\phi_0(\bar{X}) + 2q(1-q)\phi_1(\bar{X}) + q^2\phi_2(\bar{X}),$$

or

$$K_P(\bar{X}) = \frac{(1-q)^2}{1+e^{-(\beta_{00}+\beta_{10}\bar{X})}} + \frac{2q(1-q)}{1+e^{-(\beta_{01}+\beta_{11}\bar{X})}} + \frac{q^2}{1+e^{-(\beta_{02}+\beta_{12}\bar{X})}},$$

For example, the intercept for the baseline prevalence, β_{00} can be written as a function of $K_P(\bar{X})$, β_{01}, β_{11}, β_{02}, β_{12} and q as

$$\beta_{00} = -\beta_{10}\bar{X} - log\Big(\frac{(1-q)^2}{K_P(\bar{X}) - \frac{2q(1-q)}{1+e^{-(\beta_{01}+\beta_{11}\bar{X})}} - \frac{q^2}{1+e^{-(\beta_{02}+\beta_{12}\bar{X})}}} - 1\Big), \qquad (5.4)$$

The remaining parameters in the model are β_{10}, β_{01}, β_{11}, β_{02}, β_{12} and q. Once the other parameters have been estimated, β_{00} can be obtained from equation (5.4). The parameters can be estimated by maximizing the likelihood function numerically and approximate standard errors can be obtained by evaluating the inverse of information matrix or by using the nonparametric bootstrap. As is discussed below, some choices of the parameters cause the term in large parentheses in (5.4) to be negative, which creates problems for the numerical maximization.

It is possible to make inferences about the size and type of genetic effect and the extent of effect of the variable X on penetrance by comparing the fit of a general model with that of a reduced model with fewer effects, using a likelihood ratio test (LRT). To assess whether the penetrances depend on X, the hypotheses of interest are

$$\begin{array}{ll} H_0: & \beta_{10} = \beta_{11} = \beta_{12} = 0 \\ H_a: & \text{at least one } \beta_{1i} \neq 0 \text{ for } i = 0, 1, 2. \end{array}$$

The alternative hypothesis has six parameters and null has three so the likelihood ratio test statistic, Λ, is asymptotically distributed as χ^2 with three degrees of freedom.

Specific genetic models give relationships between different penetrances and this is also true when they are modelled using a continuous variable.

In the dominant model, the homozygote and heterozygote relative risks are equal, implying the penetrances for the genotype Aa and aa to be the same i.e., $\phi_2 = \phi_1$.

When the genotypic frequencies are modelled as a logistic function of the continuous variable,

$$\frac{1}{1+e^{-(\beta_{02}+\beta_{12}X)}} = \frac{1}{1+e^{-(\beta_{01}+\beta_{11}X)}}$$

or

$$\beta_{02} + \beta_{12}X = \beta_{01} + \beta_{11}X$$

Therefore, $\beta_{01} = \beta_{02}$ and $\beta_{11} = \beta_{12}$, i.e., the intercepts for one or two variant alleles are equal and the slopes for one or two variant alleles are equal for the dominant model.

Under a recessive model, having two copies of the variant allele leads to an increased risk of disease susceptibility and the heterozygote relative risk is one, i.e., $\phi_0 = \phi_1$.

When the genotypic frequencies are modelled as a logistic function of the continuous variable,

$$\frac{1}{1+e^{-(\beta_{00}+\beta_{10}X)}} = \frac{1}{1+e^{-(\beta_{01}+\beta_{11}X)}}$$

or

$$\beta_{00} + \beta_{10}X = \beta_{01} + \beta_{11}X$$

Therefore, $\beta_{00} = \beta_{01}$ and $\beta_{10} = \beta_{11}$, i.e., the intercepts for none or one variant alleles are equal and the slopes for none or one variant alleles are equal for the recessive model.

In an additive disease model, the difference between the homozygote and heterozygote penetrance is the same as the difference between the baseline and homozygote penetrance, i.e., $\phi_2 - \phi_1 = \phi_1 - \phi_0$ or $\phi_2 = 2\phi_1 - \phi_0$ or

$$\phi_1 = \frac{\phi_0 + \phi_2}{2}.$$

When the genotypic frequencies are modelled as a logistic function of the continuous variable, the additive model implies

$$\frac{1}{1 + e^{-(\beta_{01}+\beta_{11}X)}} = \frac{1}{2}\Big(\frac{1}{1 + e^{-(\beta_{00}+\beta_{10}X)}} + \frac{1}{1 + e^{-(\beta_{02}+\beta_{12}X)}}\Big),$$

which does not give a simple relationship among the intercepts and slopes.

In a multiplicative model, the homozygote relative risk is the square of the heterozygote relative risk, i.e. $\phi_2 = \phi_1^2/\phi_0$ or $\phi_1 = \sqrt{\phi_0\phi_2}$.

When the genotypic frequencies are modelled as a logistic function of the continuous variable, the multiplicative model implies

$$\Big(\frac{1}{1 + e^{-(\beta_{01}+\beta_{11}X)}}\Big)^2 = \Big(\frac{1}{1 + e^{-(\beta_{00}+\beta_{10}X)}}\Big)\Big(\frac{1}{1 + e^{-(\beta_{02}+\beta_{12}X)}}\Big),$$

which does not simplify further.

Some preliminary simulations have been attempted for this model using R software and the built-in function “optim” to obtain the maximum likelihood estimates.

Level and power of the proposed hypothesis test for age-related penetrance was approximated by simulation based on an assumed disease prevalence of $K_P = 0.02$ at mean age 45 years. For the simulations, 1000 replicated data sets were used for each combination of two different values of the minor allele frequency q (0.1, 0.2), two different sample sizes for cases and controls, n (300, 1000). The variable X was considered as age, and assumed to be distributed as normal with mean 45 years and SD of 20 years. The level and power are approximated using the proportion of hypotheses rejected when the data are generated under the appropriate null and alternative hypotheses, respectively. The data were simulated using $\phi_0 = 0.018$, $\phi_1 = 0.022$ and $\phi_2 = 0.166$ (estimates of penetrances for kidney stones data at the R990G locus as described in Section 4.8).

In order to assess the level of the test, the data were generated assuming the intercepts, β_{0i}s to be $\beta_{0i} = logit(\phi_i)$, $i = 0, 1, 2$, giving $\beta_{00} = -3.9768$, $\beta_{01} = -3.8045$, $\beta_{02} = -1.6135$, and the slopes, β_{1i}'s as, $\beta_{10} = \beta_{11} = \beta_{12} = 0$.

In order to assess the power of the test, the data were generated assuming the the intercepts, β_{0i}'s as $\beta_{0i} = logit(\phi_i)$, $i = 0, 1, 2$, giving $\beta_{00} = -3.9768$, $\beta_{01} = -3.8045$, $\beta_{02} = -1.6135$, no effect of age on the AA genotype, i.e., $\beta_{10} = 0$ and small but increasing effect of the number of variants, i.e., $\beta_{11} = 0.01$ and $\beta_{12} = 0.02$.

For some combinations of the values for allele frequencies, slope and intercept, the term inside the parentheses in (5.4) becomes negative. However, when the slopes are constrained to be within ± 0.25 for $q = 0.1$ or $q = 0.2$, the term stays positive and the method converges.

Table 5.2 gives the type I error and power obtained by simulation for test of effect of the variable X, age, on penetrances at $\alpha = 0.05$. The level of the test decreases with increase in minor allele frequency, q. The power increases with the sample size, n and the minor allele frequency q. The level is close to 0.05 for $q = 0.1$ when q is assumed known, otherwise it is less than 0.05 and smaller for $q = 0.2$ than $q = 0.1$.

Table 5.2: Type I errors and power obtained by simulation for test of the variable X, age on penetrances at $\alpha = 0.05$

	q estimated		q fixed	
q	$n = 300$	$n = 1000$	$n = 300$	$n = 1000$
	Level of the test			
0.1	0.044	0.039	0.051	0.047
0.2	0.030	0.030	0.030	0.031
	Power of the test			
0.1	0.240	0.813	0.289	0.800
0.2	0.276	0.872	0.270	0.883

When the data are generated under the null hypothesis, Table 5.3 gives the mean and standard error of the parameter estimates, and the p-value for significance of the bias when q is also estimated and Table 5.4 when q is not estimated. An examination of the distribution of the parameter estimates reveals that the distribution of the estimates has a sharp peak and that they have larger variance when estimated under the alternative hypothesis (Figures G.1, G.2, G.3, G.4 in Appendix G and Figures G.5, G.6, G.7, G.8 in Appendix G). It also shows that there is a significant bias in the estimates when data are generated and fitted under the null hypothesis for estimation of type I error.

Table 5.3: Mean ± standard error of parameter estimates for the simulations when q is also estimated under H_0.

	True	$n = 300$		$n = 1000$	
	Value	Mean ± SE	p	Mean ± SE	p
H_0					
β_{00}	-3.9768	-4.0021 ± 0.0026	< 0.0001	-4.0028 ± 0.0008	< 0.0001
β_{01}	-3.8045	-3.8315 ± 0.0091	0.0030	-3.8240 ± 0.0028	< 0.0001
β_{02}	-1.6135	-1.6237 ± 0.0220	0.6438	-1.6307 ± 0.0063	0.0067
q	0.1	0.1000 ± 0.0007	1.0000	0.0996 ± 0.0002	0.0447
H_a					
β_{00}	-3.9768	-3.8197 ± 0.1308	0.2296	-3.8774 ± 0.0379	0.0087
β_{10}	0	-0.0043 ± 0.0030	0.1521	-0.0029 ± 0.0009	0.0007
β_{01}	-3.8045	-3.6572 ± 0.1314	0.2622	-3.7017 ± 0.0379	0.0066
β_{11}	0	-0.0043 ± 0.0030	0.1544	-0.0029 ± 0.0009	0.0007
β_{02}	-1.6135	-1.2907 ± 0.1707	0.0586	-1.4575 ± 0.0477	0.0011
β_{12}	0	-0.0065 ± 0.0039	0.0929	-0.0035 ± 0.0011	0.0011
q	0.1	0.1005 ± 0.0007	0.4814	0.0998 ± 0.0002	0.3230
H_0					
β_{00}	-3.9768	-4.2251 ± 0.0043	< 0.0001	-4.2248 ± 0.0013	< 0.0001
β_{01}	-3.8045	-4.0507 ± 0.0068	< 0.0001	-4.0429 ± 0.0020	< 0.0001
β_{02}	-1.6135	-1.8681 ± 0.0124	< 0.0001	-1.8732 ± 0.0037	< 0.0001
q	0.2	0.1989 ± 0.0009	0.2249	0.1986 ± 0.0003	< 0.0001
H_a					
β_{00}	-3.9768	-3.9386 ± 0.0761	0.6154	-3.9416 ± 0.0254	0.1655
β_{10}	0	-0.0062 ± 0.0016	0.0001	-0.0063 ± 0.0006	< 0.0001
β_{01}	-3.8045	-3.7689 ± 0.0745	0.6328	-3.7565 ± 0.0251	0.0555
β_{11}	0	-0.0062 ± 0.0016	0.0001	-0.0065 ± 0.0006	< 0.0001
β_{02}	-1.6135	-1.5178 ± 0.0898	0.2865	-1.5364 ± 0.0297	0.0095
β_{12}	0	-0.0072 ± 0.0019	0.0002	-0.0074 ± 0.0007	< 0.0001
q	0.2	0.1995 ± 0.0009	0.5836	0.1989 ± 0.0003	0.0001

Table 5.4: Mean ± standard error of parameter estimates for the simulations when q is not estimated under H_0.

	True	$n = 300$		$n = 1000$	
	Value	Mean ± SE	p	Mean ± SE	p
H_0, $q = 0.1$					
β_{00}	-3.9768	-4.0012 ± 0.0021	< 0.0001	-4.0006 ± 0.0006	< 0.0001
β_{01}	-3.8045	-3.8419 ± 0.0071	< 0.0001	-3.8347 ± 0.0021	< 0.0001
β_{02}	-1.6135	-1.6524 ± 0.0134	0.0037	-1.6471 ± 0.0041	< 0.0001
H_a, $q = 0.1$					
β_{00}	-3.9768	-3.7503 ± 0.1316	0.0853	-3.9071 ± 0.0437	0.1108
β_{10}	0	-0.0057 ± 0.0029	0.0506	-0.0021 ± 0.0010	0.0295
β_{01}	-3.8045	-3.5812 ± 0.1320	0.0907	-3.7272 ± 0.0440	0.0792
β_{11}	0	-0.006 ± 0.0029	0.0404	-0.0024 ± 0.0010	0.0137
β_{02}	-1.6135	-1.2657 ± 0.1710	0.0419	-1.4798 ± 0.0551	0.0152
β_{12}	0	-0.0071 ± 0.0038	0.0589	-0.0030 ± 0.0012	0.0135
H_0, $q = 0.2$					
β_{00}	-3.9768	-4.2202 ± 0.0035	< 0.0001	-4.2216 ± 0.0011	< 0.0001
β_{01}	-3.8045	-4.0511 ± 0.0053	< 0.0001	-4.0484 ± 0.0016	< 0.0001
β_{02}	-1.6135	-1.9012 ± 0.0064	< 0.0001	-1.8928 ± 0.0019	< 0.0001
H_a, $q = 0.2$					
β_{00}	-3.9768	-3.8514 ± 0.0757	0.0976	-3.9732 ± 0.0253	0.8857
β_{10}	0	-0.0087 ± 0.0017	< 0.0001	-0.0055 ± 0.0006	< 0.0001
β_{01}	-3.8045	-3.6835 ± 0.0757	0.1098	-3.7944 ± 0.0250	0.6872
β_{11}	0	-0.0087 ± 0.0017	< 0.0001	-0.0056 ± 0.0006	< 0.0001
β_{02}	-1.6135	-1.4232 ± 0.0912	0.0368	-1.5965 ± 0.0300	0.5694
β_{12}	0	-0.0106 ± 0.0021	< 0.0001	-0.0063 ± 0.0007	< 0.0001

When the data are generated under the alternative hypothesis to compute power, Table 5.5 gives the mean and standard deviation of the estimates when q is also estimated and Table 5.6 describes the distribution of the estimates when q is not estimated. An examination of the distribution of the estimates shows them to have a sharp peak and to have a larger variance when fitted under the alternative hypothesis. There is significant bias in the estimates when the model is fitted under the alternative hypothesis for the estimation of power. (Figures H.1, H.2, H.3, H.4 in Appendix H and Figures H.5, H.6, H.7, H.8 in Appendix H)

Table 5.5: Mean ± standard error of parameter estimates for the simulations when q is also estimated under H_a.

	True	$n = 300$		$n = 1000$	
	Value	Mean ± SE	p	Mean ± SE	p
H_0					
β_{00}	-3.9768	-4.1834 ± 0.0030	< 0.0001	-4.1776 ± 0.0009	< 0.0001
β_{01}	-3.8045	-3.5329 ± 0.0086	< 0.0001	-3.5421 ± 0.0026	< 0.0001
β_{02}	-1.6135	-0.9230 ± 0.0228	< 0.0001	-0.9535 ± 0.0066	< 0.0001
q	0.1	0.0992 ± 0.0007	0.2363	0.0991 ± 0.0002	< 0.0001
H_a					
β_{00}	-3.9768	-4.0579 ± 0.0739	0.2725	-4.0145 ± 0.0225	0.0935
β_{10}	0.0	-0.0028 ± 0.0016	0.0746	-0.0036 ± 0.0005	< 0.0001
β_{01}	-3.8045	-3.8846 ± 0.0744	0.2814	-3.8418 ± 0.0228	0.1022
β_{11}	0.01	0.0073 ± 0.0016	0.0879	0.0065 ± 0.0005	< 0.0001
β_{02}	-1.6135	-1.6995 ± 0.1143	0.4520	-1.5974 ± 0.0325	0.6189
β_{12}	0.02	0.0176 ± 0.0025	0.3382	0.0144 ± 0.0007	< 0.0001
q	0.1	0.0994 ± 0.0007	0.3825	0.0991 ± 0.0002	< 0.0001
H_0					
β_{00}	-3.9768	-4.5706 ± 0.0051	< 0.0001	-4.5663 ± 0.0016	< 0.0001
β_{01}	-3.8045	-3.9250 ± 0.0062	< 0.0001	-3.9290 ± 0.0019	< 0.0001
β_{02}	-1.6135	-1.4149 ± 0.0127	< 0.0001	-1.4299 ± 0.0037	< 0.0001
q	0.2000	0.1935 ± 0.0009	< 0.0001	0.1943 ± 0.0003	< 0.0001
H_a					
β_{00}	-3.9768	-4.1840 ± 0.0752	0.0059	-4.3376 ± 0.0294	< 0.0001
β_{10}	0.0	-0.0090 ± 0.0017	< 0.0001	-0.0054 ± 0.0007	< 0.0001
β_{01}	-3.8045	-3.9785 ± 0.0726	0.0166	-4.1334 ± 0.0293	< 0.0001
β_{11}	0.01	0.0010 ± 0.0016	< 0.0001	0.0045 ± 0.0007	< 0.0001
β_{02}	-1.6135	-1.6733 ± 0.0934	0.5222	-1.8913 ± 0.0366	< 0.0001
β_{12}	0.02	0.0062 ± 0.0021	< 0.0001	0.0106 ± 0.0008	< 0.0001
q	0.2	0.1941 ± 0.0009	< 0.0001	0.1947 ± 0.0003	< 0.0001

Table 5.6: Mean and standard deviation of parameter estimates for the simulations when q is not estimated under H_a.

	True	$n = 300$		$n = 1000$	
	Value	Mean ± SE	p	Mean ± SE	p
H_0, $q = 0.1$					
β_{00}	-3.9768	-4.174 ± 0.0026	< 0.0001	-4.1721 ± 0.0008	< 0.00011
β_{01}	-3.8045	-3.5547 ± 0.0057	< 0.0001	-3.5569 ± 0.0018	< 0.0001
β_{02}	-1.6135	-1.0115 ± 0.0116	< 0.0001	-0.9995 ± 0.0035	< 0.0001
H_a, $q = 0.1$					
β_{00}	-3.9768	-3.9473 ± 0.0835	0.7235	-4.0753 ± 0.0283	0.0005
β_{10}	0.0	-0.0052 ± 0.0019	0.0053	-0.0022 ± 0.0006	0.0005
β_{01}	-3.8045	-3.8075 ± 0.0840	0.9713	-3.9011 ± 0.0288	0.0008
β_{11}	0.01	0.0054 ± 0.0019	0.0145	0.0076 ± 0.0006	0.0002
β_{02}	-1.6135	-1.6219 ± 0.1259	0.9471	-1.6972 ± 0.0400	0.0363
β_{12}	0.02	0.015 ± 0.0028	0.0784	0.0162 ± 0.0009	< 0.0001
H_0, $q = 0.2$					
β_{00}	-3.9768	-4.5708 ± 0.0047	< 0.0001	-4.5593 ± 0.0014	< 0.0001
β_{01}	-3.8045	-3.9438 ± 0.0051	< 0.0001	-3.951 ± 0.0016	< 0.0001
β_{02}	-1.6135	-1.499 ± 0.0052	< 0.0001	-1.4974 ± 0.0015	< 0.0001
H_a, $q = 0.2$					
β_{00}	-3.9768	-4.221 ± 0.0863	0.0047	-4.4306 ± 0.0330	< 0.00011
β_{10}	0.0	-0.0081 ± 0.0019	< 0.0001	-0.0032 ± 0.0007	< 0.0001
β_{01}	-3.8045	-4.0414 ± 0.0856	0.0057	-4.265 ± 0.0330	< 0.0001
β_{11}	0.01	0.0019 ± 0.0019	< 0.0001	0.0068 ± 0.0007	< 0.0001
β_{02}	-1.6135	-1.8107 ± 0.1090	0.0706	-2.0727 ± 0.0413	< 0.0001
β_{12}	0.02	0.0075 ± 0.0024	< 0.0001	0.0131 ± 0.0009	< 0.0001

To assess the distribution of the test statistic, Λ, for assessing dependence on age, the χ^2 Q-Q plot of the test statistic was plotted (Figure 5.1). The figure indicates that the LRT statistic approximately follows a χ^2 distribution with three degrees of freedom.

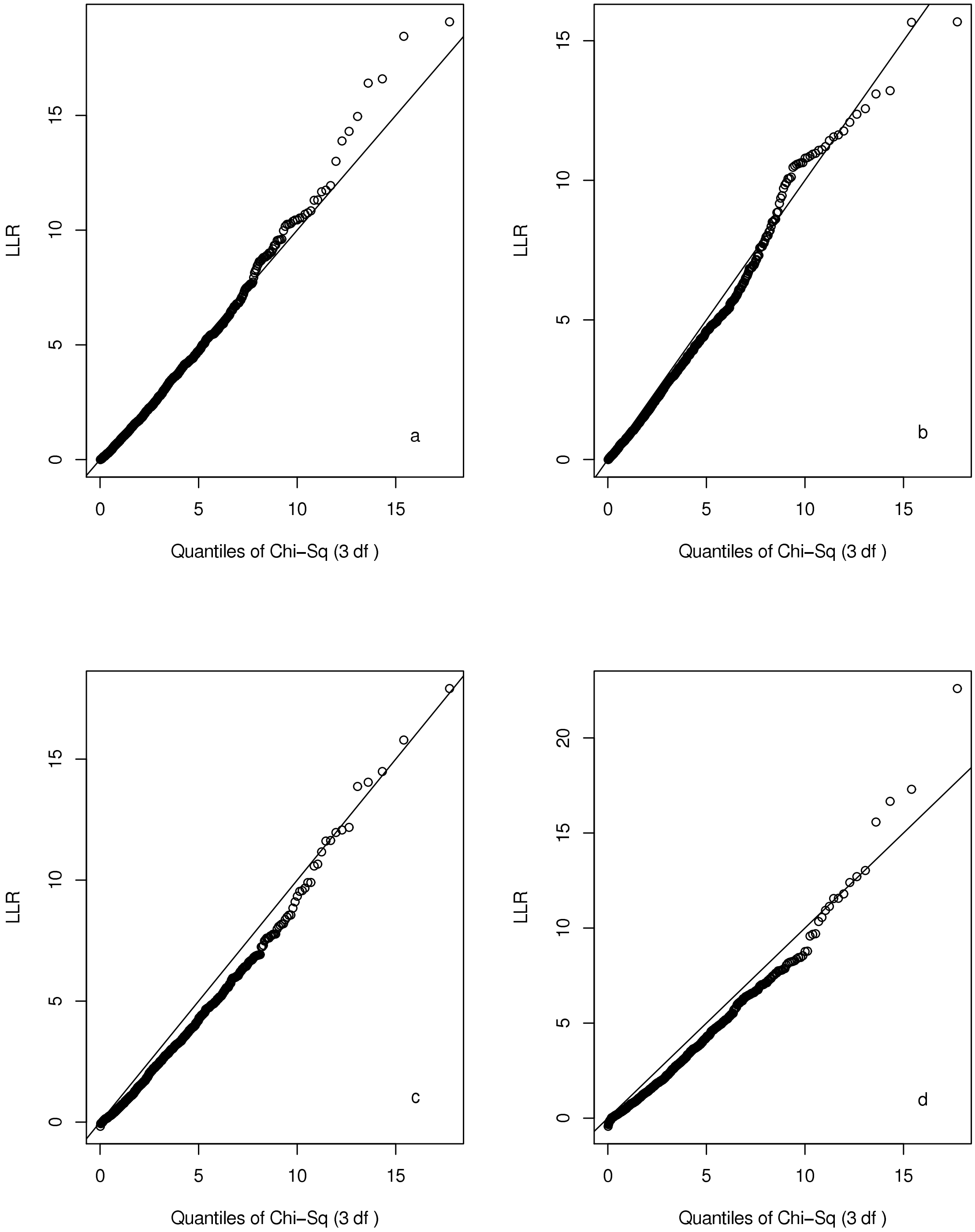


Figure 5.1: The Q-Q plots using the χ^2_3 distribution for the LLR statistics when q is not estimated
For $n = 300$ (Panels (a) and (c)) $n = 1000$ (Panels (b) and (d)) and $q = 0.1$ (Panels (a) and (b)) $q = 0.2$ (Panels (c) and (d)).

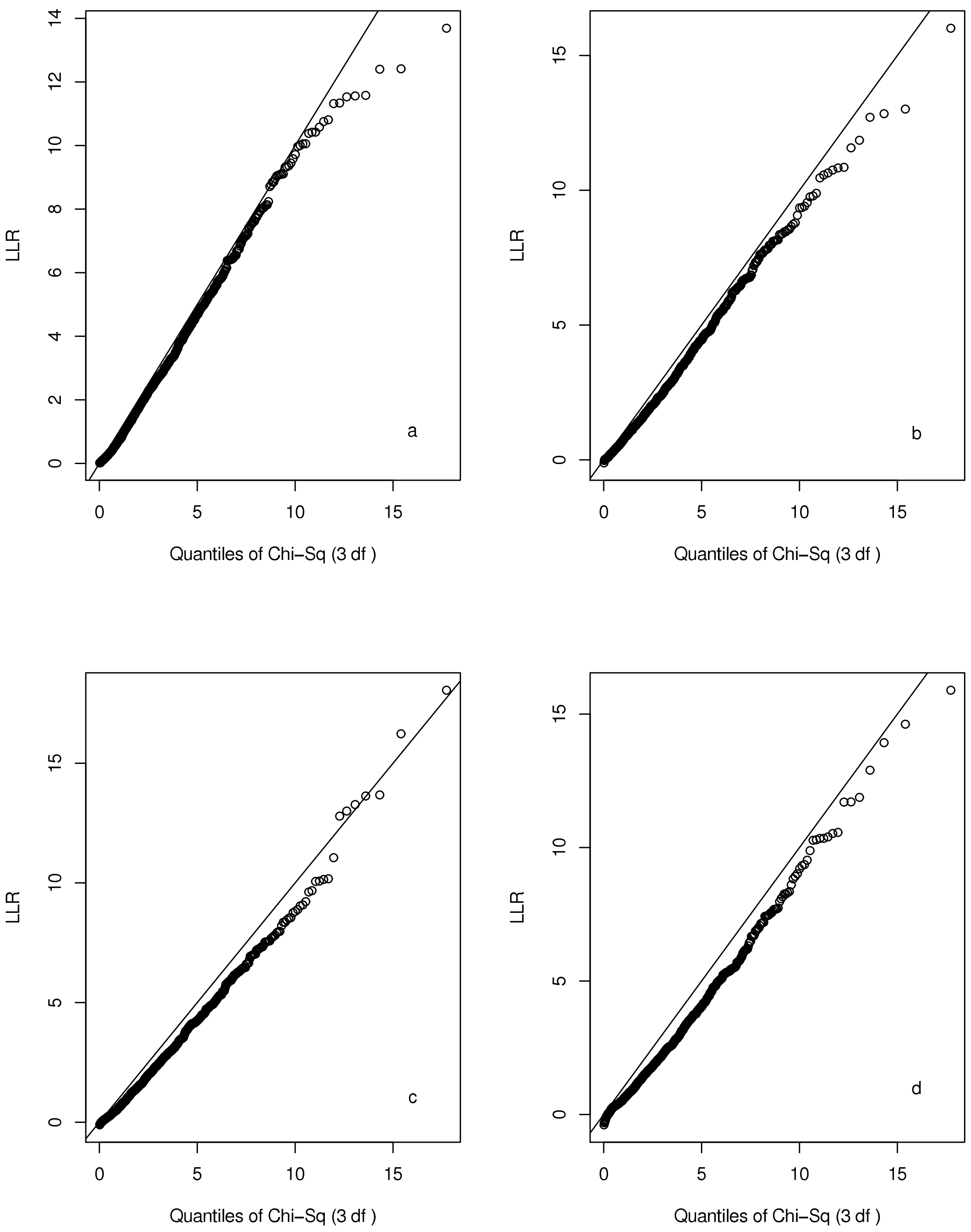


Figure 5.2: The Q-Q plots using the χ^2_3 distribution for the LLR statistics when q is also estimated
For $n = 300$ (Panels (a) and (c)) $n = 1000$ (Panels (b) and (d)) and $q = 0.1$ (Panels (a) and (b)) $q = 0.2$ (Panels (c) and (d)).

The evaluation of the likelihood function (5.3) requires calculation of the penetrance, prevalence of disease, and probabilities of the appropriate genotype for each person for some values of the intercepts β_0 and slopes β_1. These calculations are repeated until the maximum of the likelihood function is reached. This is computer intensive and it uses lot of computing resources, memory as well as time. When the allele frequency is large the term inside the log in (5.4) becomes negative for many choices of the parameter values.

In order to overcome the computation challenges, the possibility of modelling the penetrances as a linear function of the variable X was explored. In this case, the penetrances, ϕ_g can be expressed as

$$\phi_g = \beta_{0g} + \beta_{1g}X, \tag{5.5}$$

where, $g = 0, 1, 2$ is the number of variant alleles. The prevalence of the disease, $K_P(X)$ is

$$K_P(X) = (1-q)^2(\beta_{00} + \beta_{10}X) + 2q(1-q)(\beta_{01} + \beta_{11}X) + q^2(\beta_{02} + \beta_{12}X),$$

where q is the minor allele frequency. or

$$K_P(X) = [(1-q)^2\beta_{00} + 2q(1-q)\beta_{01} + q^2\beta_{02}] + [(1-q)^2\beta_{10} + 2q(1-q)\beta_{11} + q^2\beta_{12}]X. \tag{5.6}$$

Once again assuming the disease prevalence K_P to be known at mean $\bar{X}$, one of the parameters can be evaluated from the others. For example, the intercept for the baseline prevalence, β_{00} can be written as a function of $K_P(\bar{X})$, β_{01}, β_{11}, β_{02}, β_{12} and

q as

$$\beta_{00} = \frac{1}{(1-q)^2} K_P(\bar{X}) - 2q(1-q)\beta_{01} - q^2\beta_{02}] - [(1-q)^2\beta_{10} + 2q(1-q)\beta_{11} + q^2\beta_{12}]X, \quad (5.7)$$

which is much simpler expression than the one in (5.4). The remaining parameters in the model are β_{10}, β_{01}, β_{11}, β_{02}, β_{12} and q. Once the other parameters have been estimated, β_{00} can be obtained from equation (5.7). The parameters can be estimated by maximizing the likelihood function numerically and approximate standard errors can be obtained by evaluating the inverse of information matrix or by using the nonparametric bootstrap.

Even though (5.7) is a much simpler expression than (5.4), the method still requires the evaluation of the likelihood function, (5.3), which in turn involves calculating the penetrance, prevalence of disease and probabilities of the appropriate genotype for each person for some values of the β_0's and β_1's. The method is still computer intensive, using lot of computing resources, memory as well as time.

The results described above are preliminary and require further investigation. It is possible that a new formulation of the model or a new computing algorithm will produce better results.

Bibliography

Abel, L., A. Mallet, F. Demenais, G. E. Bonney, and D. C. Rao (1989). Modeling the age-of-onset function in segregation analysis: A causal scheme for leprosy so: Genetic epidemiology. *Genet Epidemiol 6*(4), 501 – 516.

Aggarwal, A. and G. Nicholson (2005). Age dependent penetrance of three different superoxide dismutase 1 (SOD 1) mutations. *Intern. J. Neuroscience 115*, 1119 – 1130.

Al-Mulla, F., J. M. Bland, D. Serratt, J. Miller, C. Chu, and G. T. Taylor (2009). Age-dependent penetrance of different germline mutations in the brca1 gene. *Journal of Clinical Pathology 62*, 350 – 356.

Alarcon, F., C. Bonaïti-Pelliè, and H. Harari-Kermadec (2009). A nonparametric method for penetrance function estimation. *Genet Epidemiol 33*(1), 38 – 44.

Alarcon, F., C. Bourgain, M. Gauthier-Villars, V. Plantè-Bordeneuve, D. Stoppa-Lyonnet, and C. Bonaïti-Pelliè (2009). PEL: an unbiased method for estimating age-dependent genetic disease risk from pedigree data unselected for family history. *Genet Epidemiol. 33*(5), 379 – 385.

Boehnke, M. and C. D. Langefeld (1998). Genetic association mapping based on discordant sib pairs: the discordant-alleles test. *Am J Hum Genet 62*, 950 – 961.

Chase, G. A., M. F. Folstein, J. C. Breitner, T. H. Beaty, and S. G. Self (1983). The use of life tables and survival analysis in testing genetic hypotheses, with an application to alzheimer's disease. *Am J Epidemiol.*

Chatterjee, N. and S. Wacholder (2001). A marginal likelihood approach for estimating penetrance from kincohort designs. *Biometrics 57*, 245 – 252.

Chatterjee, N., K. Zeymep, J. H. Shih, and M. H. Gail (2006). Casecontrol and case-only designs with genotype and family history data: Estimating relative risk, residual familial aggregation, and cumulative risk. *Biometrics 62*(1), 36 – 48.

Chen, l., L. Hsu, and K. Malone (2009). A frailty-model-based approach to estimating the age-dependent penetrance function of candidate genes using population-based case-control study designs: An application to data on the BRCA1 gene. *Biometrics 65*(4), 1105 – 1114.

Chidambaram, A., A. Chakravarti, R. E. Ferrell, and S. Iyengar (1988). Estimating the age-at-onset function using life-table methods. *Genet Epidemiol 5*, 255 – 263.

Cole, D. E. C., H. M. Trang, R. Vieth, V. D. Peltekova, A. Pierratos, B. Y. Wong, L. A. Rubin, and G. N. Hendy (1998). Calcium excretion is independently associated with the Vitamin D receptor (*VDR*) and calcium - sensing receptor (*CASR*) polymorphisms in a nephrolithiasis population. *Bone 23 Suppl1*, S248.

Crowe, R. R. and P. E. Smouse (1977). The genetic implications of age-dependent penetrance in manic-depressive illness. *J Psychiatr Res 13*, 273–285.

Cupples, L. A., N. C. Terrin, R. H. Myers, R. B. D'Agostino, and D. C. Rao (1989). Using survival methods to estimate age at onset distributions for genetic diseases with an application to huntington disease. *Genet Epidemiol 6*(2), 361 – 371.

Debniak, T., B. Gòrski, T. Huzarski, T. Byrski, C. Cybulski, A. Mackiewicz, S. Gozdecka-Grodecka, J. Gronwald, E. Kowalska, O. Haus, E. Grzybowska, M. Stawicka, M. Swiec, U. Urbaǹski, S. Niepsuj, B. Was̀ko, S. Gòz̀dz̀, P. Wandzel, C. Szczylik, D. Surdyka, A. Rozmiarek, O. Zambrano, M. Posmyk, S. Narod, and L. J (2005). A common variant of CDKN2A (p16) predisposes to breast cancer. *J Med Genet 42*(10), 763 – 765.

Deng, H., W. Chen, and R. R. Recker (2002). Transmission disequilibrium test with discordant sib pairs when parents are available. *Human genetics 110*(5), 451 – 446.

Deng, H. W., W. M. Chen, and R. R. Recker (2001). Population admixture: Detection by Hardy - Weinberg test and its quantitative effects on linkage - disequilibrium methods for localizing genes underlying complex traits. *Genetics 157*, 885 – 897.

Elston, R. C. (1973). Ascertainment and age of onset in pedigree analysis. *Hum Hered 23*, 105–112.

Ewens, W. J. and R. S. Spielman (1995). The transmission/disequilibrium test: history, subdivision, and admixture. *Am J Hum Genet 57*(2), 455 – 464.

F-de Misa, R., J. G. Hernàndez-Jimenez, G. C. Hernàndez, L. Pèrez-Mèndez, A. Aguirre-Jaime, Flores, J. S. Hernàndez, P. A. Molinero, and F. Claverie-Martin (2008). The D84E variant of the α-MSH receptor 1 gene is associated with cutaneous malignant melanoma early onset. *Journal of Dermatological Science 52*(3), 186 – 192.

Fan, R. and M. Knapp (2005). Sibship T^2 association tests of complex diseases for tightly linked markers. *Hum Genomics 2*(2), 90 – 112.

Fengzhu, S., W. D. Flanders, Q. Yang, and M. J. Khoury (1999). Transmission disequilibrium test (TDT) when only one parent is available: The 1-TDT. *Am J Epidemiol 150*(1), 97 – 104.

Gail, M., D. Pee, J. Benechou, and R. Carroll (1999). Designing studies to estimate the penetrance of an identified autosomal mutation: Cohort, casecontrol, and genotyped-proband design. *Genet Epidemiol 16*, 15 – 39.

Gail, M., D. Pee, and R. Carroll (1999). Kin-cohort designs for gene characterization. *Journal of the National Cancer Institute, Monograph 26*, 55 – 60.

Heimbuch, R. C., S. Matthysse, and K. K. Kidd (1980). Estimating age-of-onset distributions for disorders with variable onset. *Am J Hum Genet 32*, 564 – 574.

Hirschhorn, J. N., K. Lohmueller, E. Byrne, and H. K (2002). A comprehensive review of genetic association studies. *Genet Med 4*(4), 45 – 61.

Horvath, S. and N. M. Laird (1998). A discordant - sibship test for disequilibrium and linkage: no need for parental data. *Am J Hum Genet 63*, 1886 – 1897.

Hsu, L., R. L. Prentice, L. P. Zhao, and J. J. Fan (1999). On dependence estimation using correlated failure time data from case-control family studies. *Biometrika 86*, 743 – 753.

Langbehn, D. R., R. R. Brinkman, D. Falush, J. S. Paulsen, and M. R. Hayden (2004). A new model for prediction of the age of onset and penetrance for Huntingtons disease based on CAG length. *Clin Genet 65*(4), 267 – 277.

Lee, W. C. (2003). Searching for disease - susceptibility loci by testing for Hardy - Weinberg disequilibrium in a gene bank of affected individuals. *Am J Epidemiol 158*(5), 397 – 400.

Li, B. and S. M. Leal (2008). Deviations from Hardy-Weinberg equilibrium in parental and unaffected sibling genotype data. *Hum Hered 67*(2), 104 – 15.

Li, H., P. Yang, and A. G. Schwartz (1998). Analysis of age of onset data from casecontrol family studies. *Biometrics 54*, 1030 – 1339.

Lunt, P. W., D. A. S. Compston, and P. S. Harper (1989). Estimation of age dependent penetrance in facioscapulohumeral muscular dystrophy by minimising ascertainment bias. *J Med Genet 26*(12), 755 – 760.

Meyer, J. M. and L. J. Eaves (1988). Estimating genetic parameters of survival distributions: a multifactorial model. *Genet Epidemiol 5*(4), 265 – 275.

Moore, D. F., N. Chatterjee, D. Pee, and M. H. Gail (2001). Pseudo-likelihood estimates of the cumulative risk of an autosomal dominant disease from a kin-cohort study. *Genet Epidemiol 20*, 210 – 227.

Morgan, T. M., H. M. Krumholz, R. P. Lifton, and J. A. Spertus (2007). Nonvalidation of reported genetic risk factors for acute coronary syndrome in a large-scale replication study. *JAMA 297*(14), 1551 – 1561.

Nielsen, D. M., M. G. Ehm, and B. S. Weir (1998). Detecting marker - disease association by testing for Hardy - Weinberg disequilibrium at a marker locus. *Am J Hum Genet 63*, 1531 – 1540.

Patel, M. S., D. E. C. Cole, J. D. Smith, G. A. Hawker, B. Wong, H. Trang, R. Vieth, P. Meltzer, and L. A. Rubin (2000). Alleles of the estrogen receptor alpha gene and an estrogen receptor cotranscriptional activator gene, amplified in breast cancer-1 (AIB1), are associated with quantitative calcaneal ultrasound. *J Bone Miner Res 15*(11), 2231 – 2239.

Risch, N. (1983). Estimating morbidity risks with variable age of onset: review of methods and a maximum likelihood approach. *Biometrics*.

Risch, N. and K. Merikangas (1996). The future of genetic studies of complex human diseases. *Science 273*, 1516 – 1517.

Rubin, L. A., G. A. Hawker, V. D. Peltekova, L. J. Fielding, R. Ridout, and D. E. C. Cole (1999). Determinants of peak bone mass: clinical and genetic analyses in a young female Canadian cohort. *J Bone Miner Res 14*(4), 633 – 643.

Schaid, D. J. and S. J. Jacobsen (1999). Biased tests of association: comparison of allele frequencies when departing from the Hardy - Weinberg proportions. *Am J Epidemiol 149*, 706 – 711.

Self, S. G. and K. Y. Liang (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Stat Assoc 82*(398), 605 – 610.

Sham, P. C. and D. Curtis (1995). An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann Hum Genet 59*(Pt 3), 323 – 336.

Shih, J. H. and N. Chatterjee (2002). Analysis of survival data from case-control family studies. *Biometrics 58*, 502 – 509.

Spielman, R. S. and W. J. Ewens (1998). A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet 62*(2), 450 – 458.

Spielman, R. S., R. E. McGinnis, and W. J. Ewens (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet 52*(3), 506 – 516.

Strahan, N. V., E. A. Murphy, N. J. Fortuin, P. C. Come, and J. O. Humphries (1983). Inheritance of the mitral valve prolapse syndrome. *Am J Med 74*, 967 – 972.

Sturt, E. (1986). Application of survival analysis to the inception of dementia. *Psychol Med 16*, 583 – 593.

Wacholder, S., P. Hartge, J. P. Struewing, D. Pee, M. McAdams, L. Brody, and M. Tucker (1998). The kin-cohort study for estimating penetrance. *Am J Epidemiol 148*(7), 623 – 630.

Weir, B. S. (1996). *Genetic Data Analysis 2: Methods for Discrete Population Genetic Data*. Sinauer Associates.

Whittemore, A. S. (1995). Logistic regression of family data from case-control studies. *Biometrika 82*, 57 – 67.

Wickramaratne, P. J., B. A. Prusoff, K. R. Merikangas, and M. M. Weissman (1986). The use of survival models with non-proportional hazard functions to investigate age of onset in family studies. *J Chronic Dis 39*, 389 – 397.

Wittke-Thompson, J. K., A. Pluzhnikov, and N. J. Cox (2005). Rational inferences about departures from Hardy - Weinberg equilibrium. *The Am J Hum Genet 76*, 967 – 986.

Xu, J., A. Turner, J. Little, E. R. Bleecker, and D. A. Meyers (2002). Positive results in association studies are associated with departure from Hardy - Weinberg equilibrium: hint for genotyping error? *Hum Genet 111*(2), 573 – 574.

Yan, T., Y. Yang, X. Cheng, M. DeAngelis, J. Hoh, and H. Zhang (2009). Genotypic association analysis using discordant-relative-pairs. *Annals of Human Genetics 73*, 84 – 94.

Yun, F. H., B. Y. Wong, M. Chase, A. Y. Shuen, L. Canaff, K. Thongthai, K. Siminovitch, G. N. Hendy, and D. E. C. Cole (2007). Genetic variation at the calcium - sensing receptor (CASR) locus: Implications for clinical molecular diagnostics. *Clin Biochem 40*(8), 551 – 561.

Zhao, L. P., L. Hsu, S. Holte, Y. Chen, F. Quiaoit, and R. L. Prentice (1998). Combined association and aggregation analysis of data from case-control family studies. *Biometrika 85*, 299 – 315.

Appendix A

Hardy-Weinberg Disequilibrium Due To Association In Affected Individuals And Their Parents

Consider an affected person and its parent. The genotypic frequencies for the affected child and its parent are derived below as are the Hardy-Weinberg coefficients.

A.1 Genotypic Frequencies

The joint probability of an affected child and its parent's genotypes is obtained by summing over all probabilities for the other parent's genotype, g_3,

$$\begin{aligned} P(g_1 \cap g_2) &= \sum_{g_3} P(g_1 \cap g_2 | g_3) P(g_3) \\ &= \sum_{g_3} P(g_1 | g_2 \cap g_3) P(g_2) P(g_3). \qquad \text{(A.1)} \end{aligned}$$

Table A.1 summarizes the possible genotypes of the parents, their probability assuming Hardy-Weinberg equilibrium in the population and the conditional probability of a child's genotype given their parents' genotypes.

Table A.1: Possible parental genotypes types, their probabilities with probabilities of possible offspring genotypes

Parent 1 genotype g_2	Parent 2 genotype g_3	Probability of parent's genotype $P(g_2)P(g_3)$	Conditional Probability of offspring genotype $P(g_1 \mid g_2 \cap g_3)$		
			AA	Aa	aa
AA	AA	p^4	1	0	0
AA	Aa	$2\ p^3\ q$	0.5	0.5	0
AA	aa	$p^2\ q^2$	0	1	0
Aa	AA	$2\ p^3\ q$	0.5	0.5	0
Aa	Aa	$4\ p^2\ q^2$	0.25	0.5	0.25
Aa	aa	$2\ p\ q^3$	0	0.5	0.5
aa	AA	$p^2\ q^2$	0	1	0
aa	Aa	$2\ p\ q^3$	0	0.5	0.5
aa	aa	q^4	0	0	1

For example, the joint probability of the affected child having genotype AA and its parent having genotype AA is

$$P(AA \cap AA) = P(AA|AA \cap AA)P(AA)P(AA) + P(AA|AA \cap Aa)P(AA)P(Aa).$$

Substituting values from Table A.1 gives the joint probability

$$\begin{aligned} P(AA \cap AA) &= 1p^2p^2 + \frac{1}{2}p^2 2pq \\ &= p^3(p+q) \\ &= p^3. \end{aligned}$$

If the genotype of the parent is AA or aa, the genotype of the offspring cannot be aa or AA, respectively. Therefore,

$$P(aa \cap AA) = P(AA \cap aa) = 0.$$

Similarly, other probabilities can be computed (Table A.2).

Table A.2: Joint probability of genotypes of an affected child and its parent

	Parent			
Affected	AA	Aa	aa	Total
AA	p^3	p^2q	0	p^2
Aa	p^2q	pq	pq^2	$2pq$
aa	0	pq^2	q^3	q^2
Total	p^2	$2pq$	q^2	1

The conditional probability of the pair of genotypes given that the child is affected and the parent has status d_2, using equation (A.1) is

$$P(g_1 \cap g_2 | \mathcal{A}_1 \cap d_2) = \frac{P(\mathcal{A}_1|g_1)P(d_2|g_2)}{P(d_2|\mathcal{A}_1)K_P} \sum_{g_3} P(g_1|g_2 \cap g_3)P(g_2)P(g_3). \tag{A.2}$$

Affected Child-Parent Pair

When the parent is also affected, i.e., $d_2 = \mathcal{A}_2$, then $P(d_2|\mathcal{A}_1) = P(\mathcal{A}_2|\mathcal{A}_1)$ in (A.2) is the recurrence risk of the parent being affected given that their child is affected, denoted by K_R, and equation (A.2) becomes

$$P(g_1 \cap g_2 | \mathcal{A}_1 \cap \mathcal{A}_2) = \frac{P(\mathcal{A}_1|g_1)P(\mathcal{A}_2|g_2)}{K_P K_R} \sum_{g_3} P(g_1|g_2 \cap g_3)P(g_2)P(g_3). \tag{A.3}$$

For example, the joint probability of both the affected child and their affected parent having genotype AA is

$$P(AA \cap AA | \mathcal{A}_1 \cap \mathcal{A}_2) = \frac{P(\mathcal{A}_1|AA)P(\mathcal{A}_2|AA)}{K_P K_R} \sum_{g_3} P(g_1 = AA|AA \cap g_3)P(AA)P(g_3).$$

From Table A.2,

$$\begin{aligned} P(AA \cap AA|\mathcal{A}_1 \cap \mathcal{A}_2) &= \frac{\alpha\alpha}{K_P K_R} p^3 \\ &= \frac{\alpha^2 p^3}{K_P K_R}. \end{aligned}$$

Similarly, other probabilities can be obtained (Table A.3).

Table A.3: Joint probability of genotypes of an affected child-parent pair

	Parent			
Affected	AA	Aa	aa	Total
AA	$\frac{\phi_0^2 p^3}{K_P K_R}$	$\frac{\phi_0\phi_1 p^2 q}{K_P K_R}$	0	$\frac{\phi_0 p^2}{K_P K_R} SP_{AA}$
Aa	$\frac{\phi_0\phi_1 p^2 q}{K_P K_R}$	$\frac{\phi_1^2 pq}{K_P K_R}$	$\frac{\phi_1\phi_2 pq^2}{K_P K_R}$	$\frac{\phi_1 pq}{K_P K_R} SP_{Aa}$
aa	0	$\frac{\phi_1\phi_2 pq^2}{K_P K_R}$	$\frac{\phi_2^2 q^3}{K_P K_R}$	$\frac{\phi_2 q^2}{K_P K_R} SP_{aa}$
Total	$\frac{\phi_0 p^2}{K_P K_R} SP_{AA}$	$\frac{\phi_1 pq}{K_P K_R} SP_{Aa}$	$\frac{\phi_2 q^2}{K_P K_R} SP_{aa}$	1

For simplicity of presentation, the expressions have been abbreviated using

$$SP_{AA} = \phi_0 p + \phi_1 q,$$

$$SP_{Aa} = \phi_0 p + \phi_1 + \phi_2 q,$$

and

$$SP_{aa} = \phi_1 p + \phi_2 q.$$

Note that this table is symmetric and that the marginal probabilities are the same. Using the fact that the probabilities add up to one, the recurrence risk for a parent,

K_R can be obtained from the above table as

$$K_R = \frac{\alpha^2}{K_P}(p^3 + 2p^2\beta q + \beta^2 pq + 2\beta pq^2\gamma + \gamma^2 q^3).$$

Discordant Child-Parent Pair

When the parent is not affected, i.e., $d_2 = \bar{\mathcal{A}}_2$, then $P(d_2|\mathcal{A}_1) = P(\bar{\mathcal{A}}_2|\mathcal{A}_1) = 1 - K_R$, and $P(d_2|g_2) = P(\bar{\mathcal{A}}_2|g_2)$ and equation (A.2) becomes

$$P(g_1 \cap g_2|\mathcal{A}_1 \cap \bar{\mathcal{A}}_2) = \frac{P(\mathcal{A}_1|g_1)P(\bar{\mathcal{A}}_2|g_2)}{K_P(1-K_R)} \sum_{g_3} P(g_1|g_2 \cap g_3)P(g_2)P(g_3). \tag{A.4}$$

For example, the joint probability of both the affected child and their affected parent having genotype AA is

$$P(AA \cap AA|\mathcal{A}_1 \cap \bar{\mathcal{A}}_2) = \frac{P(\mathcal{A}_1|AA)P(\bar{\mathcal{A}}_2|AA)}{K_P(1-K_R)} \sum_{g_3} P(g_1|g_2 \cap g_3)P(g_2)P(g_3).$$

Substituting from Table A.2,

$$P(AA \cap AA|\mathcal{A}_1 \cap \bar{\mathcal{A}}_2) = \frac{\alpha(1-\alpha)}{K_P(1-K_R)}p^3.$$

Similarly the other joint probabilities can also be calculated (Table A.4).

Table A.4: Joint probability of genotypes of a discordant child-parent pair

Parent				
Affected	AA	Aa	aa	Total
AA	$\frac{\phi_0(1-\phi_0)p^3}{K_P(1-K_R)}$	$\frac{\phi_0(1-\phi_1)p^2q}{K_P(1-K_R)}$	0	$\frac{\phi_0p^2}{K_P(1-K_R)}SNP_{AA}$
Aa	$\frac{\phi_1(1-\phi_0)p^2q}{K_P(1-K_R)}$	$\frac{\phi_1(1-\phi_1)pq}{K_P(1-K_R)}$	$\frac{\phi_1(1-\phi_2)pq^2}{K_P(1-K_R)}$	$\frac{\phi_1pq}{K_P(1-K_R)}SNP_{Aa}$
aa	0	$\frac{\phi_2(1-\phi_1)pq^2}{K_P(1-K_R)}$	$\frac{\phi_2(1-\phi_2)q^3}{K_P(1-K_R)}$	$\frac{\phi_2q^2}{K_P(1-K_R)}SNP_{aa}$
Total	$\frac{(1-\phi_0)p^2}{K_P(1-K_R)}SP_{AA}$	$\frac{(1-\phi_1)pq}{K_P(1-K_R)}SP_{Aa}$	$\frac{(1-\phi_2)q^2}{K_P(1-K_R)}SP_{aa}$	1

For simplicity of presentation, the expressions have been abbreviated using SP_{AA}, SP_{Aa} and S_{aa} described above and

$$SNP_{AA} = (1-\phi_0)p + (1-\phi_1)q$$

$$SNP_{Aa} = (1-\phi_0)p + (1-\phi_1) + (1-\phi_2)q$$

and

$$SNP_{aa} = (1-\phi_1)p + (1-\phi_2)q$$

A.2 Hardy-Weinberg Coefficient

The Hardy-Weinberg coefficient, D, measures the excess homozygosity and is given by

$$D = P_{aa} - q^2,$$

where the minor allele frequency can be obtained from the genotypic frequencies using the relationship

$$q = P_{aa} + \frac{1}{2}P_{Aa}.$$

Affected Child-Parent Pair

When the parent is also affected, the allele frequencies of the affected child and parent (Table A.3) are the same

$$q_{1\mathcal{A}} = q_{2\mathcal{A}} = \frac{\alpha^2 q}{2K_P K_R}[\beta^2 p + (3pq\gamma + p^2)\beta + 2\gamma^2 q^2]$$

and the HWD coefficient is

$$D_{i\mathcal{A}} = -\frac{\alpha^4 p^2 q^2}{4K_P^2 K_R^2}\{q[q\beta(\beta-4) - 4p]\gamma^2 + 2\beta[q\beta(\beta-p) - 2p^2]\gamma + \beta^4 + \beta^2 p(2\beta + p)\},$$

for $i = 1, 2$.

Discordant Child-Parent Pair

When the parent is unaffected, the allele frequency of the affected child can be obtained from Table A.4

$$q_{1\mathcal{A}} = \frac{-\alpha q}{2K_P(1-K_R)}[\alpha\beta p(p + q\gamma + \beta) - 2(1 - \alpha\gamma q)(p\beta + q\gamma)]$$

and for the unaffected parent the allele frequency is

$$q_{2\mathcal{A}} = \frac{\alpha q}{2K_P(1-K_R)}[p(1-\alpha\beta)(p+\beta) + q(1-2\alpha\gamma)(p\beta + q\gamma) + pq\beta(1-\alpha\gamma) + q\gamma].$$

The HWD coefficients for the discordant child-parent pair in this case are

$$D_{1\mathcal{A}} = \frac{-\alpha^4 p^2 q^2}{4K_P^2(1-K_R)^2}\{\beta^2(\beta+p)^2 - 2\beta\gamma(2p^2 + pq\beta - \beta^2 q) - q\gamma^2[4p + \beta q(4-\beta)]$$
$$- \frac{4}{\alpha}(\beta^2 - \gamma)[p + \beta + \gamma q - \frac{1}{\alpha}]\}$$

and

$$\begin{aligned}D_{2\mathcal{A}} = \frac{-\alpha^4p^2q^2}{4K_P^2(1-K_R)^2}\Big(&\beta^2(\beta+p)^2 - 2\beta\gamma(2p^2+pq\beta-\beta^2q)\\ &- q\gamma^2[4p+\beta q(4-\beta)]\\ &+ \frac{2}{\alpha}\{2\gamma[pq+(1-3pq)\beta-q^2\beta^2]\\ &- \beta(\beta^2-p^2+2p^2\beta)+q\gamma^2(\beta q+2p)\}\\ &+ \frac{1}{\alpha^2}\{[\beta^2(1-2q)-2\beta(p+\gamma q)+1]\\ &- 2pq\gamma+q^2(\gamma^2+1)\}\Big).\end{aligned}$$

The HWD coefficient for the affected child is positive if $\gamma > \beta^2$ and is zero for the multiplicative model.

In order to understand the magnitude and direction of D, it was studied under some specific genetic models discussed in the next section.

A.3 Specific Genetic Models

The HWD coefficients for an affected child and its parent were studied in specific genetic models for both cases regarding the disease status of the parent. For each model, the HWD coefficient is plotted for two different values of the homozygote relative risk, γ, 1.5, 3.

Dominant Model, $\beta = \gamma, \gamma > 1$

Affected Child-Parent Pair

The HWD coefficient for the affected child and its parent is the same when the parent is also affected and is

$$D_{1\mathcal{A}} = D_{2\mathcal{A}} = \frac{-\alpha^4\gamma^2(\gamma-1)^2p^2q^2}{4K_P^2K_R^2}[q(q+2)+\frac{\gamma+3}{(\gamma-1)}]$$

The HWD coefficient is always negative.

Discordant Child-Parent Pair

When the disease status of the parent is unaffected, the HWD coefficients for the child and its parent are different and are

$$D_{1\bar{\mathcal{A}}} = \frac{-\alpha^4\gamma(\gamma-1)p^2q^2}{4K_P^2(1-K_R)^2}\{(1+q)^2\gamma^2+p(q+3)\gamma-\frac{4}{\alpha}[(1+q)\gamma+p]+\frac{4}{\alpha^2}\},$$

and

$$D_{2\bar{\mathcal{A}}} = \frac{\alpha^3(1-\alpha\gamma)(\gamma-1)p^2q^2}{4K_P^2(1-K_R)^2}[(1+q)^2\gamma^2+p(q+3)\gamma-\frac{p^2}{\alpha}(\gamma-1)].$$

Figure A.1 illustrates the direction and magnitude of HWD in the dominant genetic model. Also shown are the coefficients for unrelated affected and unaffected individuals. The HWD coefficient increases in magnitude with γ for all cases and reaches a maximum when q is between 0.3 and 0.5. Panel (a) shows that the HWD coefficient for the affected child is negative and is similar regardless of the disease status of the parent. Panel (b) shows that the coefficient is largest when the parent is affected and when the parent is unaffected, its coefficient is similar to that of an unrelated unaffected person. The HWD coefficient for the parent (panel (b)) is smaller in magnitude than the affected child (panel (a)) when the parent is unaffected ($\triangle$).

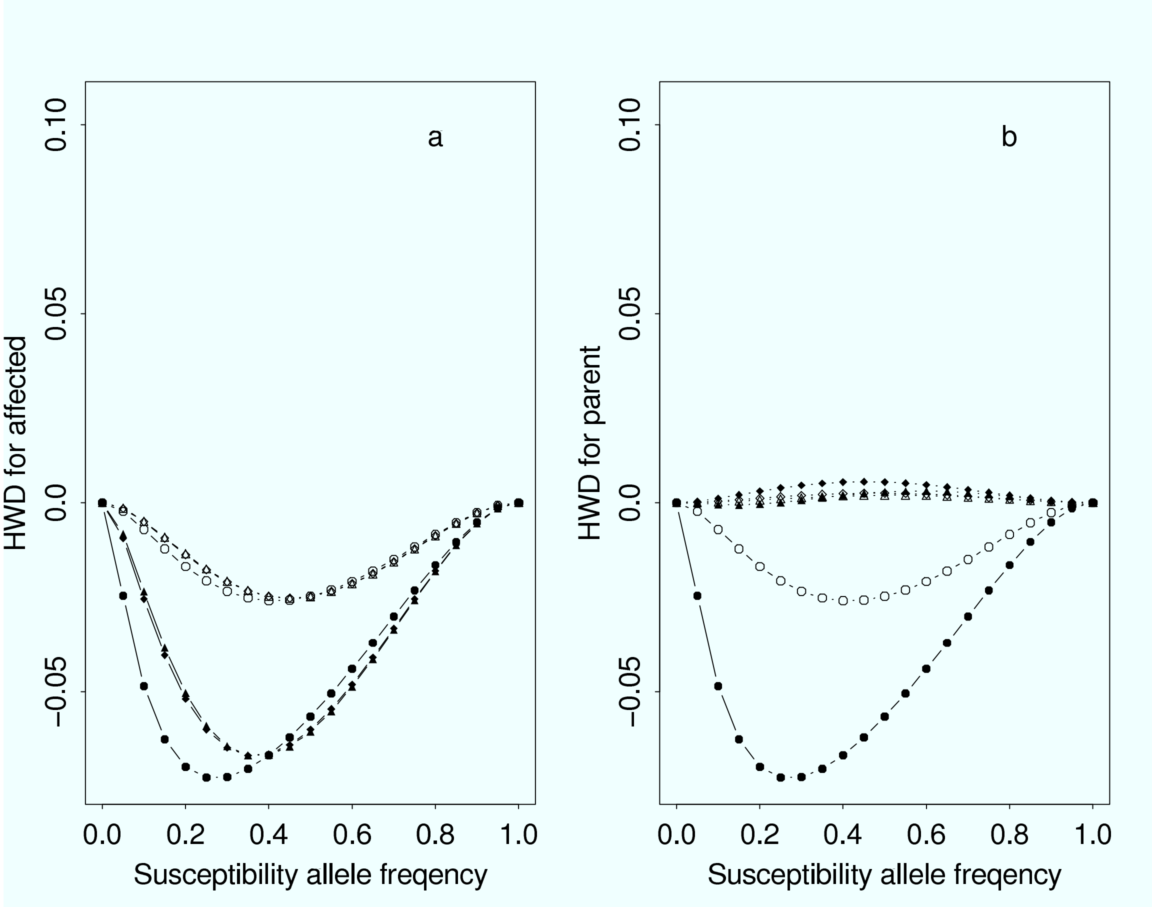


Figure A.1: HWD coefficients for the dominant genetic model as a function of the susceptibility-allele frequency for an affected child (a) and their parent or an unrelated unaffected person (b).
$K_P = 0.1$, disease status of parent (affected, unaffected) = (○, △), unrelated affected/unaffected ◇, open/filled symbol for $\gamma = 1.5/3$.

Recessive Model, $\beta = 1$, $\gamma > 1$

Affected Child-Parent Pair

When the parent is also affected, the HWD coefficient for both the affected child and its parent is the same and is

$$D_{1\mathcal{A}} = D_{2\mathcal{A}} = \frac{\alpha^4(\gamma-1)p^2q^2}{4K_P^2K_R^2}[\gamma q(4-q)+(2-q)^2],$$

which is always positive.

Discordant Child-Parent Pair

When the parent is unaffected, the HWD coefficient is different for the affected child and its parent,

$$D_{1\bar{A}} = \frac{\alpha^4(\gamma-1)p^2q^2}{4K_P^2(1-K_R)^2}\{\gamma q(4-q)+(2-q)^2-\frac{4}{\alpha}[(\gamma-1)q+2]+\frac{4}{\alpha^2}\}$$

and

$$D_{2\bar{A}} = \frac{-\alpha^2(1-\alpha)^2(\gamma-1)^2p^2q^2}{4K_P^2(1-K_R)^2}2[q^2+\frac{4\alpha}{(\gamma-1)(1-\alpha)}(p+\gamma q)].$$

The coefficient is always negative for the parent.

Figure A.2 illustrates the direction and magnitude of HWD in the recessive genetic model. Also shown are the coefficients for unrelated affected and unaffected individuals. The HWD coefficient increases in magnitude with γ for all cases and reaches a maximum when q is between 0.3 and 0.5. The HWD coefficient for the affected child is positive and is similar regardless of the disease status of the parent (panel (a)). When the parent is unaffected (panel (b), $\triangle$), its HWD coefficient is smaller in magnitude than that of the affected child (panel a) and is negative. When the parent is unaffected (panel (b), $\triangle$), the coefficient is negative and is similar to that of unrelated unaffected individuals ($\diamondsuit$).

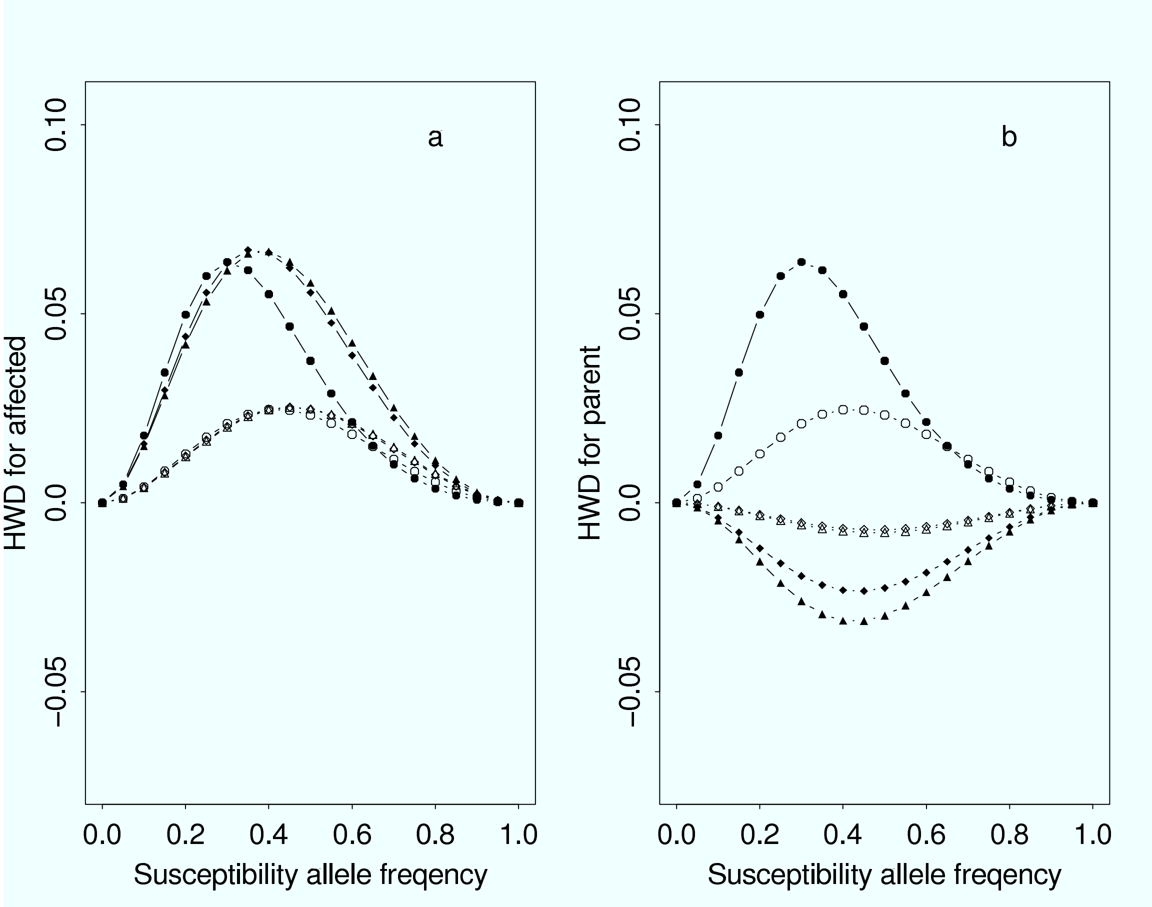


Figure A.2: HWD coefficients for the recessive genetic model for an affected child (a) and their parent or an unrelated unaffected person (b).
$K_P = 0.2$, disease status of parent (affected, unaffected) = (○, △), unrelated affected/unaffected ◇, open/filled symbol for $\gamma = 1.5/3$.

Additive Model, $\gamma = 2\beta - 1, \beta > 1$

Affected Child-Parent Pair

The HWD coefficient of the affected child and its parent is the same when the parent is also affected and is

$$D_{1\mathcal{A}} = D_{2\mathcal{A}} = \frac{-\alpha^4(\gamma-1)^2p^2q^2}{64K_P^2K_R^2}[4(\gamma-1)^2q^2 + 4(\gamma+3)(\gamma-1)q + (\gamma+9)(\gamma+1)],$$

which is always negative.

Discordant Child-Parent Pair

When the parent is unaffected, the HWD coefficient for the affected child is

$$D_{1\bar{A}} = \frac{-\alpha^4 p^2 q^2 (\gamma - 1)^2}{64 K_P^2 (1 - K_R)^2} \{(2q + 1)^2 \gamma^2 + 2\gamma(4pq + 5) + (3 - 2q)^2 \\ - \frac{8}{\alpha}[2(\gamma - 1)q + \gamma + 3] + \frac{16}{\alpha^2}\}$$

and for the parent is

$$D_{2\bar{A}} = \frac{-\alpha^4 p^2 q^2 (\gamma - 1)^2}{64 K_P^2 (1 - K_R)^2} [(2q + 1)^2 \gamma^2 + 2\gamma(4pq + 5) + (3 - 2q)^2 - \frac{4}{\alpha}(\gamma + 1) + \frac{4}{\alpha^2}].$$

Figure A.3 illustrates the direction and magnitude of the HWD coefficient in the additive genetic model. Also shown are the coefficients for unrelated affected and unaffected individuals. The HWD coefficient increases in magnitude with γ for all cases and reaches a maximum when q is between 0.3 and 0.5. Panel (a) shows that the HWD coefficient for the affected child is negative and is similar regardless of the disease status of the parent. Comparing panels (a) and (b) shows that when the parent is unaffected (panel (b), $\triangle$), its HWD coefficient is smaller in magnitude than that of affected child (panel (a)).

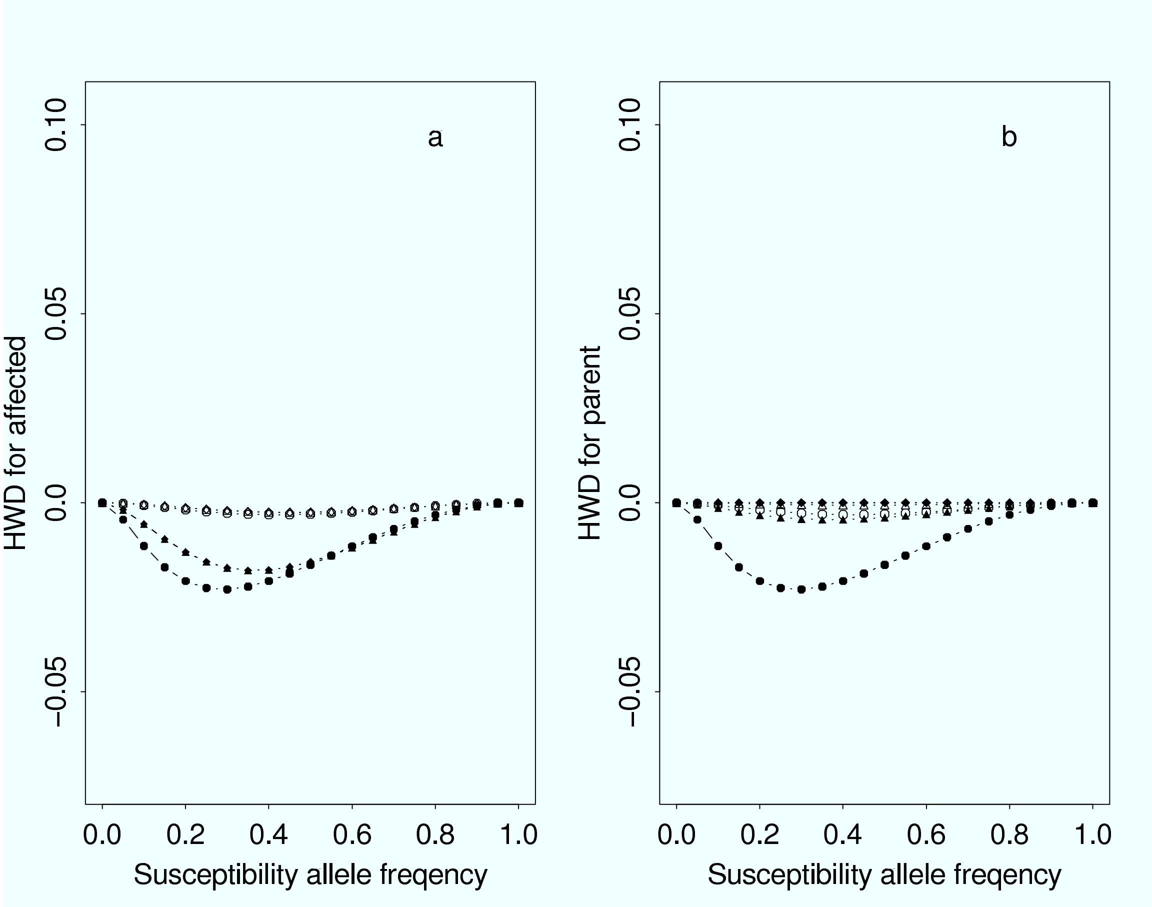


Figure A.3: HWD coefficients for the additive genetic model for an affected child (a) and their parent or an unrelated unaffected person (b).
$K_P = 0.01$, disease status of parent (affected, unaffected) = (○, △), unrelated affected/unaffected ◊, open/filled symbol for $\gamma = 1.5/3$.

Multiplicative Model, $\gamma = \beta^2, \beta > 1$

Affected Child-Parent Pair

When the parent is also affected, the HWD coefficient for both the affected child and its parent is the same and is

$$D_{1\mathcal{A}} = D_{2\mathcal{A}} = \frac{-\alpha^4\gamma p^2 q^2}{4K_P^2 K_R^2}[q^2\gamma^2 + 2(q-p)(p-\gamma q)\gamma^{1/2} + (1-6pq)\gamma + p^2].$$

Discordant Child-Parent Pair

When the parent is unaffected, the HWD coefficient for the affected child is

$$D_{1\bar{\mathcal{A}}} = \frac{-\alpha^4\gamma p^2 q^2}{4K_P^2(1-K_R)^2}[q^2\gamma^2 + 2(q-p)(p-\gamma q)\gamma^{1/2} + (1-6pq)\gamma + p^2]$$

and for the parent is

$$D_{2\bar{\mathcal{A}}} = \frac{-\alpha^2 p^2 q^2}{4K_P^2(1-K_R)^2}(1+\alpha\sqrt{\gamma})^2[q^2\gamma^2 + 2(q-p)(p-\gamma q)\gamma^{1/2} + (1-6pq)\gamma + p^2]$$

Note that for the multiplicative model,

$$D_{1\bar{\mathcal{A}}} = D_{1\mathcal{A}}\frac{K_R^2}{(1-K_R)^2},$$

so there is a simple relationship between the HWD coefficients for the affected child when the parent is affected and unaffected. When the recurrence risk is less than 0.5, the coefficient is larger when the parent is affected.

Similarly, there is a simple relationship between the HWD coefficients for the parents

$$D_{2\bar{\mathcal{A}}} = D_{2\mathcal{A}}\frac{K_R^2}{(1-K_R)^2}\Big(1+\frac{1}{\alpha\sqrt{\gamma}}\Big)^2,$$

and between the affected child and its parent when the parent is unaffected

$$D_{2\bar{\mathcal{A}}} = D_{1\bar{\mathcal{A}}}\Big(1+\frac{1}{\alpha\sqrt{\gamma}}\Big)^2.$$

So that when $K_R < 0.5$, the HWD coefficient for the affected child and its parent is larger when the parent is unaffected than when the parent is affected. If $K_R > 0.5$, the HWD coefficient for the affected child and its parent is smaller when the parent is unaffected than when the parent is affected. When the parent is unaffected, the HWD coefficient for the affected child is smaller in magnitude than that of the parent.

Figure A.4 illustrates the direction and magnitude of HWD in the multiplicative genetic model for an affected child and its parent. Also shown are the coefficients for unrelated affected and unaffected individuals.

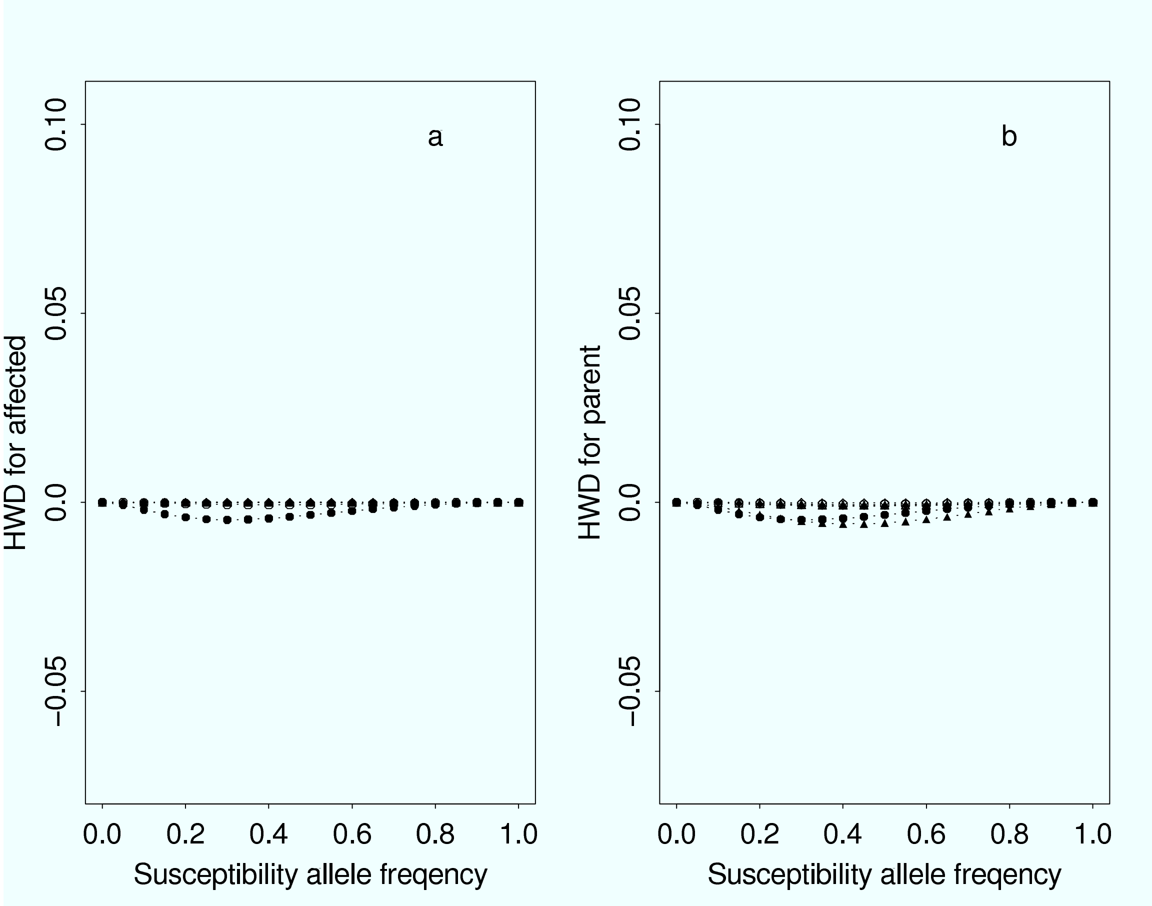


Figure A.4: HWD coefficients for the multiplicative genetic model for an affected child (a) and their parent or unrelated unaffected person (b).
K_P = 0.05, disease status of parent (affected, unaffected) = (○, △), unrelated affected/unaffected ◇, open/filled symbol for $\gamma = 1.5/3$.

The HWD coefficient increases in magnitude with γ for all cases and reaches a maximum when q is between 0.3 and 0.5. When the parent is unaffected (panel (b), $\triangle$), its HWD coefficient is negative and is larger in magnitude than that of the affected child (panel (a), $\triangle$). Panel (b) shows that the HWD coefficient for the parent is negative when the parent is affected ($\circ$) or unaffected ($\triangle$).

Appendix B

Hardy-Weinberg Disequilibrium Due To Association In Affected Individuals And Their Grandparent

Consider an affected person and its grandparent. The genotypic frequencies for the affected individual and its grandparent are derived below as are the Hardy-Weinberg coefficients.

B.1 Genotypic Frequencies

The joint probability of genotypes for an affected grandchild and its grandparent is obtained using the law of total probability,

$$P(g_1 \cap g_2) = \sum_{g_3} P(g_1 \cap g_2 | g_3) P(g_3),$$

where g_3 is the common link (child of the grandparent and parent of the affected grandchild)

Applying Bayes' theorem gives

$$P(g_1 \cap g_2) = \sum_{g_3} P(g_3) P(g_2 | g_3) P(g_1 | g_2 \cap g_3).$$

The genotype of a child depends only on its parent's genotype, therefore

$$\begin{aligned} P(g_1 \cap g_2) &= \sum_{g_3} P(g_3) P(g_2|g_3) P(g_1|g_3) \\ &= \sum_{g_3} P(g_3) \frac{P(g_2 \cap g_3)}{P(g_3)} \frac{P(g_1 \cap g_3)}{P(g_3)} \end{aligned}$$

or

$$P(g_1 \cap g_2) = \sum_{g_3} \frac{1}{P(g_3)} P(g_2 \cap g_3) P(g_1 \cap g_3). \tag{B.1}$$

The joint probabilities $P(g_2 \cap g_3)$ and $P(g_1 \cap g_3)$ are the joint genotypic probability of an affected grandchild and its parent that can be obtained from Table A.2.

For example, the joint probability of the affected grandchild having genotype AA and its grandparent having genotype AA is

$$\begin{aligned} P(AA \cap AA) &= P(AA \cap AA) P(AA \cap AA) \frac{1}{P(AA)} \\ &\quad + P(AA \cap Aa) P(AA \cap Aa) \frac{1}{P(Aa)}. \end{aligned}$$

Substituting values from Table A.2 gives the joint probability

$$\begin{aligned} P(AA \cap AA) &= p^3 p^3 \frac{1}{p^2} + p^2 q p^2 q \frac{1}{2pq} \\ &= p^4 + \frac{1}{2} p^3 q \\ &= \frac{p^3}{2} (2p + q) \\ &= \frac{p^3}{2} (1 + p). \end{aligned}$$

Similarly, other probabilities can be computed (Table B.1)

Table B.1: Joint probability of genotypes of an affected grandchild and its grandparent

Grandparent				
Affected	AA	Aa	aa	Total
AA	$\frac{p^3}{2}(1+p)$	$\frac{p^2q}{2}(1+2p)$	$\frac{p^2q^2}{2}$	p^2
Aa	$\frac{p^2q}{2}(1+2p)$	$\frac{pq}{2}(1+4pq)$	$\frac{pq^2}{2}(1+2q)$	$2pq$
aa	$\frac{p^2q^2}{2}$	$\frac{pq^2}{2}(1+2q)$	$\frac{q^3}{2}(1+q)$	q^2
Total	p^2	$2pq$	q^2	1

The equation (B.1) gives

$$P(g_1 \cap g_2|\mathcal{A}_1 \cap d_2) = \frac{P(\mathcal{A}_1|g_1)P(d_2|g_2)}{P(d_2|\mathcal{A}_1)K_P} \sum_{g_3} \frac{1}{P(g_3)} P(g_2 \cap g_3)P(g_1 \cap g_3). \tag{B.2}$$

Affected Grandchild-Grandparent Pair

When the grandparent is also affected, i.e., $d_2 = \mathcal{A}_2$, then $P(d_2|\mathcal{A}_1) = P(\mathcal{A}_2|\mathcal{A}_1)$ in (B.2) is the recurrence risk of the grandparent being affected given that their grandchild is affected, denoted by K_R, and (B.2) becomes

$$P(g_1 \cap g_2|\mathcal{A}_1 \cap \mathcal{A}_2) = \frac{P(\mathcal{A}_1|g_1)P(\mathcal{A}_2|g_2)}{K_P K_R} P(g_1 \cap g_2). \tag{B.3}$$

For example, the joint probability of the grandchild and grandparent having genotype AA conditional on them both being affected is

$$P(AA \cap AA|\mathcal{A}_1 \cap \mathcal{A}_2) = \frac{P(\mathcal{A}_1|AA)P(\mathcal{A}_2|AA)}{K_P K_R} P(AA \cap AA).$$

From Table B.1,

$$\begin{aligned} P(AA \cap AA|\mathcal{A}_1 \cap \mathcal{A}_2) &= \frac{\phi_0\phi_0 p^3}{2K_P K_R}(2p+q) \\ &= \frac{\phi_0^2 p^3}{2K_P K_R}(2p+q). \end{aligned}$$

Similarly, other probabilities can be obtained (Table B.2).

Table B.2: Joint probability of genotypes of an affected grandchild-grandparent pair

	Grandparent			
Affected	AA	Aa	aa	Total
AA	$\frac{\phi_0^2 p^3}{2K_P K_R}(1+p)$	$\frac{\phi_0\phi_1 p^2 q}{2K_P K_R}(1+2p)$	$\frac{\phi_0\phi_2 p^2 q^2}{2K_P K_R}$	$\frac{\phi_0 p^2}{2K_P K_R}SG_{AA}$
Aa	$\frac{\phi_0\phi_1 p^2 q}{2K_P K_R}(1+2p)$	$\frac{\phi_1^2 pq}{2K_P K_R}(1+4pq)$	$\frac{\phi_1\phi_2 pq^2}{2K_P K_R}(1+2q)$	$\frac{\phi_1 pq}{2K_P K_R}SG_{Aa}$
aa	$\frac{\phi_0\phi_2 p^2 q^2}{2K_P K_R}$	$\frac{\phi_1\phi_2 pq^2}{2K_P K_R}(1+2q)$	$\frac{\phi_2^2 q^3}{2K_P K_R}(1+q)$	$\frac{\phi_2 q^2}{2K_P K_R}SG_{aa}$
Total	$\frac{\phi_0 p^2}{2K_P K_R}SG_{AA}$	$\frac{\phi_1 pq}{2K_P K_R}SG_{Aa}$	$\frac{\phi_2 q^2}{2K_P K_R}SG_{aa}$	1

For simplicity of presentation, the expressions have been abbreviated using

$$SG_{AA} = \phi_0 p(1+p) + \phi_1 q(1+2p) + \phi_2 q^2,$$

$$SG_{Aa} = \phi_0 p(1+2p) + \phi_1(1+4pq) + \phi_2 q(1+2q),$$

and

$$SG_{aa} = \phi_0 p^2 + \phi_1 p + \phi_2 q(1+q).$$

Using the fact that the probabilities add up to one, the recurrence risk, K_R can be obtained from the above table as

$$K_R = \frac{\alpha^2}{2K_P}\{q^3(q+1)\gamma^2+[2pq^2(2q+1)\beta+2p^2q^2]\gamma-2p^2q(2q-3)\beta+pq(1+4pq)\beta^2+p^3(2-q)\}.$$

Discordant Grandchild-Grandparent Pair

When the grandparent is not affected, i.e., $d_2 = \bar{\mathcal{A}}_2$, then $P(d_2|\mathcal{A}_1) = P(\bar{\mathcal{A}}_2|\mathcal{A}_1) = 1 - K_R$, and $P(d_2|g_2) = P(\bar{\mathcal{A}}_2|g_2)$ and (B.4) becomes

$$P(g_1 \cap g_2|\mathcal{A}_1 \cap \bar{\mathcal{A}}_2) = \frac{P(\mathcal{A}_1|g_1)P(\bar{\mathcal{A}}_2|g_2)}{K_P(1-K_R)}P(g_1 \cap g_2). \tag{B.4}$$

For example, the joint probability of the grandchild and grandparent having genotype AA conditional on the grandchild being affected and the grandparent unaffected is

$$P(AA \cap AA|\mathcal{A}_1 \cap \bar{\mathcal{A}}_2) = \frac{P(\mathcal{A}_1|AA)P(\bar{\mathcal{A}}_2|AA)}{K_P(1-K_R)}P(AA \cap AA).$$

Substituting from Table B.1,

$$P(AA \cap AA|\mathcal{A}_1 \cap \bar{\mathcal{A}}_2) = \frac{\phi_0(1-\phi_0)p^3}{2K_P(1-K_R)}(2p+q).$$

Similarly the other joint probabilities can also be calculated (Table B.3).

Table B.3: Joint probability of genotypes of a discordant grandchild-grandparent pair

Grandparent				
Affected	AA	Aa	aa	Total
AA	$\frac{\phi_0(1-\phi_0)p^3(1+p)}{2K_P(1-K_R)}$	$\frac{\phi_0(1-\phi_1)p^2q(1+2p)}{2K_P(1-K_R)}$	$\frac{\phi_0(1-\phi_2)p^2q^2}{2K_P(1-K_R)}$	$\frac{\phi_0 p^2}{2K_P(1-K_R)}SNG_{AA}$
Aa	$\frac{(1-\phi_0)\phi_1 p^2q(1+2p)}{2K_P(1-K_R)}$	$\frac{\phi_1(1-\phi_1)pq(1+4pq)}{2K_P(1-K_R)}$	$\frac{\phi_1(1-\phi_2)pq^2(1+2q)}{2K_P(1-K_R)}$	$\frac{\phi_1 pq}{2K_P(1-K_R)}SNG_{Aa}$
aa	$\frac{(1-\phi_0)\phi_2 p^2q^2}{2K_P(1-K_R)}$	$\frac{(1-\phi_1)\phi_2 pq^2(1+2q)}{2K_P(1-K_R)}$	$\frac{\phi_2(1-\phi_2)q^3(1+q)}{2K_P(1-K_R)}$	$\frac{\phi_2 q^2}{2K_P(1-K_R)}SNG_{aa}$
Total	$\frac{(1-\phi_0)p^2}{2K_P(1-K_R}SG_{AA}$	$\frac{(1-\phi_1)pq}{2K_P(1-K_R}SG_{Aa}$	$\frac{(1-\phi_2)q^2}{2K_P(1-K_R}SG_{aa}$	1

For simplicity of presentation, the expressions have been abbreviated using SG_{AA}, SG_{Aa} and SG_{aa} as described above and

$$SNG_{AA} = (1-\phi_0)p(1+p) + (1-\phi_1)q(1+2p) + (1-\phi_2)q^2,$$

$$SNG_{Aa} = (1-\phi_0)p(1+2p) + (1-\phi_1)(1+4pq) + (1-\phi_2)q(1+2q),$$

and

$$SNG_{aa} = (1-\phi_0)p^2 + (1-\phi_1)p(1+2q) + (1-\phi_2)q(1+q)].$$

B.2 Hardy-Weinberg Coefficient

The Hardy-Weinberg coefficient, D, measures the excess homozygosity and is given by

$$D = P_{aa} - q^2,$$

where the minor allele frequency can be obtained from the genotypic frequencies using the relationship

$$q = P_{aa} + \frac{1}{2}P_{Aa}.$$

Affected Grandchild-Grandparent Pair

When the grandparent is also affected, the allele frequencies of the affected grandchild and its grandparent (Table B.2) are the same

$$q_{i\mathcal{A}} = \frac{\alpha^2 q}{4K_PK_R}\{2q^2(1+q)\gamma^2 + [3(1+2q)\beta + 2p]pq\gamma + p(4pq+1)\beta^2 + (2p+1)p^2\beta\},$$

and the HWD coefficient is

$$D_{i\mathcal{A}} = \frac{-p^2q^2\alpha^4}{16K_P^2K_R^2}\{(1+4pq)^2\beta^4 + [6+2q(1+2q)(1+4pq)\gamma$$
$$-2q(8q^3-28q^2+30q-7)]\beta^3 + [q^2(1+2q)^2\gamma^2$$
$$-2pq(1+2q)(1+2p)\gamma + p^2(1+2p)^2]\beta^2$$
$$+[8q^2(2q^2-q-2)\gamma^2 - 8(2pq+q+1)p^2\gamma]\beta$$
$$-4p^3(2-q)\gamma - 8pq(1+pq)\gamma^2 - 4q^3(1+q)\gamma^3\}.$$

Discordant Grandchild-Grandparent Pair

When the grandparent is unaffected, the allele frequency of the affected grandchild can be obtained from Table B.3

$$q_{1\mathcal{A}} = \frac{-\alpha^2 q}{4K_P(1-K_R)}\{2q^2(1+q)\gamma^2 + pq\gamma[3\beta(1+2q)+2p] - (2q-3)p^2\beta$$
$$+p(1+4pq)\beta^2 - \frac{4}{\alpha}(p\beta+\gamma q)\}.$$

Similarly for the unaffected grandparent

$$q_{2\mathcal{A}} = \frac{-\alpha^2 q}{4K_P(1-K_R)}\{2q^2(1+q)\gamma^2 + pq\gamma[3(1+2q)\beta+2p] - (2q-3)p^2\beta$$
$$+p(1+4pq)\beta^2 - \frac{1}{\alpha}[(3q+1)q\gamma + (6q+1)p\beta + 3p^2]\}.$$

The HWD coefficients for the discordant grandchild-grandparent pair in this case are

$$\begin{aligned}D_{1\mathcal{A}} = \frac{-p^2q^2\alpha^4}{16K_P^2(1-K_R)^2}\{&(1+4pq)\beta^3[(1+4pq)\beta+2q(1+2q)\gamma+2p(3-2q)]\\&+[q(1+2q)\gamma-p(3-2q)]^2\beta^2+8\beta\gamma[q^2(2q^2-q-2)\gamma\\&+(2q^2-3q-1)p^2]-4\gamma[q^3(1+q)\gamma^2+(2-q)p^3]\\&-8pq(pq+1)\gamma^2\\&-\frac{8}{\alpha}(\beta^2-\gamma)[q(1+2q)\gamma+(1+4pq)\beta+p(3-2q)-\frac{2}{\alpha}]\},\end{aligned}$$

and

$$
\begin{aligned}
D_{2\mathcal{A}} = \frac{-p^2q^2\alpha^4}{16K_P^2(1-K_R)^2}\Big(& \quad (1+4pq)\beta^3[(1+4pq)\beta+2q(1+2q)\gamma+2p(3-2q)] \\
+ & \quad [q(1+2q)\gamma-p(3-2q)]^2\beta^2 \\
+ & \quad 8\beta\gamma[q^2(2q^2-q-2)\gamma+(2q^2-3q-1)p^2] \\
+ & \quad 4\gamma[-4q^3(1+q)\gamma^2-8pq(pq+1)\gamma+(q-2)p^3] \\
+\frac{1}{\alpha}\{ & \quad -2(1+4pq)^2\beta^3+4[p(3-2q)(6q^2-3q-1) \\
& \quad +q(1+2q)(6q^2-9q+2)\gamma]\beta^2-2[q^2(12q^2-7)\gamma^2 \\
& \quad +2\gamma(q^2(11-12q(p+1))+q-2)+p^2(-12q(p+1) \\
& \quad +5)]\beta+4\gamma[p(2-3p^2q)+q(2-3pq^2)\gamma \\
& \quad +q^3(1+q)\gamma^2]+4p^3(p+1)\} \\
+\frac{1}{\alpha^2}\{ & \quad (2q-1)^2\beta^2-2(q-p)(\gamma q-p)\beta+q\gamma(-2p+\gamma q)+p^2\}\Big).
\end{aligned}
$$

In order to understand the magnitude and direction of D, it was studied under some specific genetic models discussed in the next section.

B.3 Specific Genetic Models

The HWD coefficients for an affected grandchild and its grandparent were studied in specific genetic models when the grandparent is affected or unaffected. For each model, the HWD coefficient is plotted for two different values of the homozygote relative risk, γ, 1.5, 3.

Dominant Model, $\beta = \gamma, \gamma > 1$

Affected Grandchild-Grandparent Pair

When the grandparent is also affected, the HWD coefficient is the same for the affected grandchild and its grandparent, and is given by

$$\begin{aligned} D_{1\mathcal{A}} = D_{2\mathcal{A}} = \frac{-p^2q^3\alpha^4\gamma}{16K_P^2K_R^2}(\gamma-1)^2\{&4(\gamma-1)q^2(q-5)+3(7\gamma-12)q+2(5\gamma+14) \\ &+\frac{1}{q(\gamma-1)}(\gamma^2+7\gamma+8)\}. \end{aligned}$$

Discordant Grandchild-Grandparent Pair

When the grandparent is unaffected, the HWD coefficient for the affected grandchild is

$$\begin{aligned} D_{1\bar{\mathcal{A}}} = \frac{-p^2q^3\alpha^4\gamma}{16K_P^2(1-K_R)^2}(\gamma-1)^2\{&4(\gamma-1)q^2(q-5)+3(7\gamma-12)q+2(5\gamma+14) \\ &+\frac{1}{(\gamma-1)q}(\gamma^2+7\gamma+8)+\frac{16}{\alpha^2(\gamma-1)q} \\ &-\frac{8}{\alpha}[-2q+5+\frac{3+\gamma}{(\gamma-1)q}]\}, \end{aligned}$$

and for the grandparent is

$$\begin{aligned} D_2\bar{\mathcal{A}} = \frac{p^2q^3\alpha^3}{16K_P^2(1-K_R)^2}(\gamma-1)^2(1-\alpha\gamma)\{&4(\gamma-1)q^2(q-5)+3(7\gamma-12)q \\ &+2(5\gamma+14)+\frac{1}{q(\gamma-1)}(\gamma^2+7\gamma+8) \\ &-\frac{p^2}{q\alpha}\}. \end{aligned}$$

Figure B.1 illustrates the direction and magnitude of HWD in the dominant genetic model for the affected grandchild and its grandparent. Also shown are the coefficients for unrelated affected and unaffected individuals.

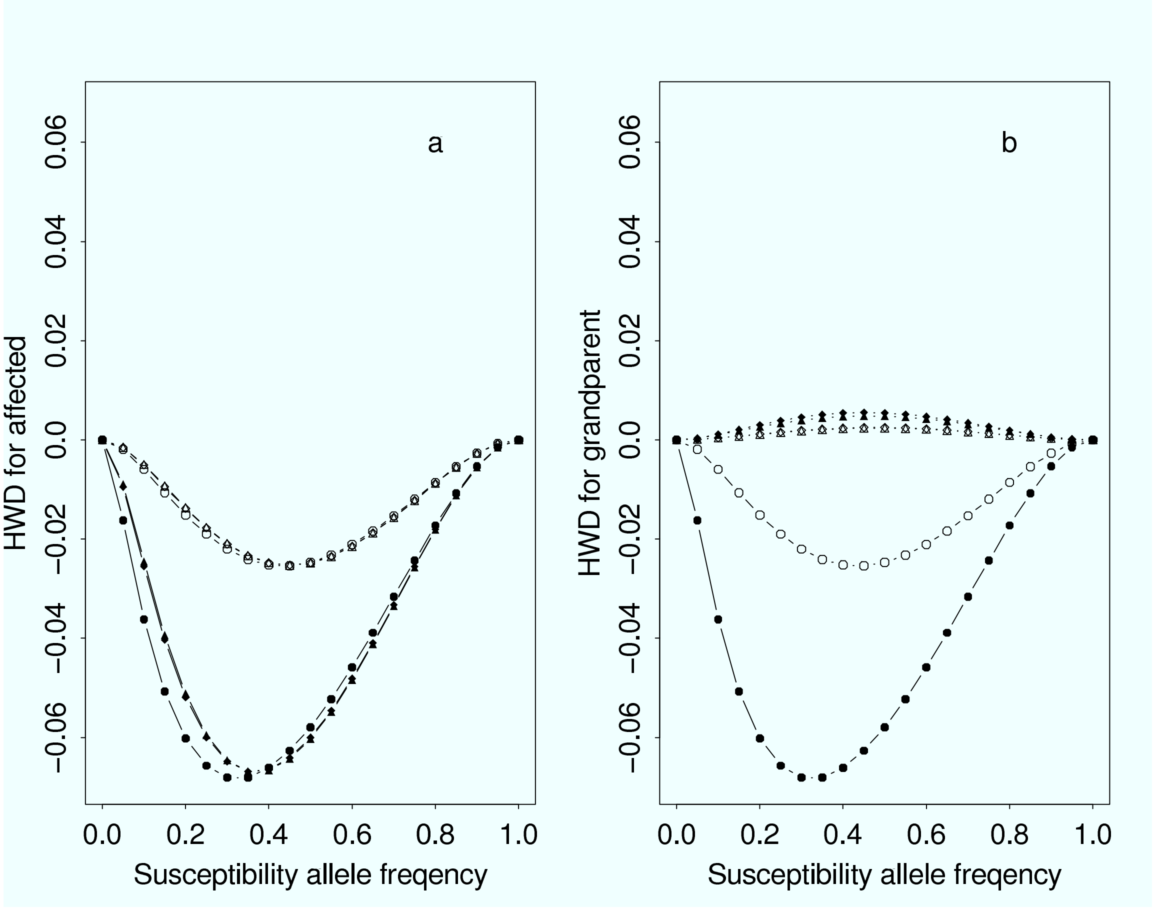


Figure B.1: HWD coefficients for the dominant genetic model as a function of the susceptibility-allele frequency for an affected grandchild (a) and their grandparent or unrelated unaffected person (b).
$K_P = 0.1$, disease status of grandparent (affected, unaffected) = (○, △), unrelated affected/unaffected ◊, open/filled symbol for $\gamma = 1.5/3$

The HWD coefficient increases in magnitude with γ for all cases and reaches a maximum when q is between 0.3 and 0.5. Panel (a) shows that the HWD coefficient of the affected grandchild is negative and similar regardless of the disease status of the grandparent. Panel (b) shows that the HWD coefficient of the grandparent is small when it is unaffected (△). It also shows that the shape and magnitude of the HWD coefficient for unrelated unaffected individuals (◊) is similar to that of the grandparent when it is unaffected (△). The HWD coefficients for the affected grandchild (panel (a)) is larger in magnitude than that of the grandparent (panel (b)) when the grandparent is unaffected (△).

Recessive Model, $\beta = 1$, $\gamma > 1$

Affected Grandchild-Grandparent Pair

When the grandparent is also affected, the HWD coefficient is the same for the affected grandchild and its grandparent and is positive.

$$D_{1\mathcal{A}} = D_{2\mathcal{A}} = \frac{p^2q^2\alpha^4}{16K_P^2K_R^2}(\gamma-1)^2[4(\gamma-1)q^3(1+q)+q(15q+8)+\frac{16}{(\gamma-1)}].$$

Discordant Grandchild-Grandparent Pair

When the grandparent is not affected, the HWD coefficients are different for affected grandchild and its grandparent and are given by

$$D_{1\bar{\mathcal{A}}} = \frac{p^2q^2\alpha^4}{16K_P^2(1-K_R)^2}(\gamma-1)^2[4(\gamma-1)q^3(1+q)+q(15q+8)+\frac{16}{\gamma-1}-\frac{8q(1+2q)}{\alpha} - \frac{32}{\alpha(\gamma-1)}+\frac{16}{\alpha^2(\gamma-1)}],$$

and

$$D_{2\bar{\mathcal{A}}} = \frac{-p^2q^2\alpha^3(1-\alpha)(\gamma-1)^2}{16K_P^2(1-K_R)^2}[4(\gamma-1)q^3(q+1)+q(15q+8)+\frac{16}{\gamma-1}+\frac{q^2}{\alpha}].$$

The HWD coefficient for the grandparent is negative.

Note that

$$D_{2\bar{\mathcal{A}}} = -\frac{q^2(\gamma-1)}{16}D_{1\bar{\mathcal{A}}},$$

therefore, the HWD coefficient for the grandparent is smaller in magnitude than that of the affected grandchild and opposite in sign.

Figure B.2 illustrates the direction and magnitude of HWD in the recessive genetic model for an affected grandchild and the grandparent. Also shown are the coefficients for unrelated affected and unaffected individuals.

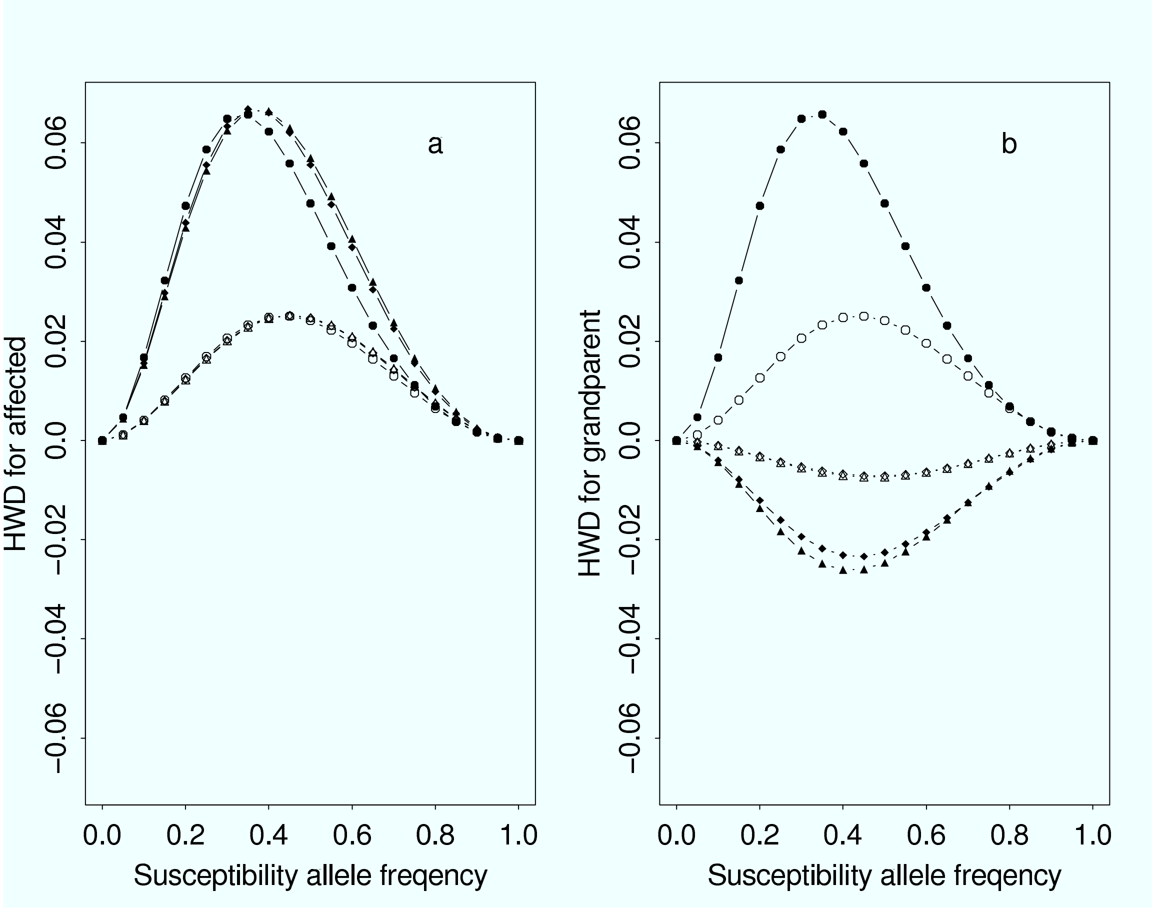


Figure B.2: HWD coefficients for the recessive genetic model for an affected grandchild (a) and their grandparent or unrelated unaffected person (b).
$K_P = 0.2$, disease status of grandparent (affected, unaffected) = (○, △), unrelated affected/unaffected ◇, open/filled symbol for $\gamma = 1.5/3$

The HWD coefficient increases in magnitude with γ for all cases and reaches a maximum when q is between 0.3 and 0.5. The HWD coefficient for the affected grandchild (panel (a)) is positive whereas that of the grandparent (panel (b)) when they are unaffected ($\triangle$), and that of an unrelated unaffected individual ($\Diamond$) is negative. The HWD coefficient is similar for unrelated unaffected individuals (panel (b), $\Diamond$) and for the grandparent when they are unaffected ($\triangle$). Panel (a) shows that the HWD coefficients for the affected grandchild is similar regardless of the disease status of the grandparent.

Additive Model, $\gamma = 2\beta - 1, \beta > 1$

Affected Grandchild-Grandparent Pair

When the grandparent is also affected, the HWD coefficient for the affected grandchild and its grandparent are the same

$$D_{1\mathcal{A}} = \frac{-p^2q^2\alpha^4}{256K_P^2K_R^2}(\gamma-1)^2[(6q+1)^2\gamma^2 + 18(4pq+1)\gamma + (6q-7)^2]$$

Discordant Grandchild-Grandparent Pair

When the grandparent is not affected, the HWD coefficient for the affected grandchild is

$$\begin{aligned} D_{1\bar{\mathcal{A}}} = \frac{-p^2q^2\alpha^4}{256K_P^2(1-K_R)^2}(\gamma-1)^2\{ \quad & (6q+1)^2\gamma^2 + 18(4pq+1)\gamma + (6q-7)^2 \\ & -\frac{16}{\alpha}[(1+6q)\gamma + 7 - 6q] + \frac{64}{\alpha^2}\}, \end{aligned}$$

and for the grandparent is

$$D_2\bar{\mathcal{A}} = \frac{-p^2q^2\alpha^4}{256K_P^2(1-K_R)^2}(\gamma-1)^2\{(6q+1)^2\gamma^2+18(4pq+1)\gamma+(6q-7)^2-\frac{4}{\alpha}(\gamma+1)+\frac{4}{\alpha^2}\}.$$

Figure B.3 illustrates the direction and magnitude of HWD in the additive genetic model for the affected grandchild and the grandparent. Also shown are the coefficients for unrelated affected and unaffected individuals.

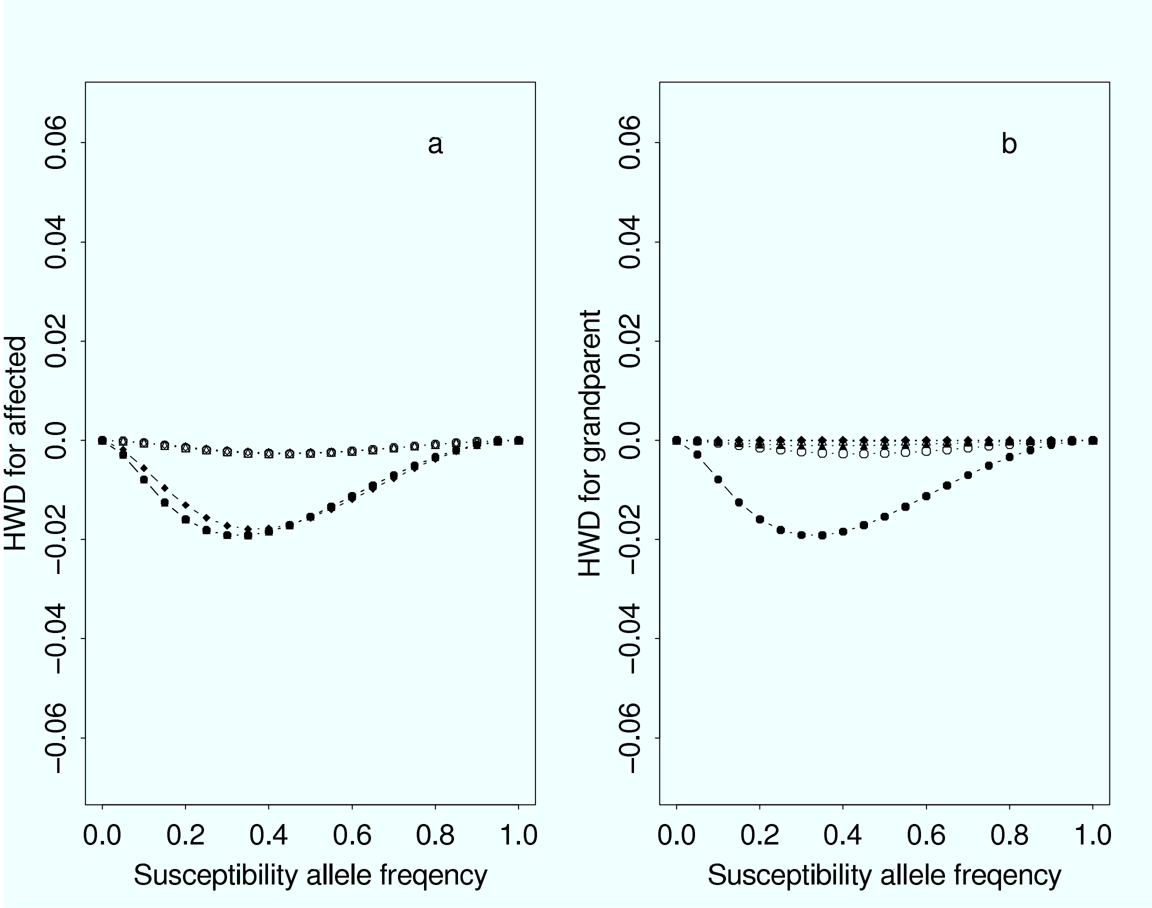


Figure B.3: HWD coefficients for the additive genetic model for an affected grandchild (a) and their grandparent or unrelated unaffected person (b).
$K_P = 0.01$, disease status of grandparent (affected, unaffected) = ($\circ$, $\triangle$), unrelated affected/unaffected $\Diamond$, open/filled symbol for $\gamma = 1.5/3$

The HWD coefficient increases in magnitude with γ for all cases and reaches a maximum when q is between 0.3 and 0.5. Panel (a) shows that the HWD coefficient of the affected grandchild is similar regardless of the disease status of the grandparent.

Panel (b) shows that the HWD coefficient of the grandparent is small when it is unaffected (△). The HWD coefficient for the affected grandchild (panel (a)) is larger in magnitude than that of the grandparent (panel (b)) when the disease status of the grandparent is unaffected (△).

Multiplicative Model $\gamma = \beta^2, \beta > 1$

Affected Grandchild-Grandparent Pair

When the grandparent is also affected, the HWD coefficient is the same for the affected grandchild and its grandparent and is given by

$$D_{1\mathcal{A}} = D_{2\mathcal{A}} = \frac{-p^2q^2\alpha^4\gamma}{16K_P^2K_R^2}[q^2\gamma^2 + (1-6pq)\gamma + 2(p-\gamma q)(q-p)\sqrt{\gamma} + p^2].$$

Discordant Grandchild-Grandparent Pair

When the grandparent is not affected, the HWD coefficient is different for the affected grandchild and its grandparent

$$D_{1\bar{\mathcal{A}}} = \frac{-p^2q^2\alpha^4\gamma}{16K_P^2(1-K_R)^2}[q^2\gamma^2 + (1-6pq)\gamma + 2(p-\gamma q)(q-p)\sqrt{\gamma} + p^2],$$

and

$$\begin{aligned} D_{2\bar{\mathcal{A}}} = \frac{-p^2q^2\alpha^3}{16K_P^2(1-K_R)^2}\{&4q^3(1+q)\gamma^3 + 2q^2(7-12q^2)\gamma^{5/2} + 4q[4-5q(1+3pq)]\gamma^2 \\ &+ [6-20pq(4pq+1)]\gamma^{3/2} - 4p[5q(3q^2-6q+2)+1]\gamma \\ &- 2(12q^2-24q+5)p^2\sqrt{\gamma} + 8 + 4q(q^3-5q^2+9q-7) \\ &+ \frac{\alpha^2\gamma+1}{\alpha}[q^2\gamma^2 + (1-6pq)\gamma - 2(\gamma q - p)(q-p)\sqrt{\gamma} + p^2]\}. \end{aligned}$$

Note that

$$D_{1\bar{\mathcal{A}}} = \frac{K_R^2}{(1-K_R)^2} D_{1\mathcal{A}}$$

so there is a relationship in the HWD coefficients for the affected grandchild when the grandparent is affected or unaffected.

Figure B.4 illustrates the direction and magnitude of HWD in the multiplicative genetic model for the affected grandchild and grandparent. Also shown are the coefficients for unrelated affected and unaffected individuals. The HWD coefficient is small in all cases but increases in magnitude with γ for all cases and reaches a maximum when q is between 0.3 and 0.5. Panel (a) shows that the HWD coefficient for the affected grandchild is slightly negative when the disease status of grandparent is unaffected ($\triangle$). The HWD coefficient of grandparent (panel (b)) is largest when it is unaffected ($\triangle$).

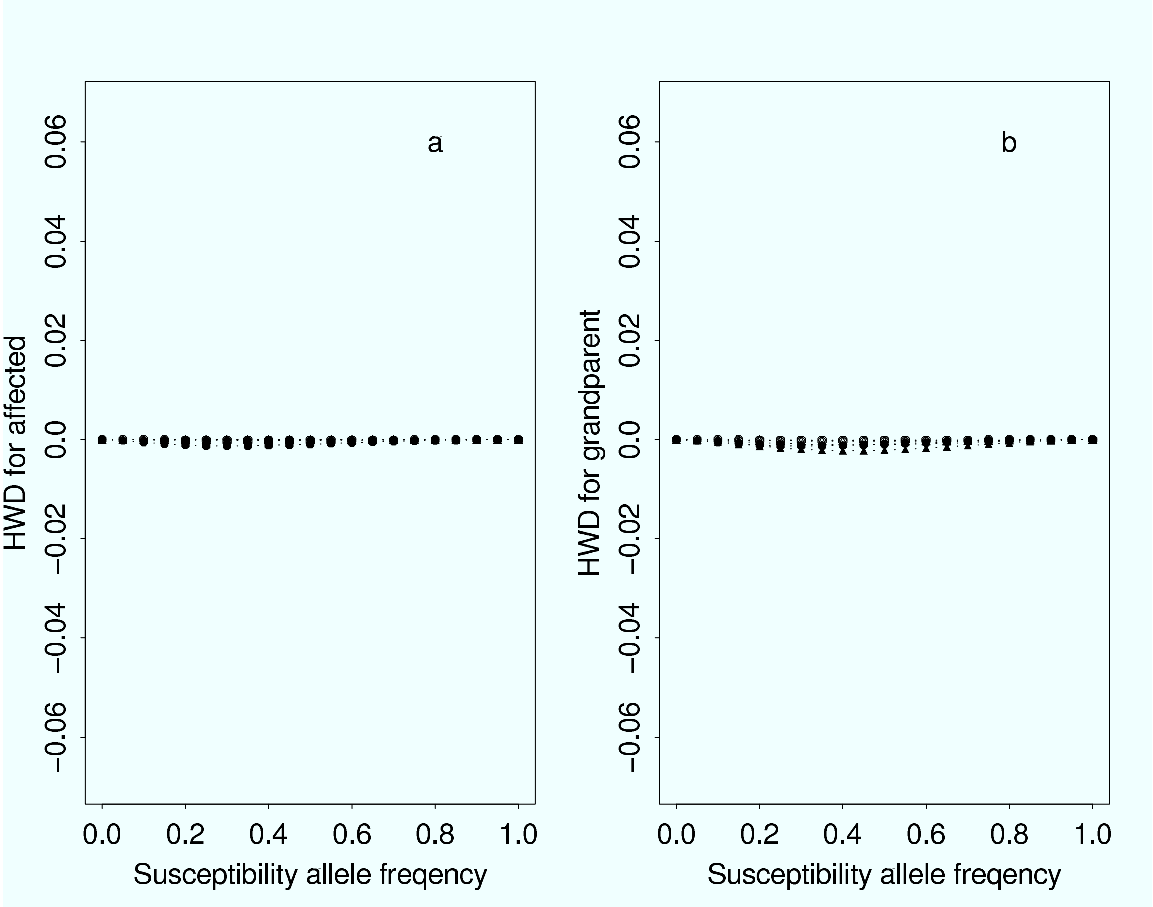


Figure B.4: HWD coefficients for the multiplicative genetic model for an affected grandchild (a) and their grandparent or unrelated unaffected person (b).
$K_P = 0.05$, disease status of grandparent (affected, unaffected) = ($\circ$, $\triangle$), unrelated affected/unaffected $\Diamond$, open/filled symbol for $\gamma = 1.5/3$

Appendix C

Distribution Of LRT For The Partially Conditional Test Of Association For Discordant Relative Pairs

To assess the distribution of the likelihood ratio test statistic for the conditional test of association for discordant relatives under the null hypothesis, the Q-Q plots of the test statistic were obtained for the χ_2^2 distribution (Figure C.1, Figure C.2 and Figure C.3). The panels in each figure correspond to the corresponding row the Table 3.7. The figures confirm that LRT statistic follows a χ^2 distribution with two degrees of freedom under the null hypothesis.

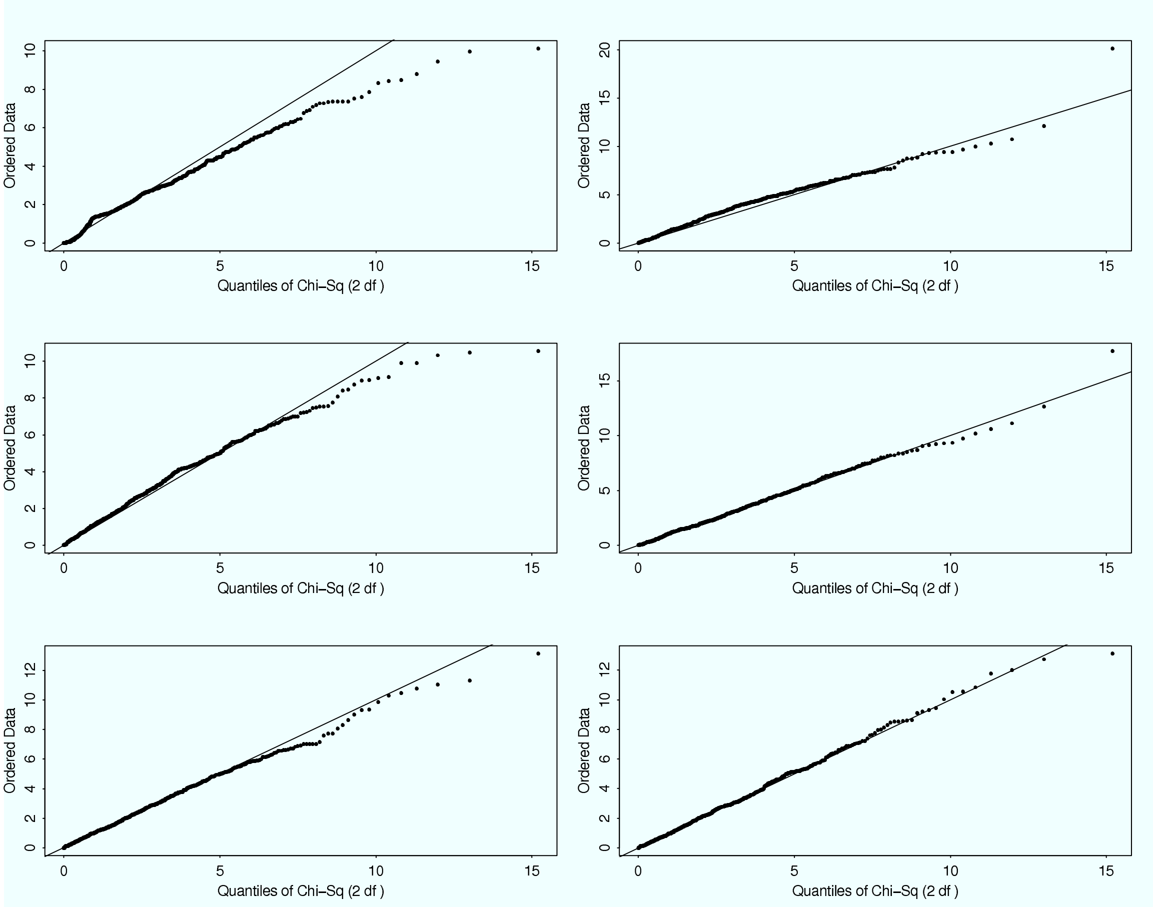


Figure C.1: Q-Q plots using the χ^2_3 distribution LLR for discordant sibling pair For $n = 300$ (Panels (a), (c) (e)), $n = 1000$ (Panels (b), (d) (f)) and $q = 0.05$ (Panels (a), (b)), $q = 0.1$ (Panels (c), (d)) and $q = 0.3$ (Panels (e), (f))

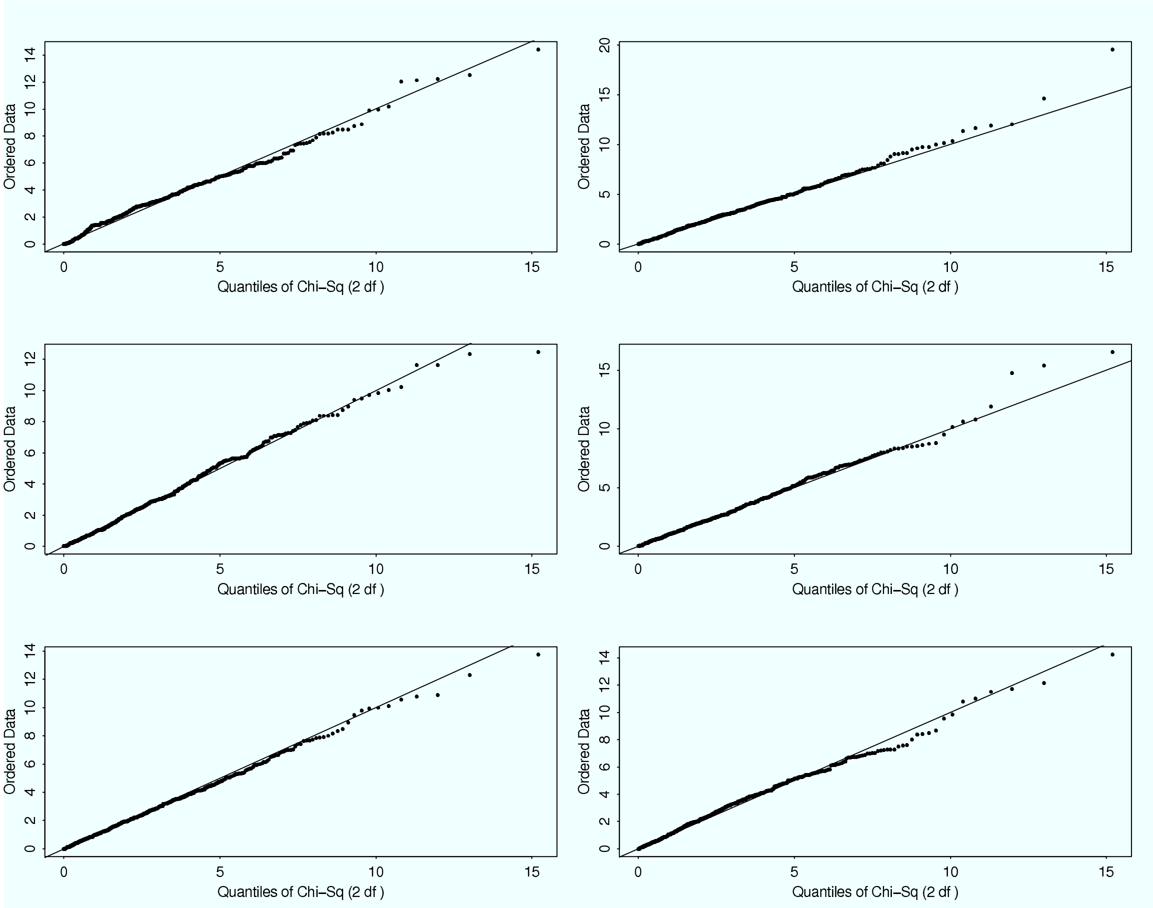


Figure C.2: Q-Q plots using the χ_2^2 distribution LLR for discordant child-parent pair For $n = 300$ (Panels (a), (c) (e)), $n = 1000$ (Panels (b), (d) (f)) and $q = 0.05$ (Panels (a), (b)), $q = 0.1$ (Panels (c), (d)) and $q = 0.3$ (Panels (e), (f))

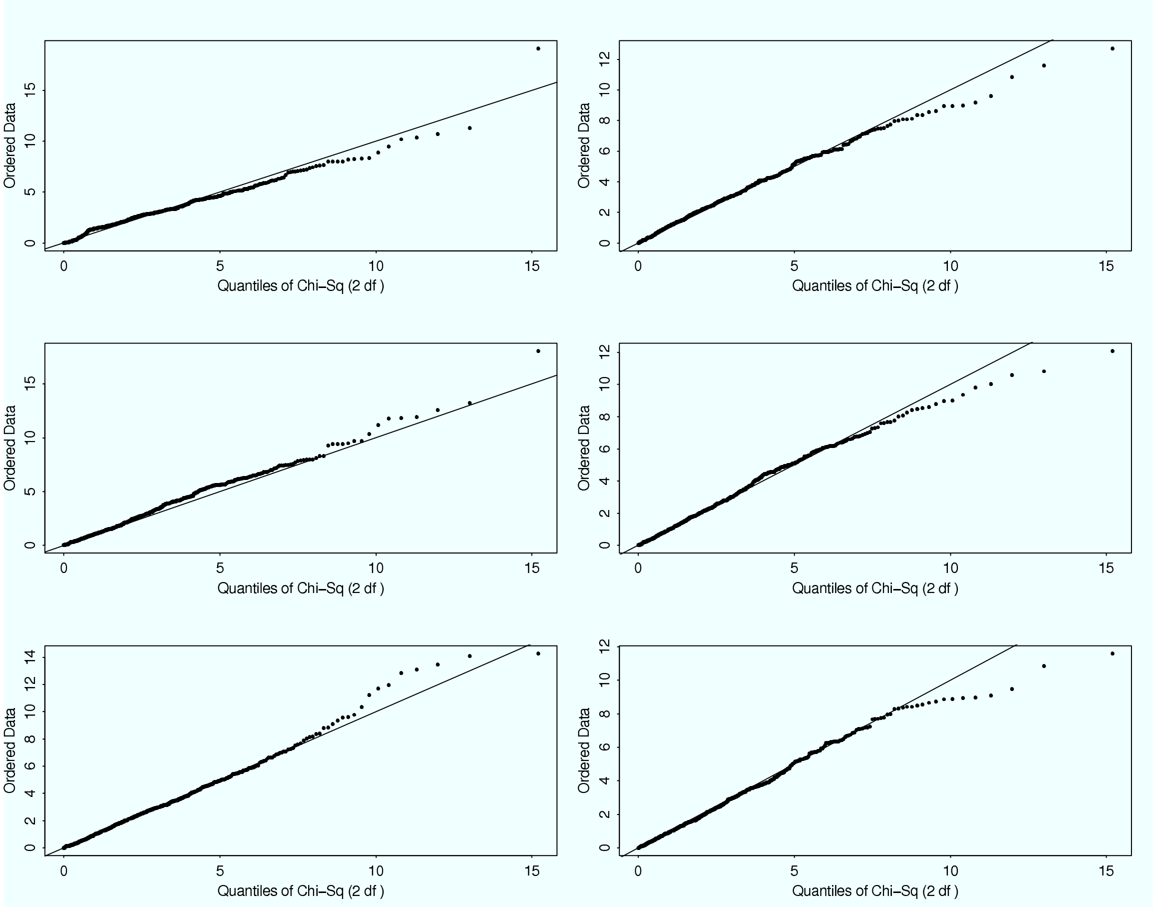


Figure C.3: Q-Q plots using the χ_2^2 distribution LLR for discordant grandchild-grandparent pair
For $n = 300$ (Panels (a), (c) (e)), $n = 1000$ (Panels (b), (d) (f)) and $q = 0.05$ (Panels (a), (b)), $q = 0.1$ (Panels (c), (d)) and $q = 0.3$ (Panels (e), (f))

Appendix D

Distribution Of LRT To Test For the Fully Conditioned Test Of Association For Discordant Relative Pairs

To assess the distribution of the likelihood ratio test statistic for the conditional test of association for discordant relatives under the null hypothesis, the Q-Q plots of the test statistic were obtained for the χ_2^2 distribution (Figure D.1, Figure D.2 and Figure D.3). The panels in each figure correspond to the corresponding row the Table 3.13. The figures confirm that LRT statistic follows a χ^2 distribution with two degrees of freedom under the null hypothesis.

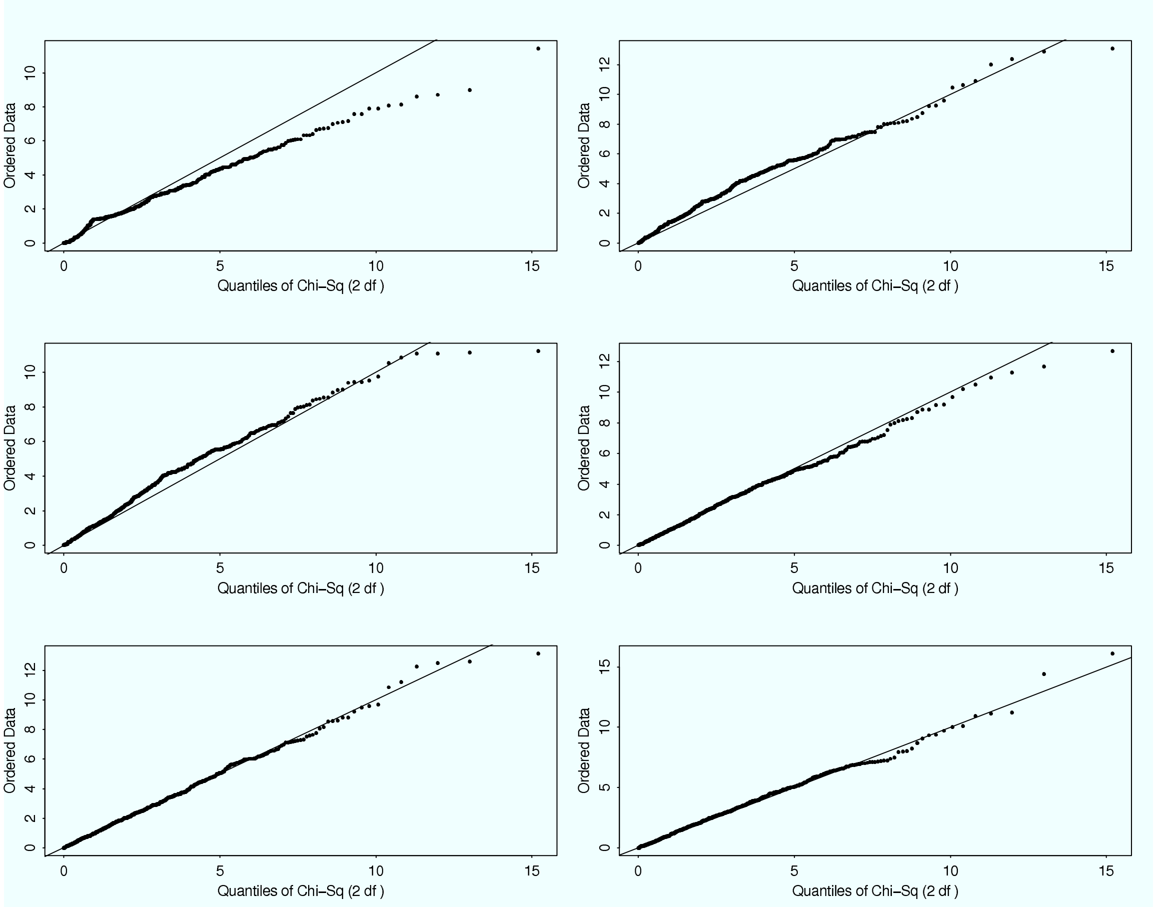


Figure D.1: Q-Q plots using the χ^2_3 distribution LLR for discordant sibling pair For $n = 300$ (Panels (a), (c) (e)), $n = 1000$ (Panels (b), (d) (f)) and $q = 0.05$ (Panels (a), (b)), $q = 0.1$ (Panels (c), (d)) and $q = 0.3$ (Panels (e), (f))

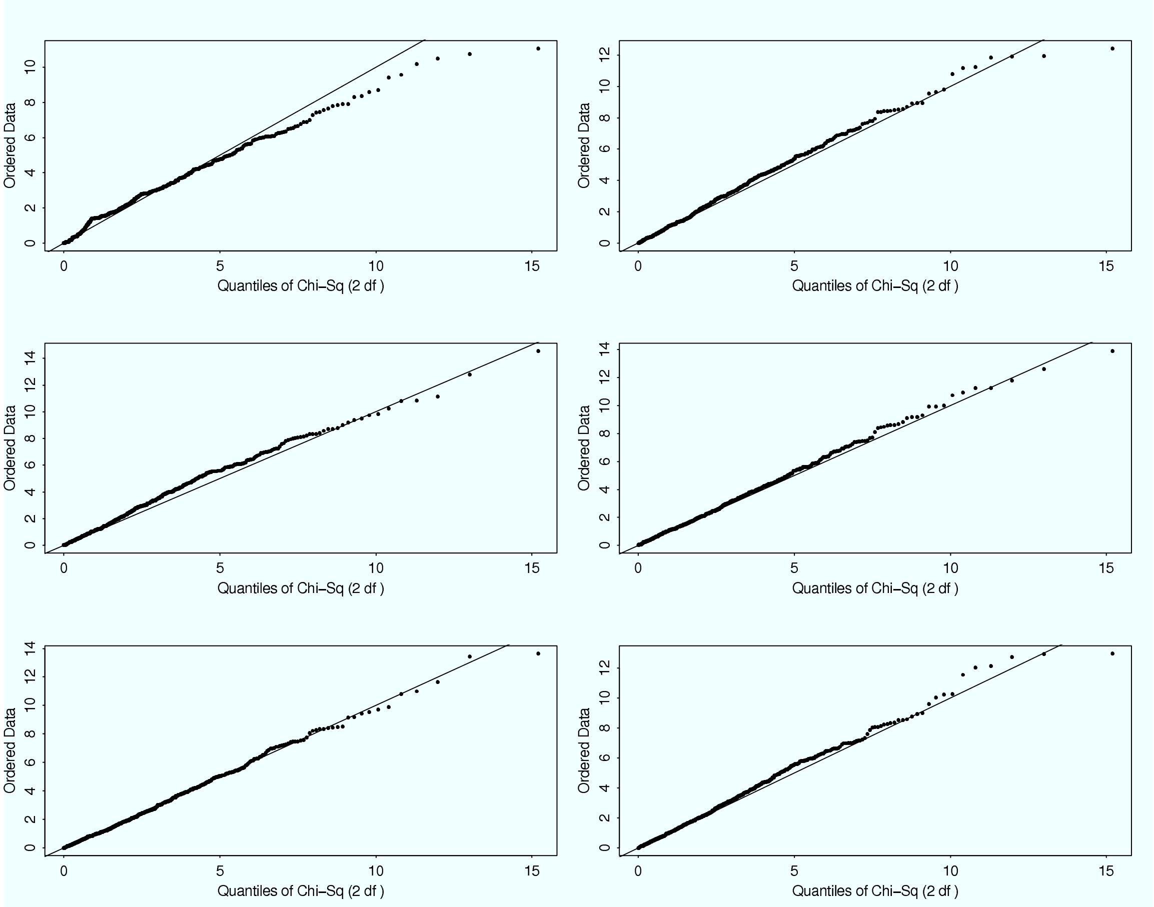


Figure D.2: Q-Q plots using the χ^2_2 distribution LLR for discordant child-parent pair For $n = 300$ (Panels (a), (c) (e)), $n = 1000$ (Panels (b), (d) (f)) and $q = 0.05$ (Panels (a), (b)), $q = 0.1$ (Panels (c), (d)) and $q = 0.3$ (Panels (e), (f))

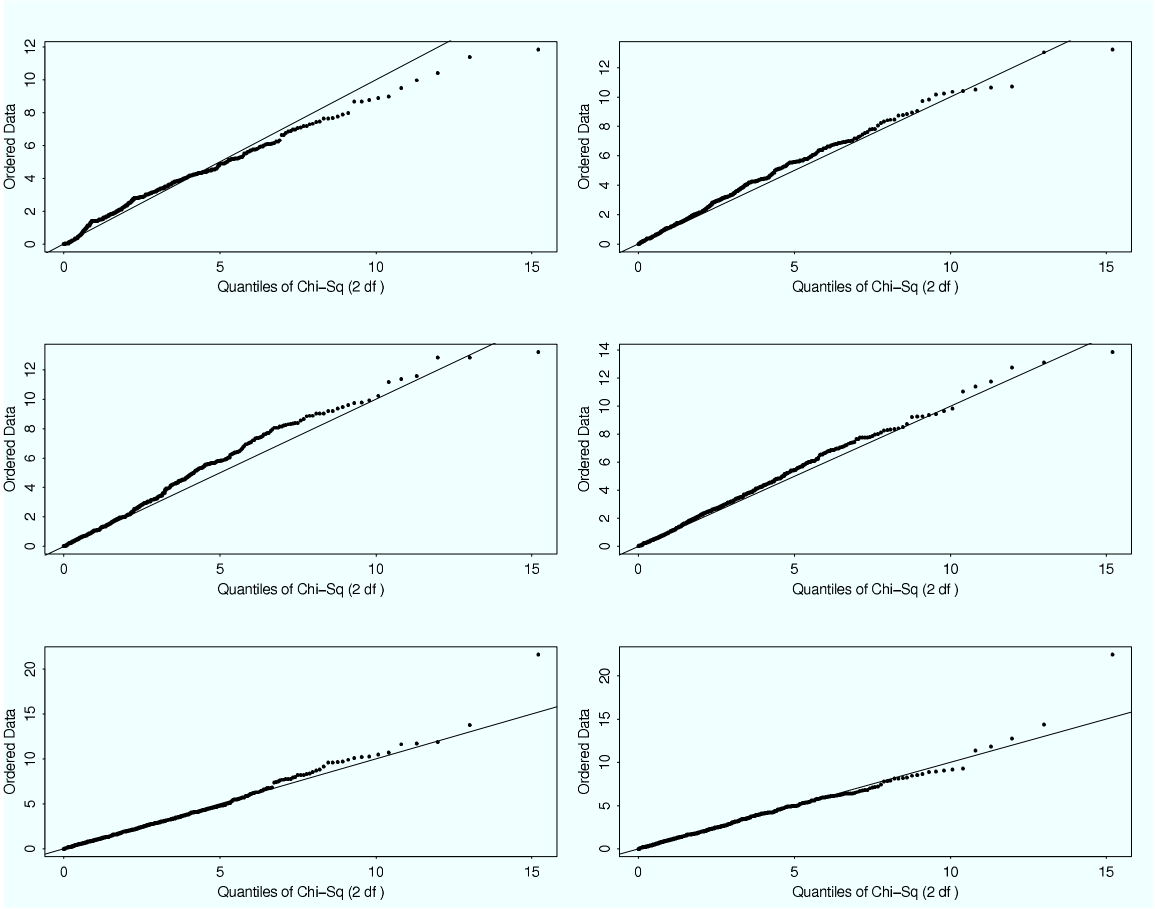


Figure D.3: Q-Q plots using the χ^2_2 distribution LLR for discordant grandchild-grandparent pair
For $n = 300$ (Panels (a), (c) (e)), $n = 1000$ (Panels (b), (d) (f)) and $q = 0.05$ (Panels (a), (b)), $q = 0.1$ (Panels (c), (d)) and $q = 0.3$ (Panels (e), (f))

Appendix E

Distribution Of LRT To Test For Genetic Effects In The Presence Of Stratification

To assess the distribution of the likelihood ratio test statistic for the test of genetic effects in the presence of stratification, i.e., H_1, stratification effects only, $(\beta = \gamma = 1)$ vs. H_a, both genetic and stratification effects, $(0 < \varepsilon < 1, \beta, \gamma)$, the Q-Q plots of the test statistic were obtained for the χ^2_2 distribution (Figure E.1, Figure E.2 and Figure E.3). The panels in each figure correspond to the 18 cases in the Table 4.2. The figures confirm that LRT statistic follows a χ^2 distribution with two degrees of freedom.

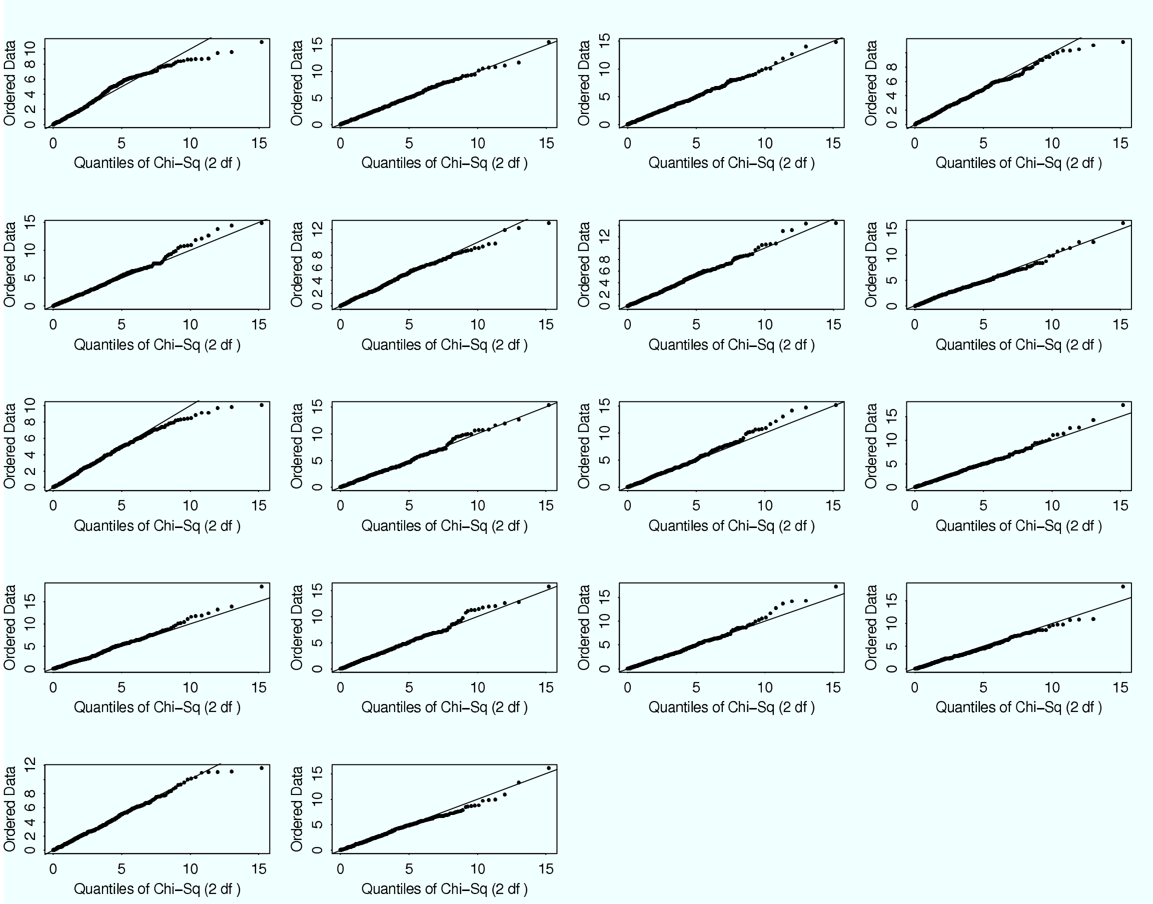


Figure E.1: Q-Q plots using the χ^2_2 distribution for the LRT statistic of H_1 vs. H_a, n = 300, r = 1.

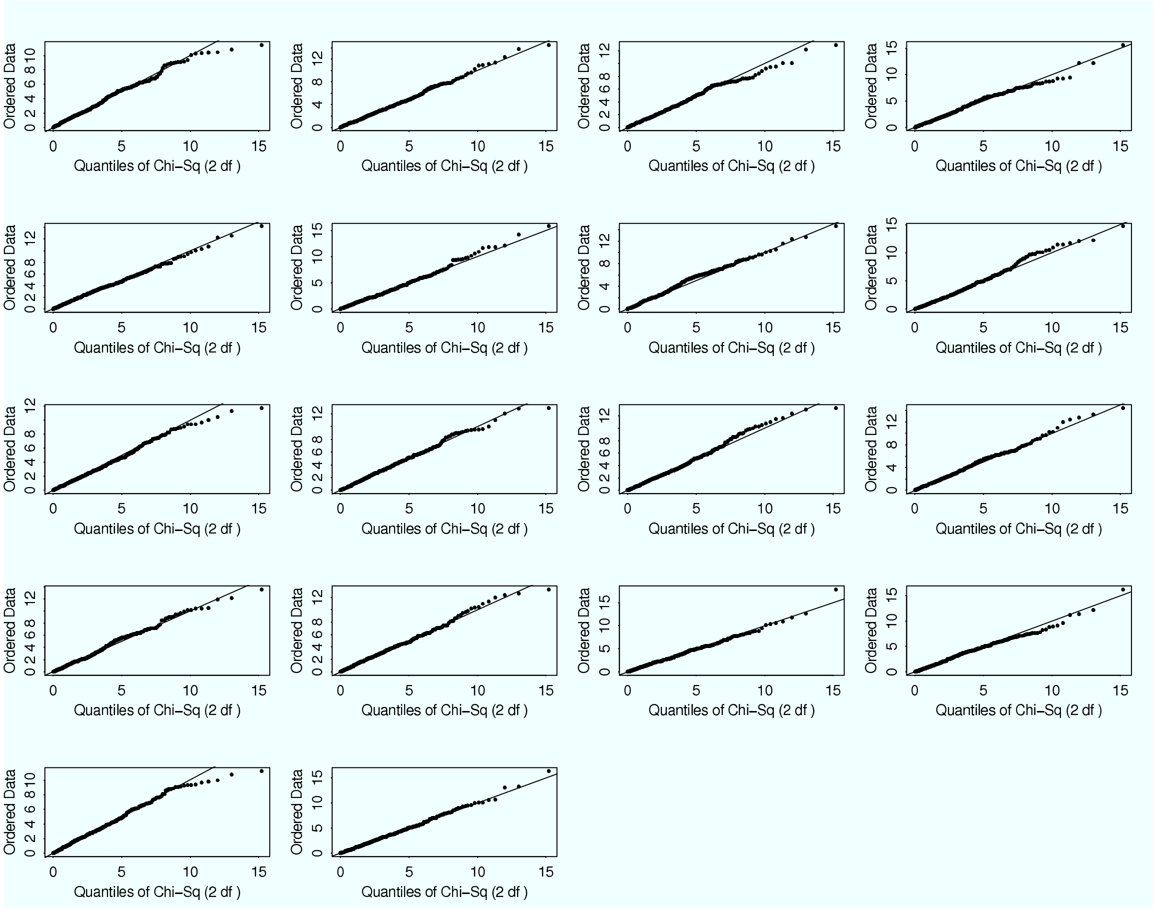


Figure E.2: Q-Q plots using the χ_2^2 distribution for the LRT statistic of H_1 vs. H_a, n = 300, r = 4.

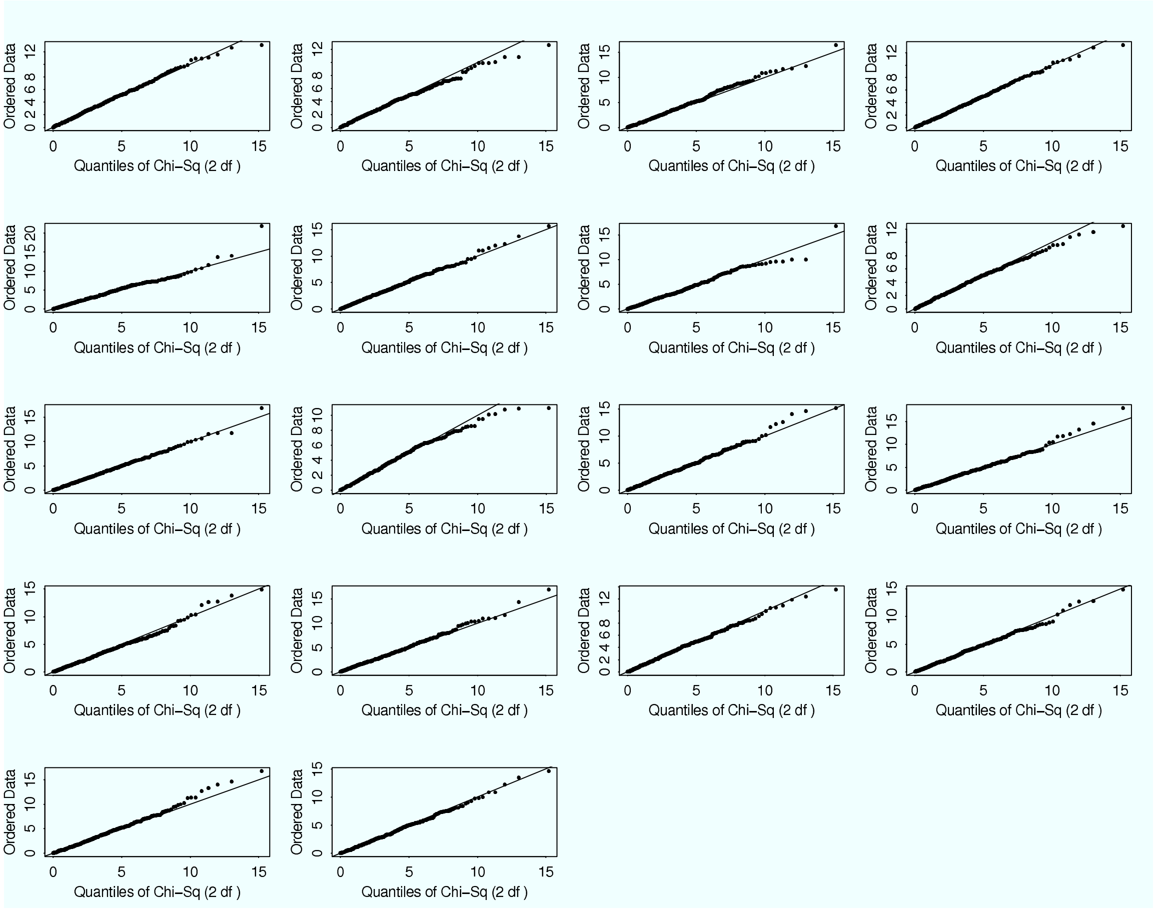


Figure E.3: Q-Q plots using the χ^2_2 distribution for the LRT statistic of H_1 vs. H_a, n = 1000, r = 1.

Appendix F

Distribution Of LRT To Test For Stratification In The Presence Of Genetic Effects

The likelihood ratio test statistic for the test of stratification in the presence of genetic effects, i.e., H_2, genetic effects only, ($\varepsilon = 0$) vs. H_a, both genetic and stratification effects, ($0 < \varepsilon < 1, \beta, \gamma$), should be distributed as a 50:50 mixture of a χ_0^2 (point mass at 0) and a χ_1^2 (Self and Liang, 1987). To verify this, the proportion of times the population proportion estimate, $\hat{\varepsilon}$, under H_a is zero was calculated (Table F.1) and the Q-Q plots of the non-zero values of the test statistic for the χ_1^2 distribution were obtained (Figure F.1, Figure F.2 and Figure F.3). Table F.1 illustrates that $\hat{\varepsilon}$ is zero approximately 50% of the time and Figures F.1, F.2 and F.3, confirm that the distribution of the non-zero values of the LRT statistic follow a χ^2 distribution with one degree of freedom. The panels in each figure correspond to the 18 cases in the Table 4.2.

Table F.1: Proportion of times the estimate of stratification proportion under H_a is zero.

Parameters			**Type 1 Errors**		
q	β	γ	$n_D = 300$ $r = 1$	$n_D = 300$ $r = 4$	$n_D = 1000$ $r = 1$
0.05	1	3	0.531	0.522	0.514
0.05	3	1	0.512	0.541	0.529
0.05	3	3	0.527	0.544	0.527
0.05	3	6	0.53	0.541	0.528
0.05	3	9	0.539	0.536	0.516
0.1	1	3	0.501	0.505	0.493
0.1	3	1	0.545	0.503	0.482
0.1	3	3	0.527	0.500	0.483
0.1	3	6	0.524	0.498	0.515
0.1	3	9	0.522	0.503	0.522
0.3	1	3	0.516	0.514	0.506
0.3	3	1	0.503	0.516	0.502
0.3	3	3	0.506	0.517	0.511
0.3	3	6	0.501	0.515	0.505
0.3	3	9	0.524	0.517	0.51

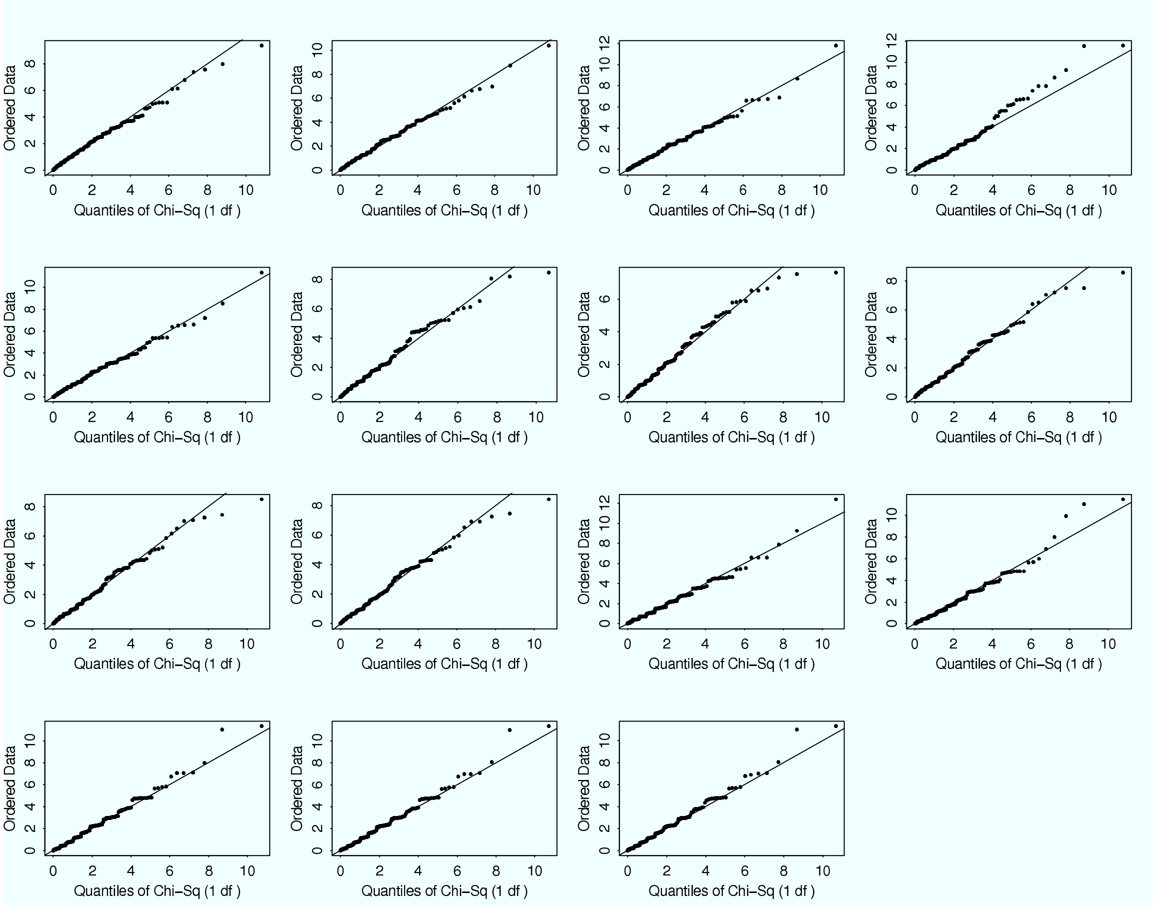


Figure F.1: Q-Q plots using the χ_1^2 distribution for the LRT statistic of H_2 vs. H_a, n = 300, r = 1.

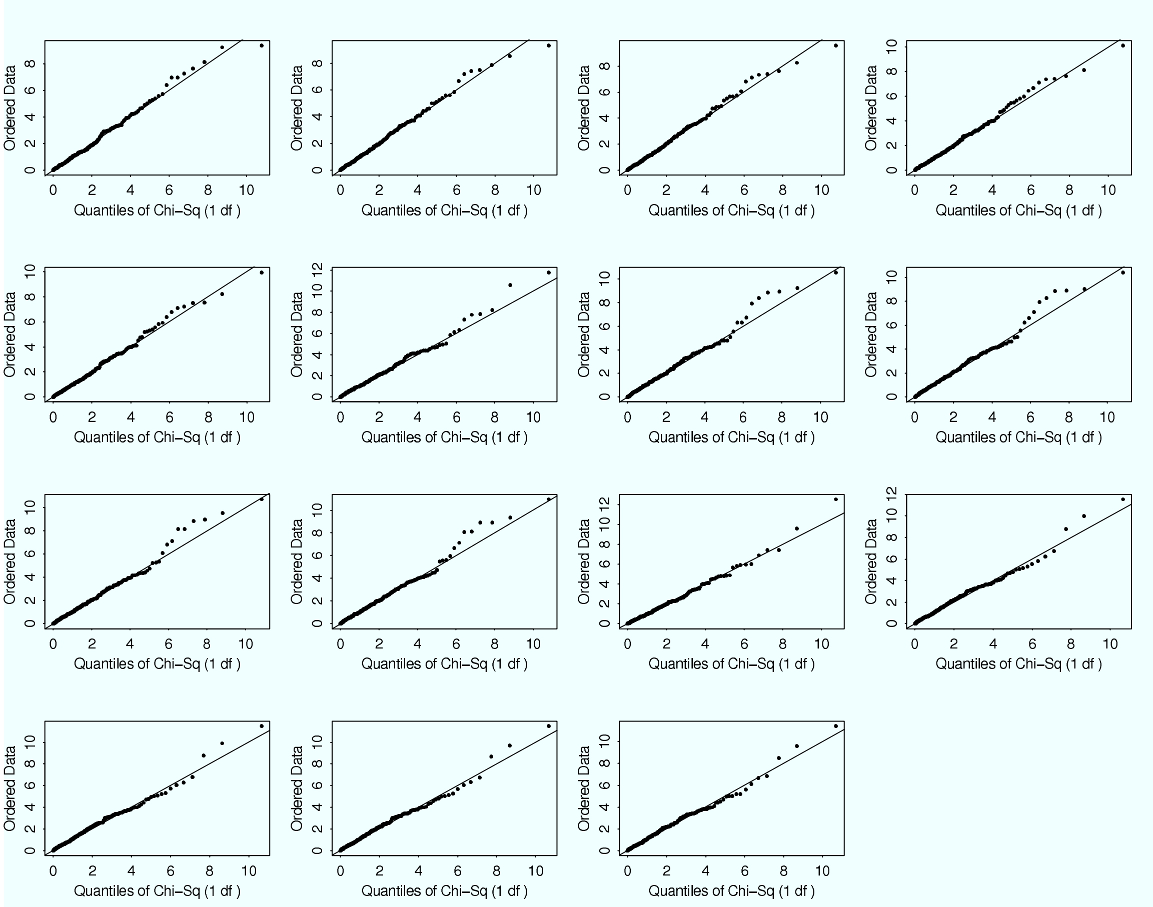


Figure F.2: Q-Q plots using the χ_1^2 distribution for the LRT statistic of H_2 vs. H_a, n = 300, r = 4.

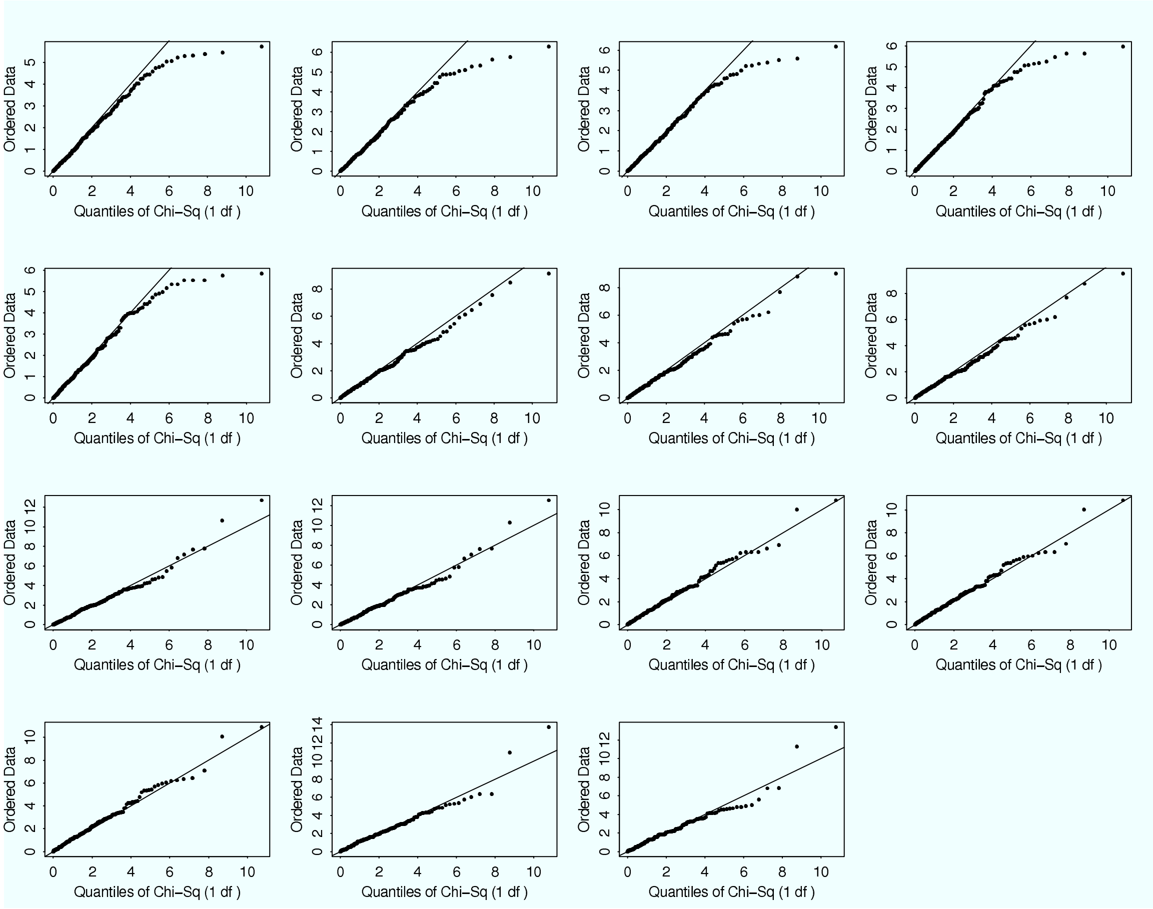


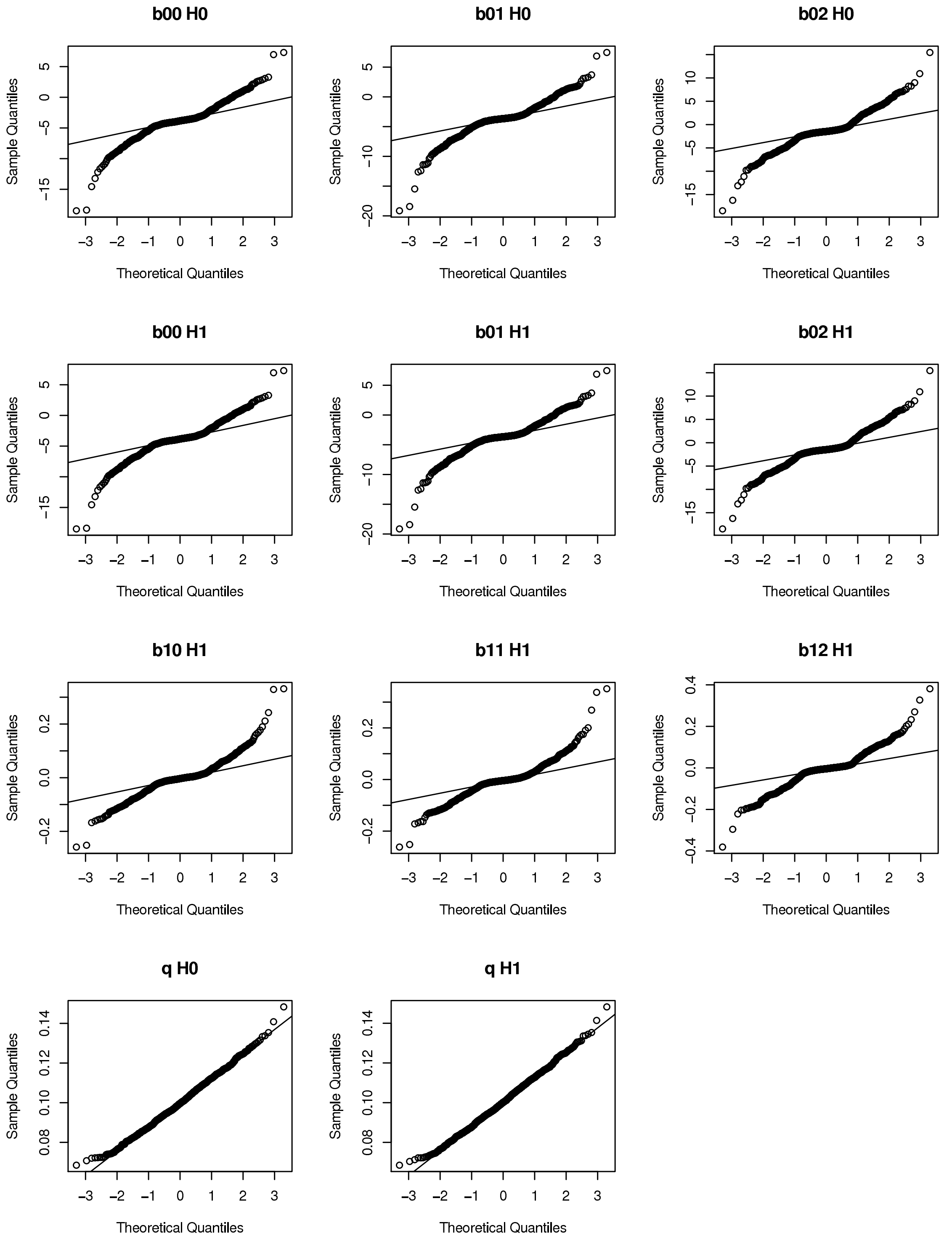
Figure F.3: Q-Q plots using the χ_1^2 distribution for the LRT statistic of H_2 vs. H_a, n = 1000, r = 1.

Appendix G

Distribution Of Estimates Of Intercepts And Slopes Of The Logistic Function For The Simulated Levels

To assess the distribution of the estimates of intercepts and slopes when the penetrance functions are modelled as a logistic function of age and data generated under null hypothesis, Q-Q plots using the $N(0,1)$ distribution were obtained. The panels in each figure correspond to the estimates of the intercepts under null hypothesis and intercepts and slopes under alternative hypothesis along with q (Figures G.1, G.2, G.3, G.4) when q is also estimated and only slopes and intercepts when q is not estimated (Figures G.5, G.6, G.7, G.8) where data are generated under null hypothesis.

The graphs depict that the estimate of q is normally distributed where as the estimates of the slopes and intercepts under null and alternative hypothesis have a sharp peak and fat tails. It also shows that the estimates have larger variance when the data is estimated under the alternative hypothesis.



b00 H0

b01 H0

b02 H0

b00 H1

b01 H1

b02 H1

b10 H1

b11 H1

b12 H1

q H0

q H1

Figure G.1: Q-Q plots using the $N(0, 1)$ distribution for the estimates of intercepts under the null and alternative hypothesis when q is also estimated for $q = 0.1$ and $n = 300$.

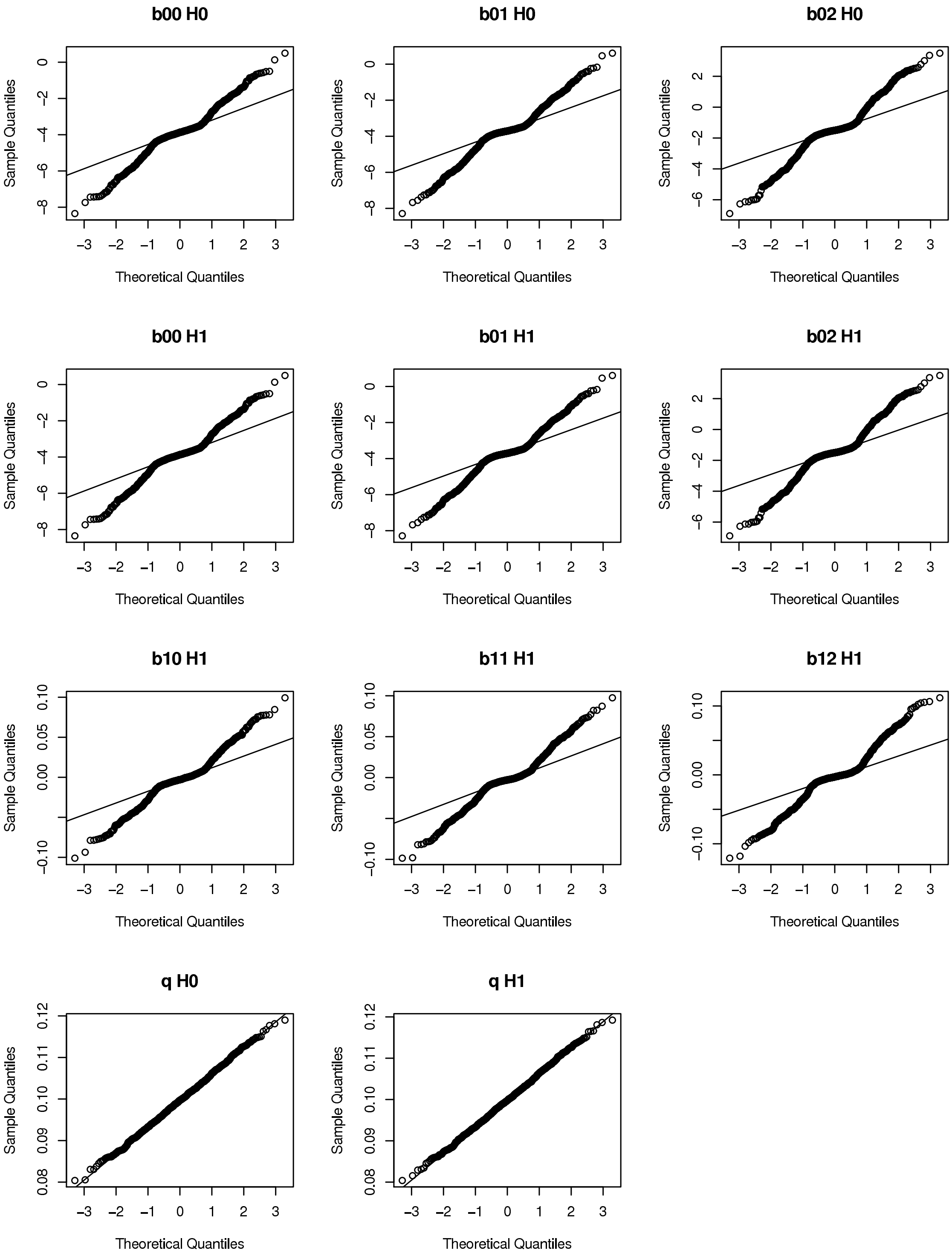


Figure G.2: Q-Q plots using the $N(0, 1)$ distribution for the estimates of intercepts under the null and alternative hypothesis when q is also estimated for $q = 0.1$ and $n = 1000$.

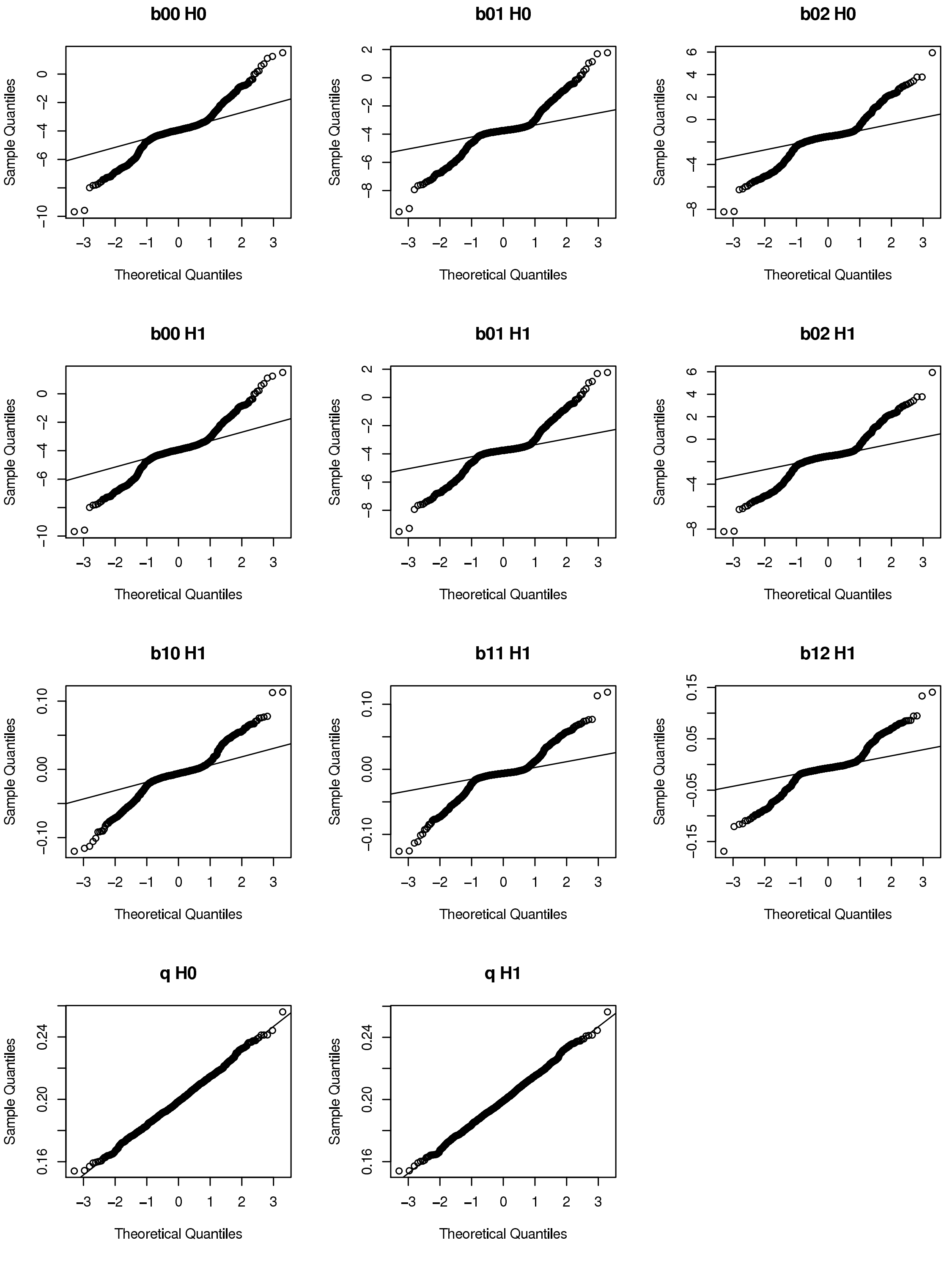


Figure G.3: Q-Q plots using the $N(0, 1)$ distribution for the estimates of intercepts under the null and alternative hypothesis when q is also estimated for $q = 0.2$ and $n = 300$.

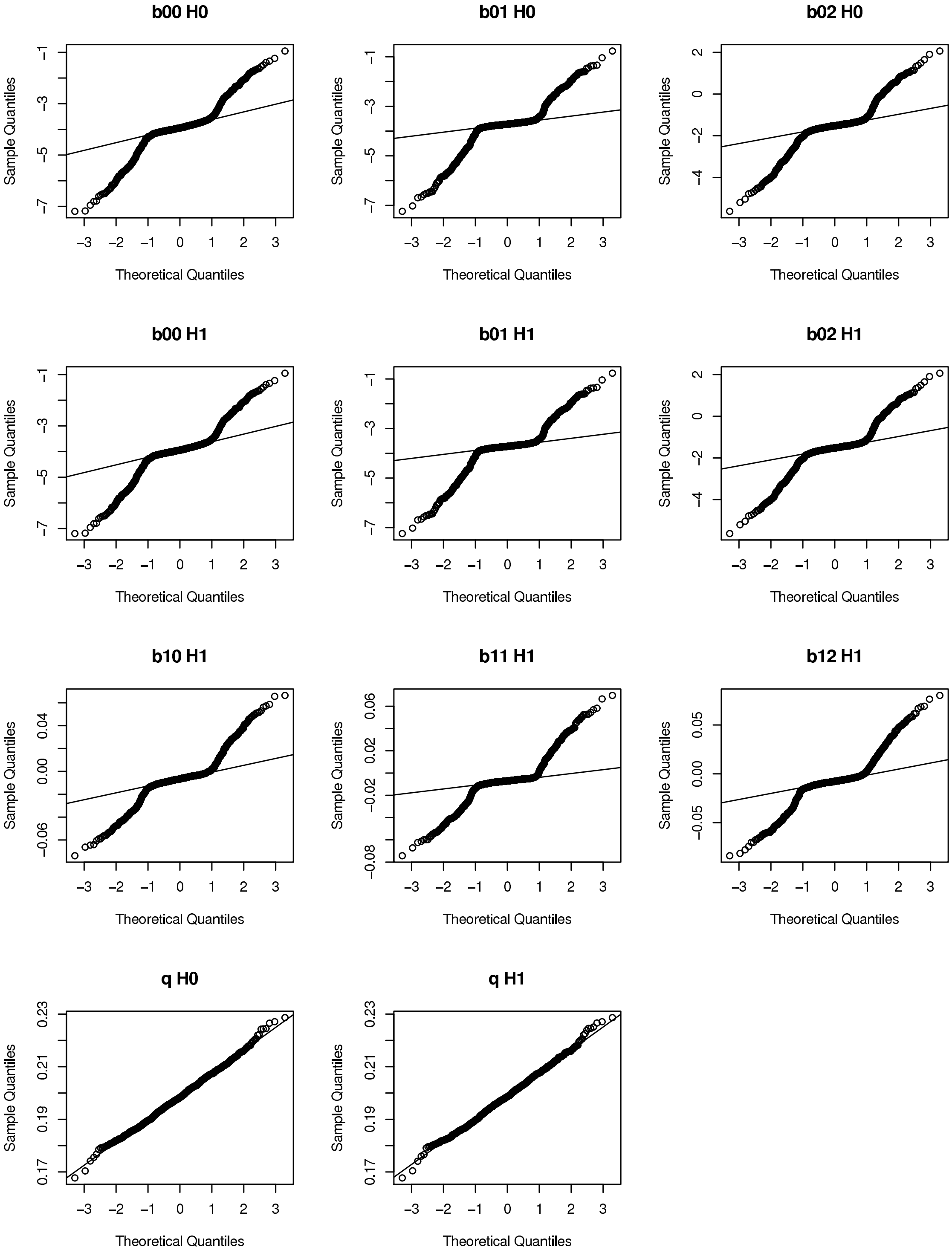


Figure G.4: Q-Q plots using the $N(0, 1)$ distribution for the estimates of intercepts under the null and alternative hypothesis when q is also estimated for $q = 0.2$ and $n = 1000$.

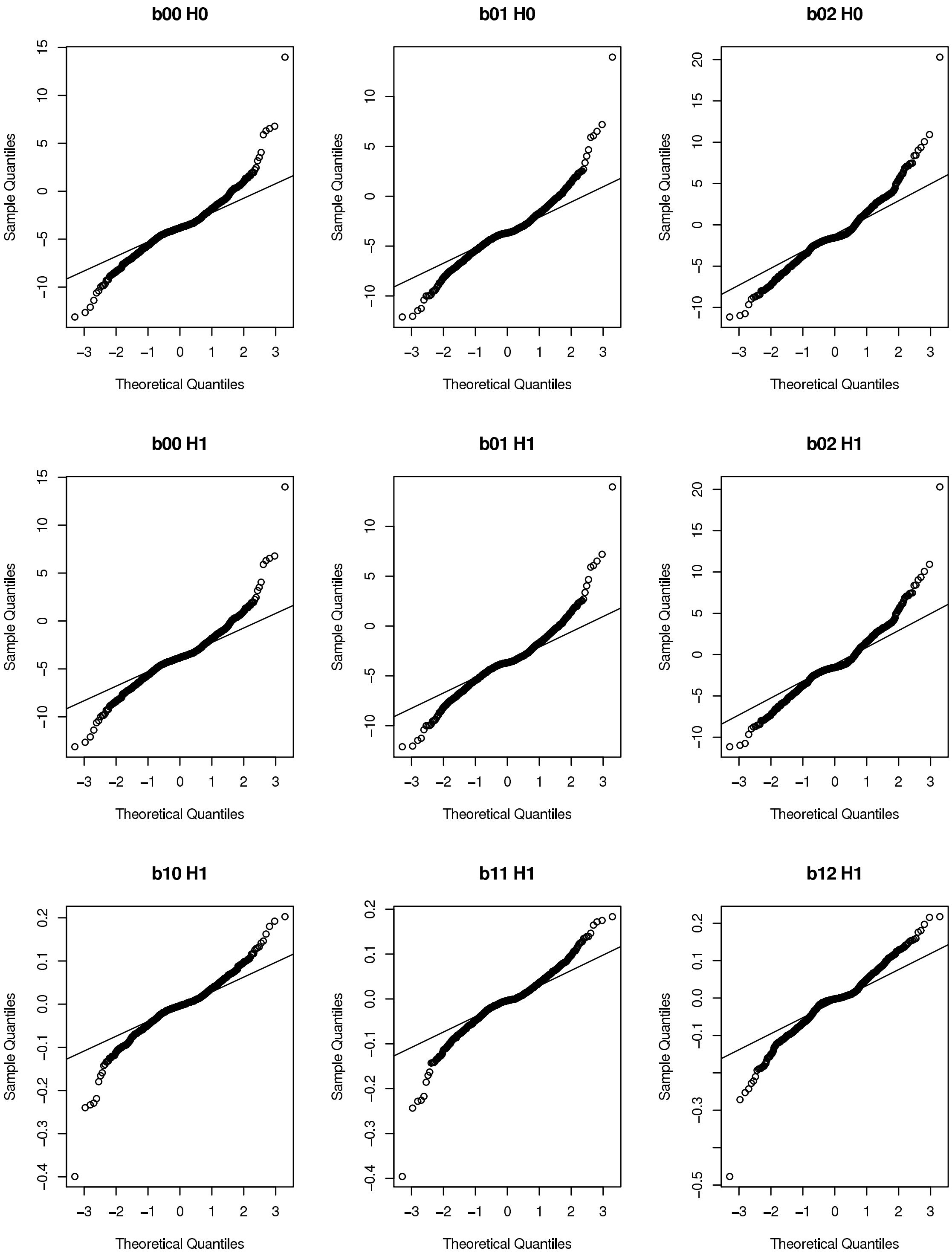


Figure G.5: Q-Q plots using the $N(0,1)$ distribution for the estimates of intercepts under the null and alternative hypothesis when q is not estimated for $q = 0.1$ and $n = 300$.

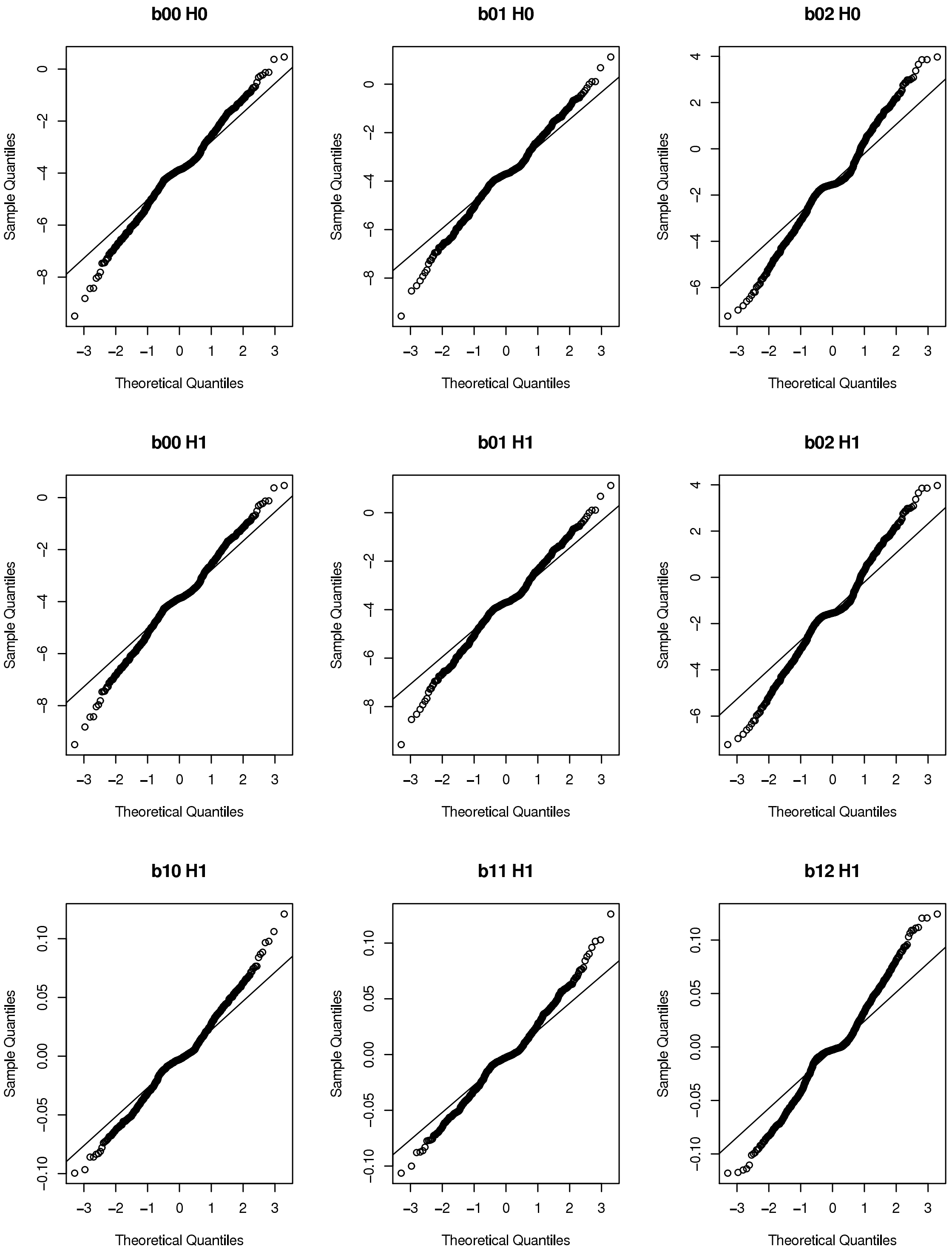


Figure G.6: Q-Q plots using the $N(0, 1)$ distribution for the estimates of intercepts under the null and alternative hypothesis when q is not estimated for $q = 0.1$ and $n = 1000$.

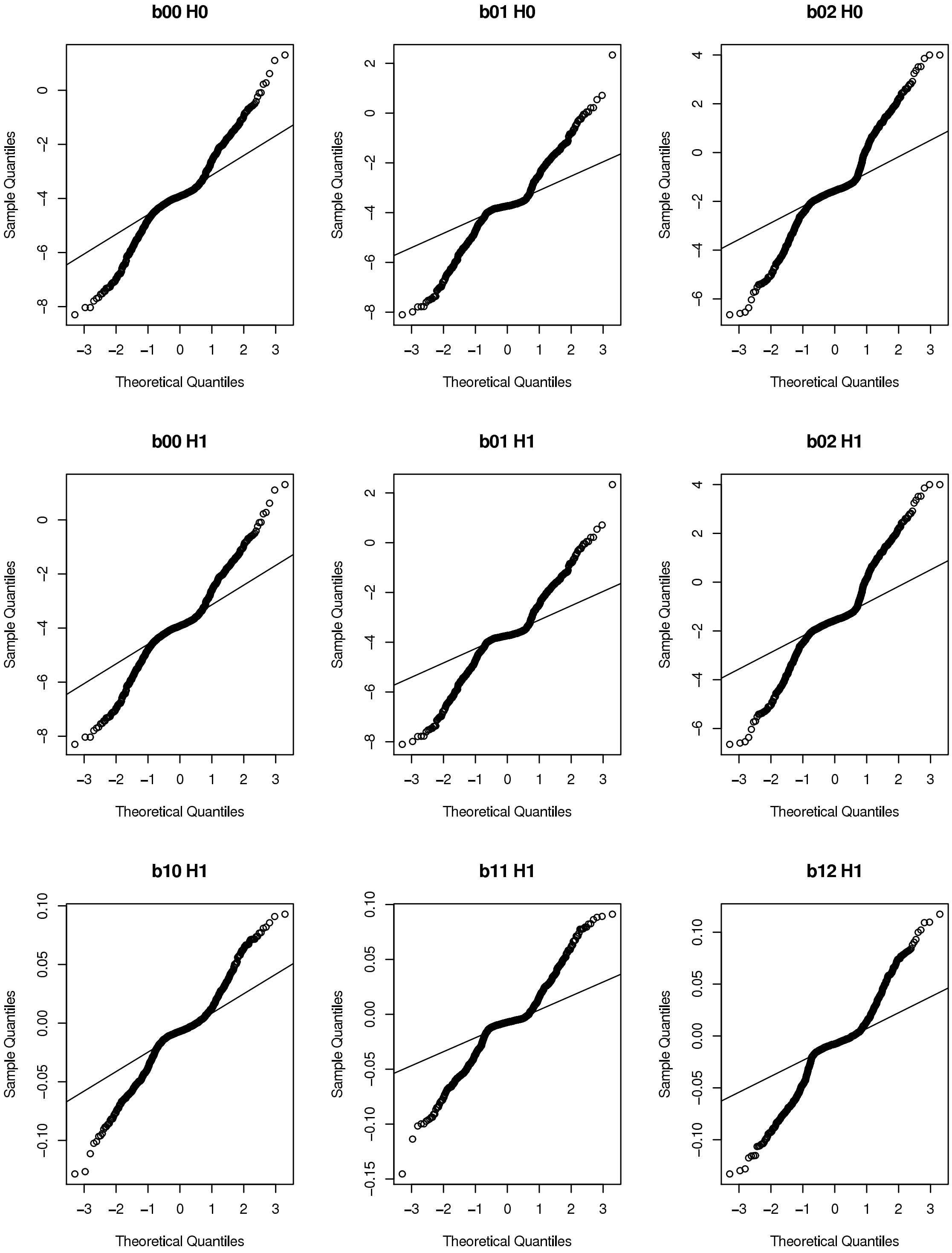


Figure G.7: Q-Q plots using the $N(0,1)$ distribution for the estimates of intercepts under the null and alternative hypothesis when q is not estimated for $q = 0.2$ and $n = 300$.

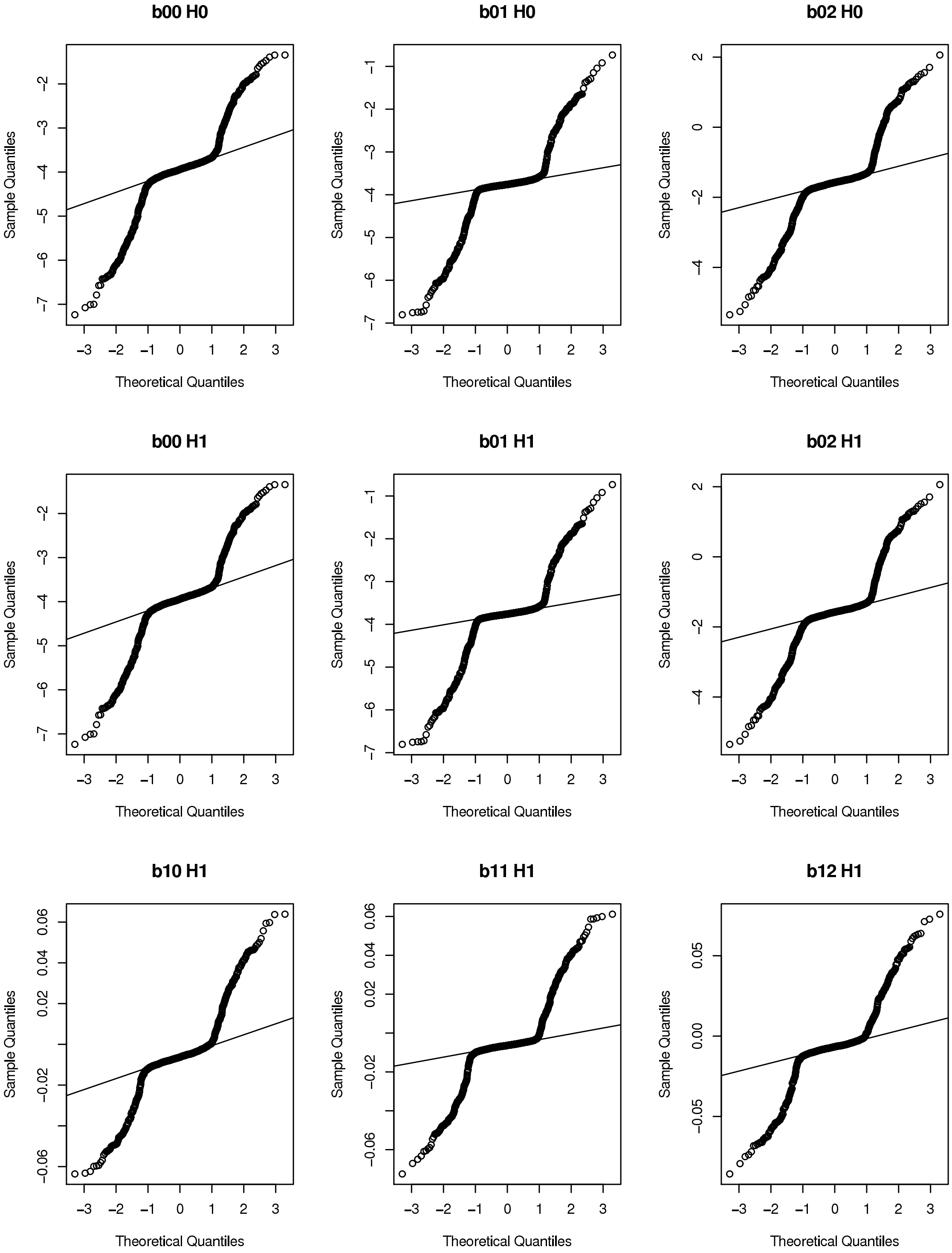


Figure G.8: Q-Q plots using the $N(0,1)$ distribution for the estimates of intercepts under the null and alternative hypothesis when q is not estimated for $q = 0.2$ and $n = 1000$.

Appendix H

Distribution Of Estimates Of Intercepts And Slopes Of The Logistic Function For The Simulated Power

To assess the distribution of the estimates of intercepts and slopes when the penetrance functions are modelled as a logistic function of age and data generated under alternative hypothesis, Q-Q plots using the $N(0,1)$ distribution were obtained. The panels in each figure correspond to the estimates of the intercepts under null hypothesis and intercepts and slopes under alternative hypothesis along with q (Figures H.1, H.2, H.3, H.4) when q is also estimated and only slopes and intercepts when q is not estimated (Figures H.5, H.6, H.7, H.8) where data are generated under null hypothesis.

The graphs depict that the estimate of q is normally distributed where as the estimates of the slopes and intercepts under null and alternative hypothesis have a sharp peak and fat tails. It also shows that the estimates have larger variance when the data is estimated under the alternative hypothesis. The figures also illustrate that the estimates are closer to normal distribution for small values of the sample size and minor allele frequency.

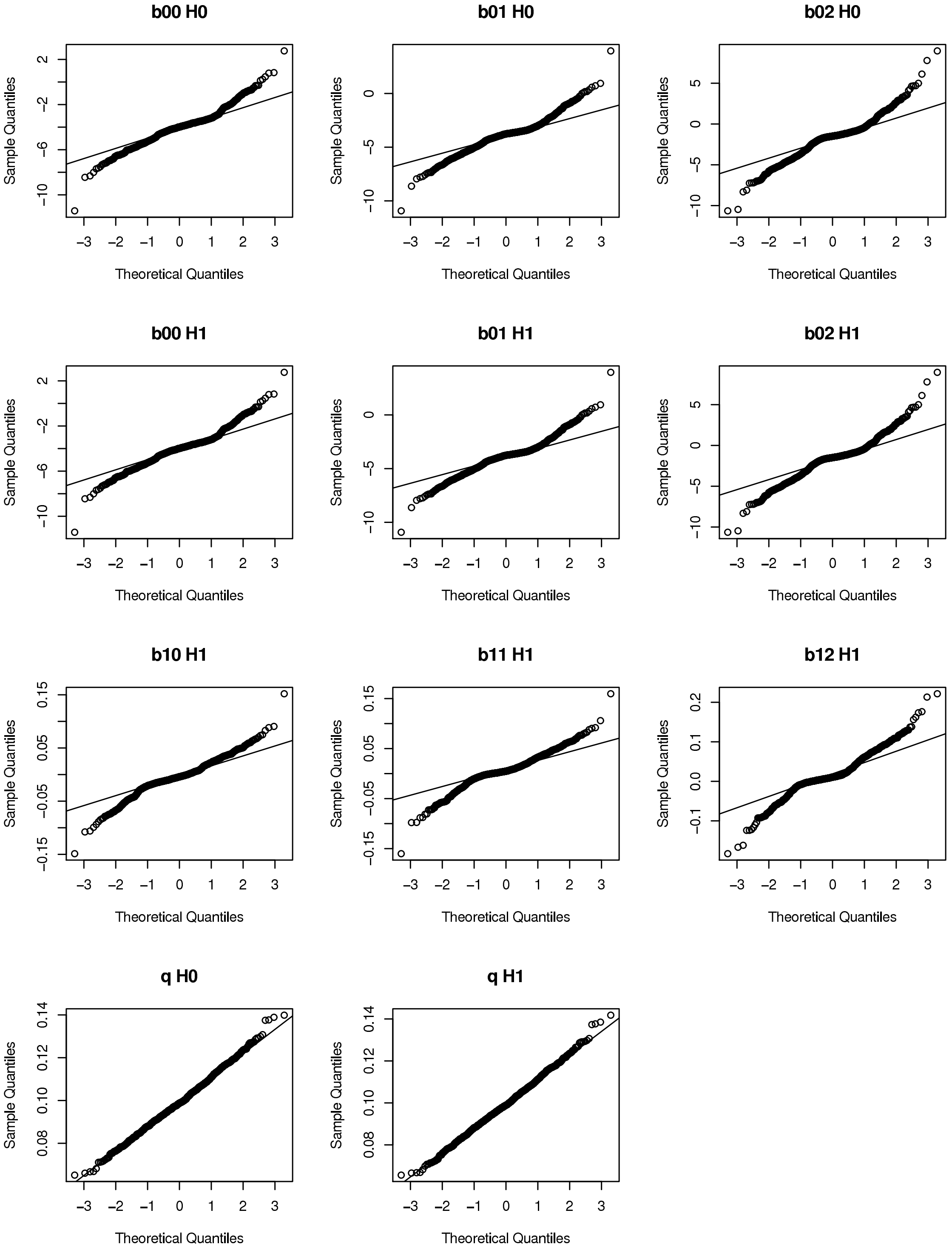


Figure H.1: Q-Q plots using the $N(0,1)$ distribution for the estimates of intercepts under the null and alternative hypothesis when q is also estimated for $q = 0.1$ and $n = 300$.

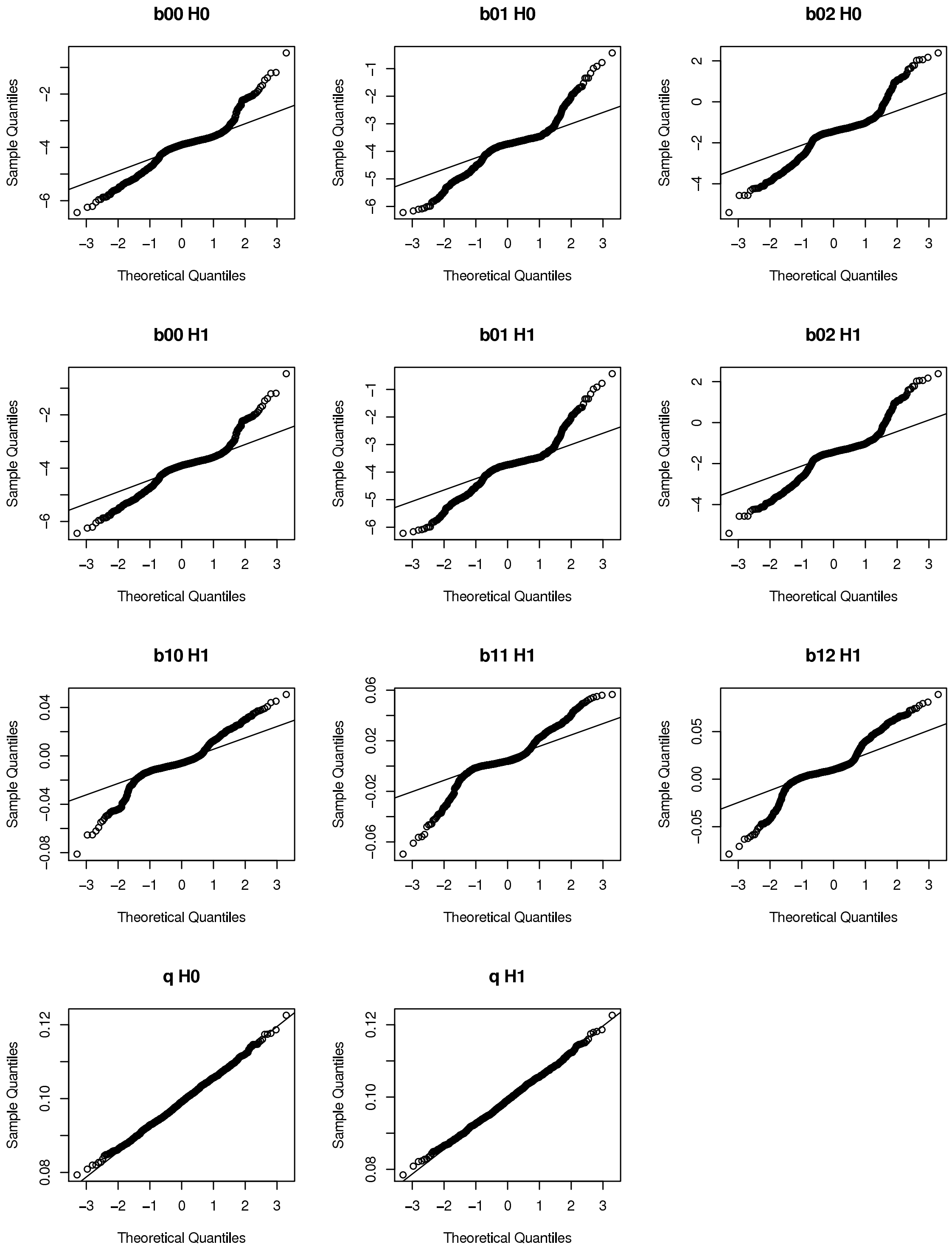


Figure H.2: Q-Q plots using the $N(0, 1)$ distribution for the estimates of intercepts under the null and alternative hypothesis when q is also estimated for $q = 0.1$ and $n = 1000$.

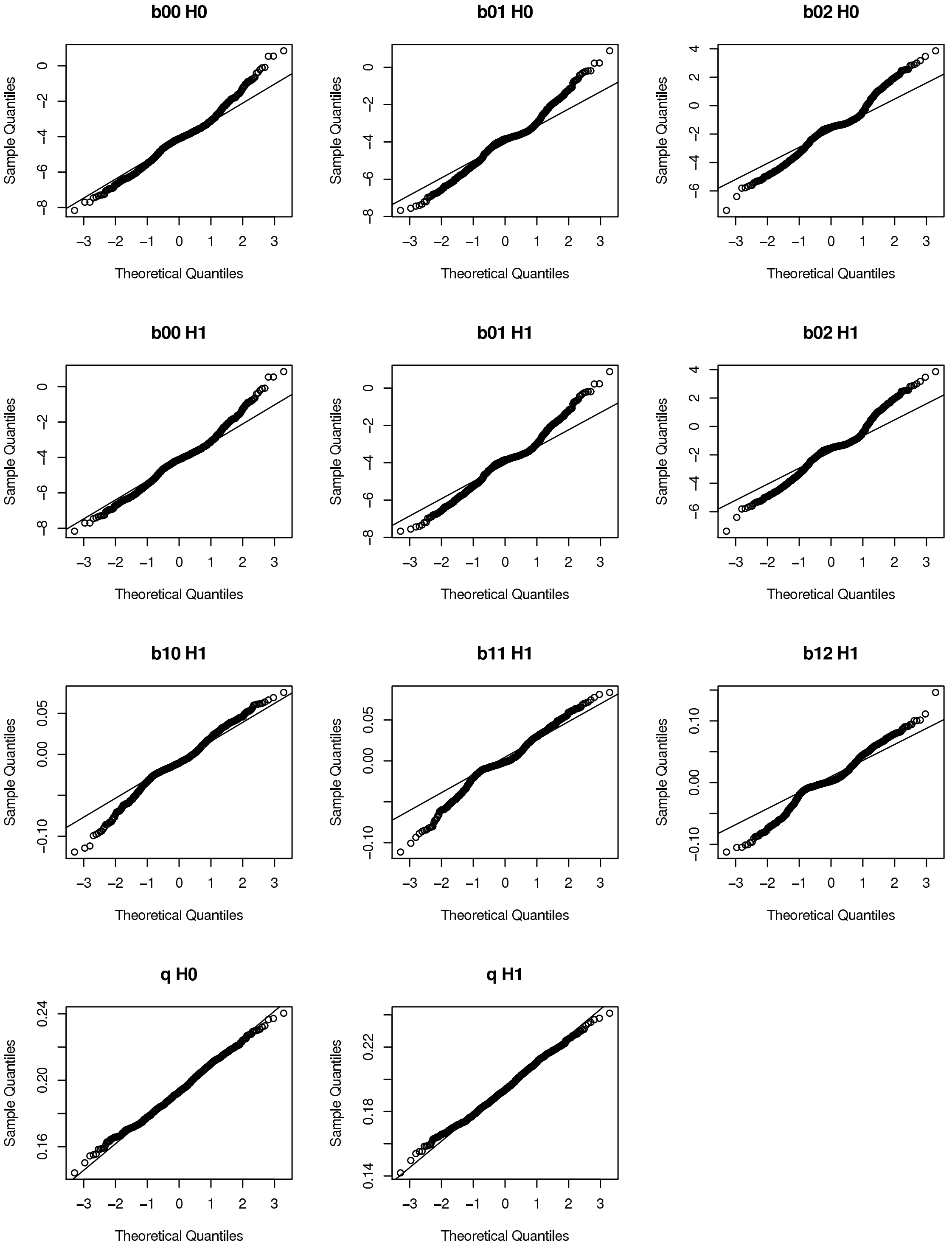


Figure H.3: Q-Q plots using the $N(0,1)$ distribution for the estimates of intercepts under the null and alternative hypothesis when q is also estimated for $q = 0.2$ and $n = 300$.

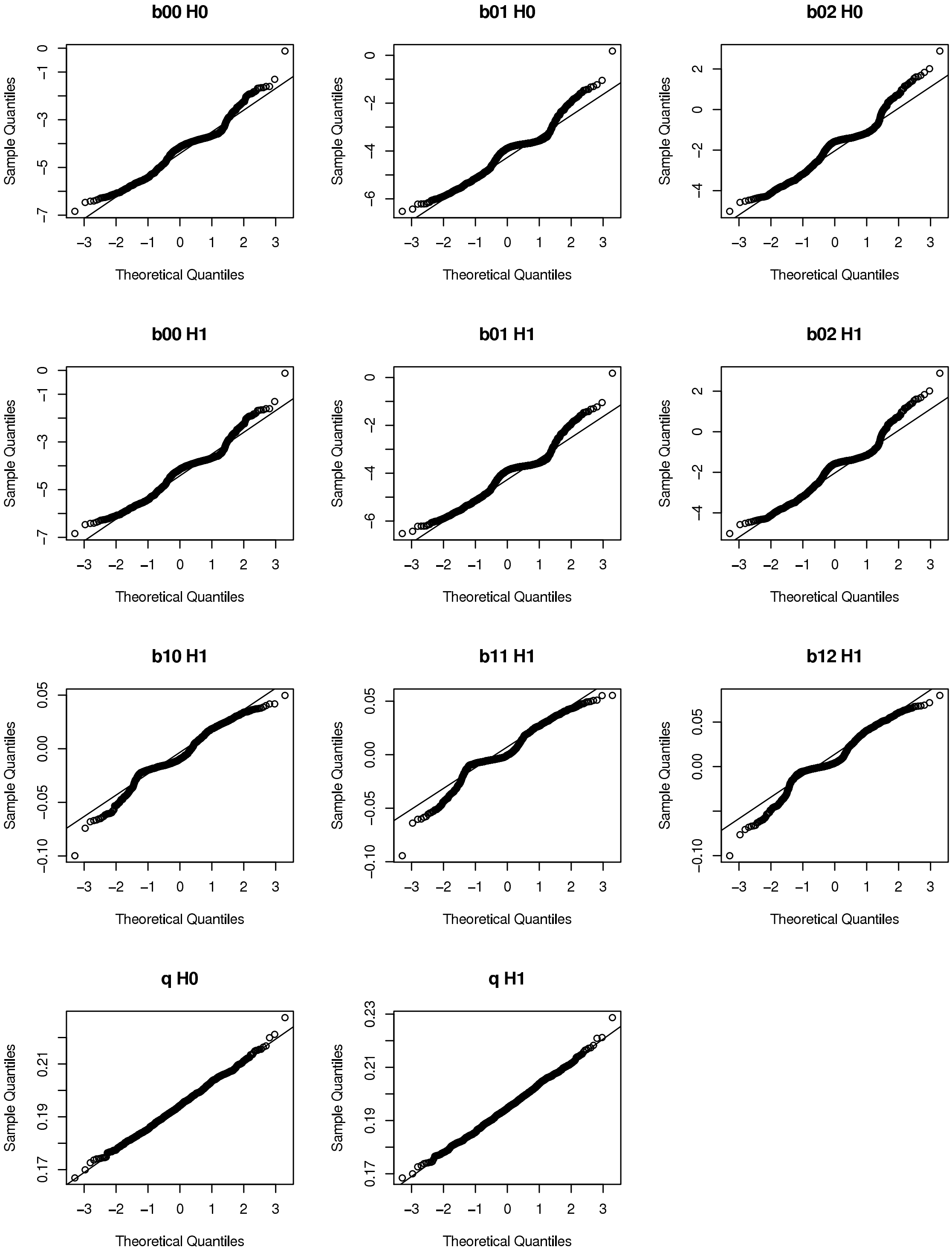


Figure H.4: Q-Q plots using the $N(0,1)$ distribution for the estimates of intercepts under the null and alternative hypothesis when q is also estimated for $q = 0.2$ and $n = 1000$.

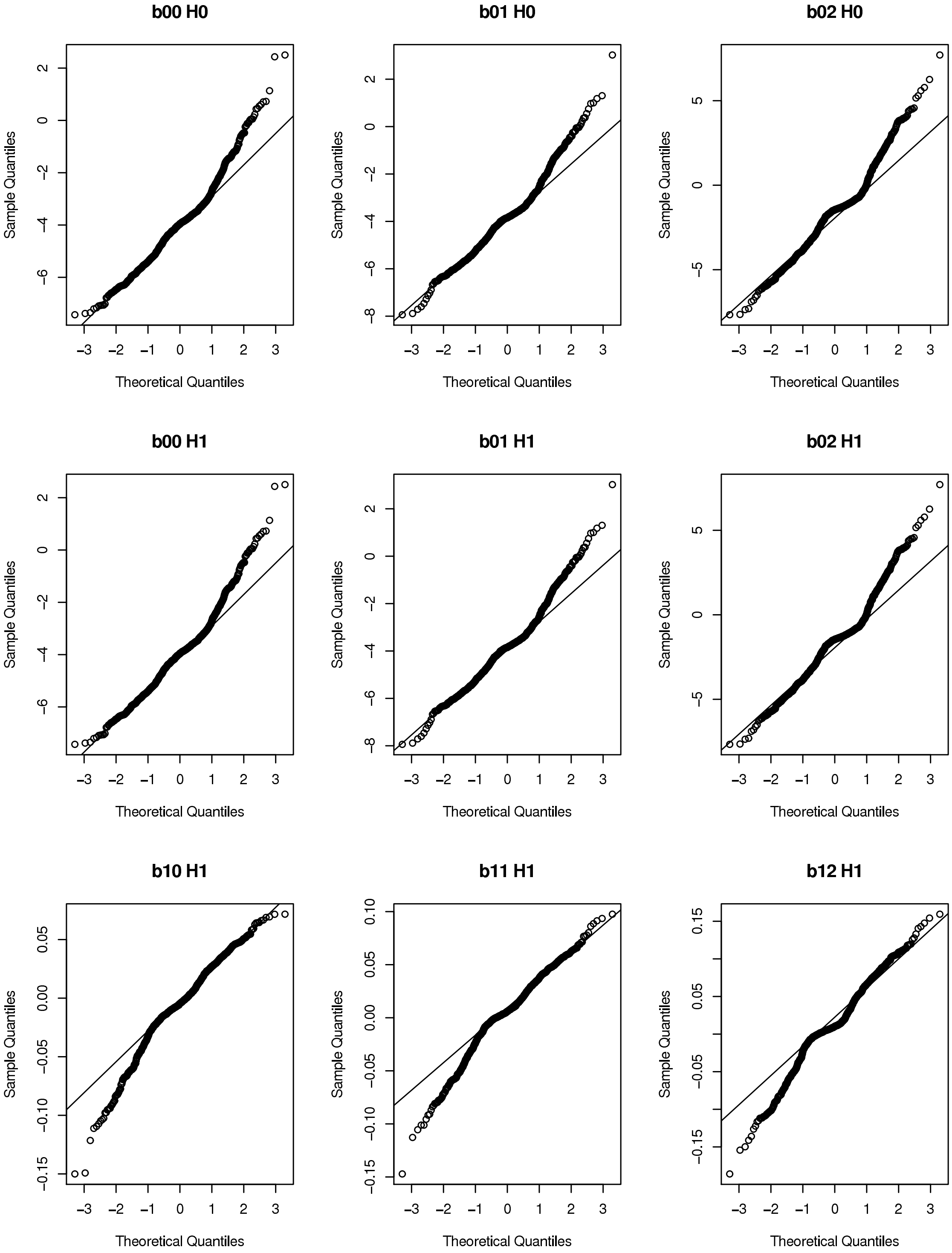


Figure H.5: Q-Q plots using the $N(0,1)$ distribution for the estimates of intercepts under the null and alternative hypothesis when q is not estimated for $q = 0.1$ and $n = 300$.

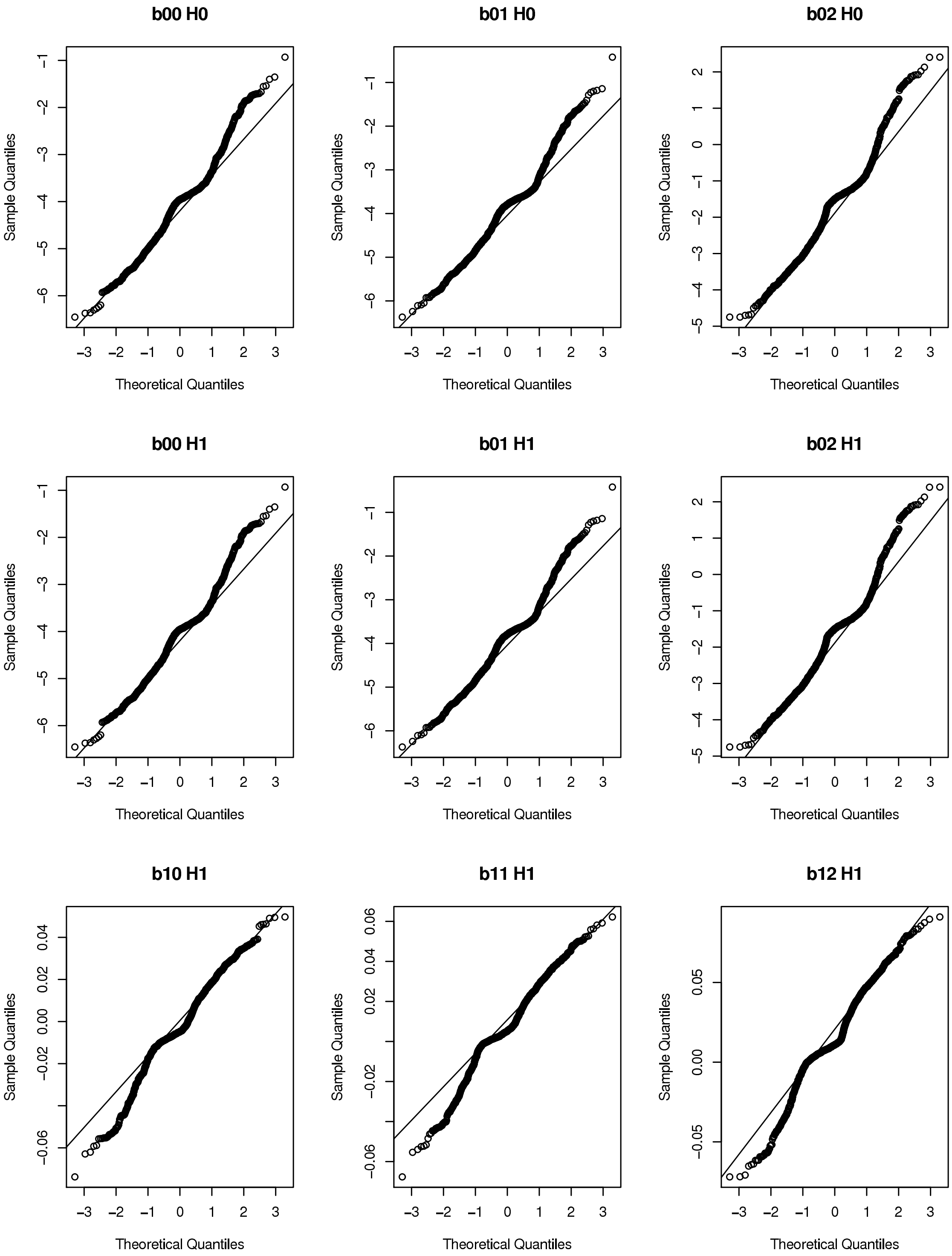


Figure H.6: Q-Q plots using the $N(0, 1)$ distribution for the estimates of intercepts under the null and alternative hypothesis when q is not estimated for $q = 0.1$ and $n = 1000$.

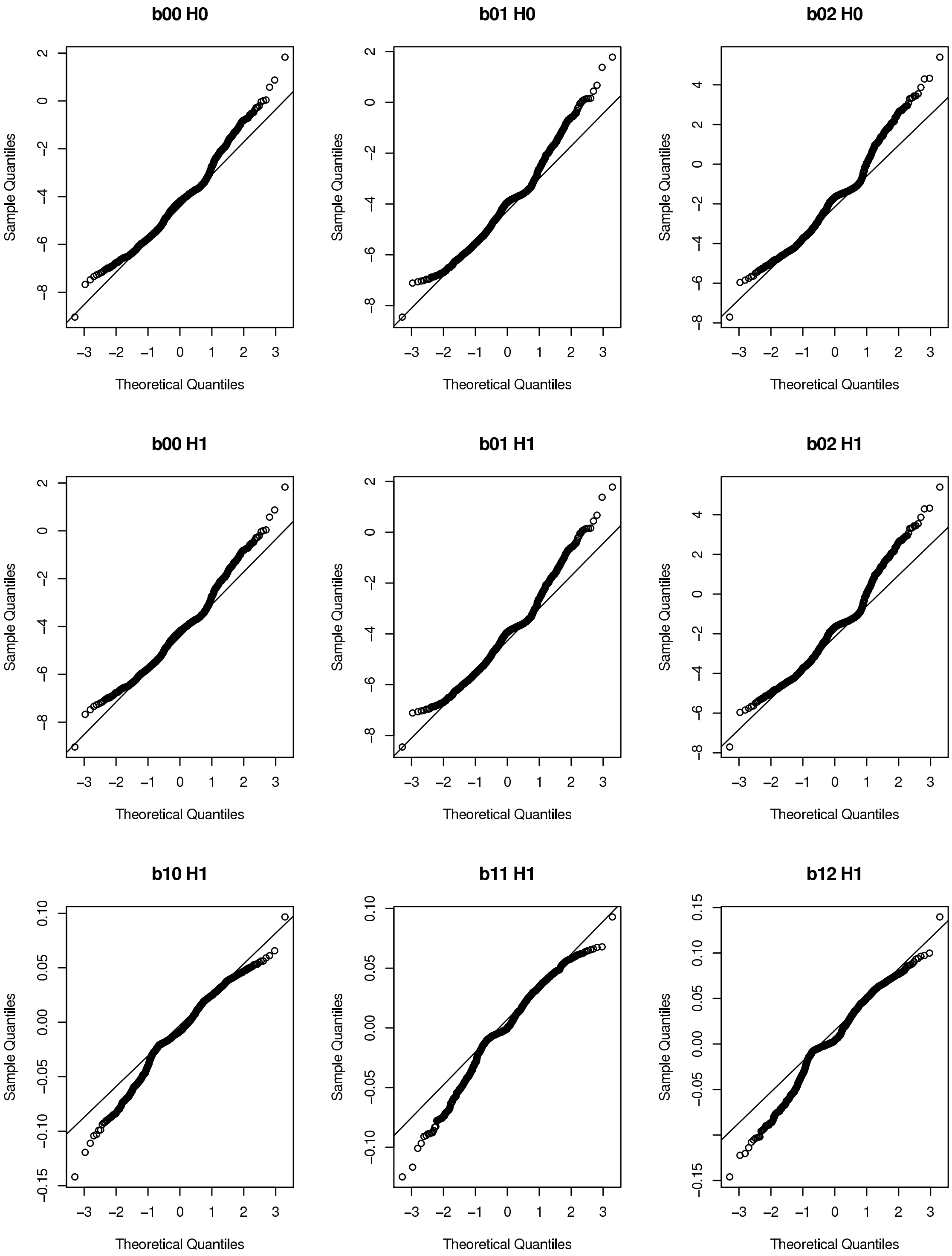
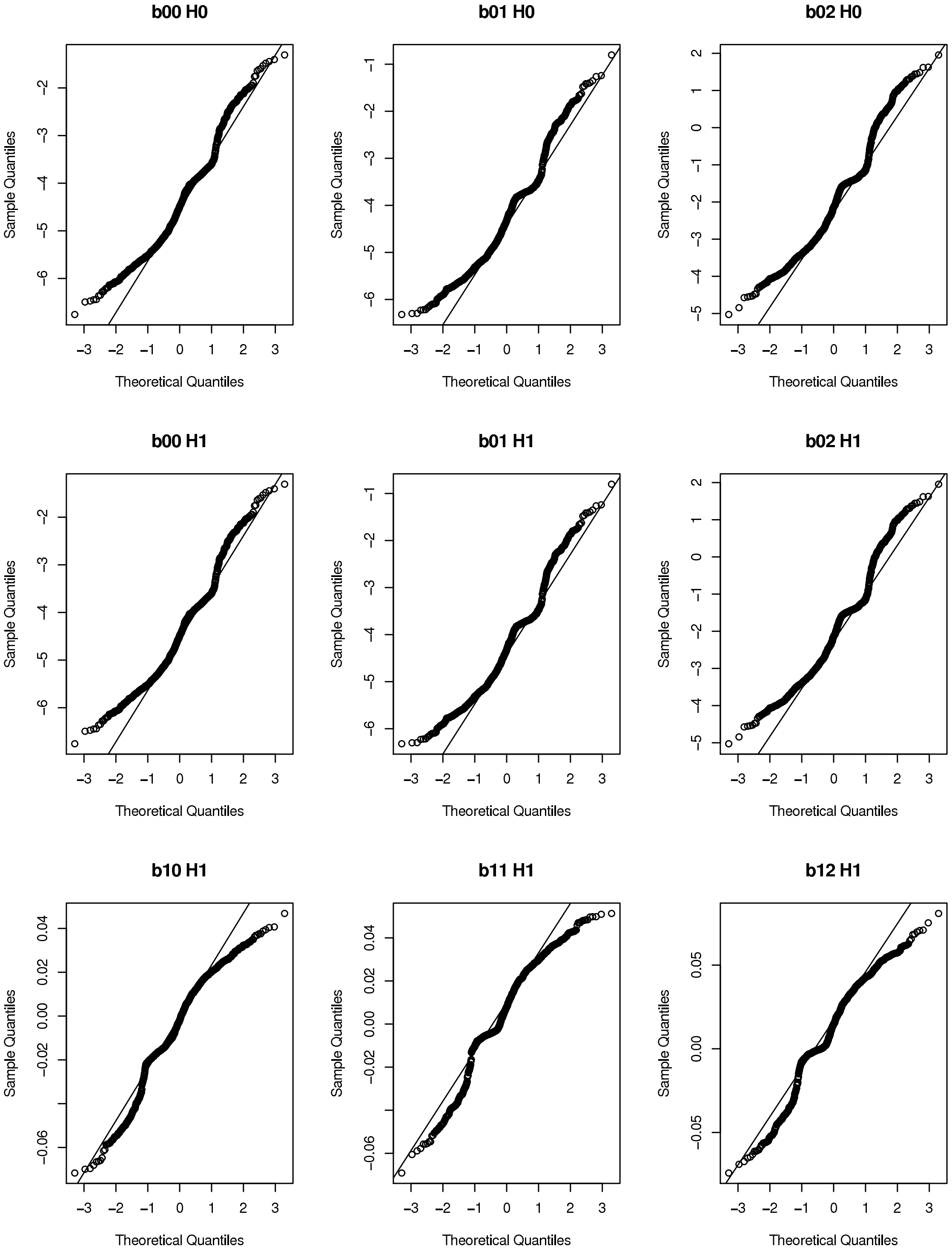


Figure H.7: Q-Q plots using the $N(0,1)$ distribution for the estimates of intercepts under the null and alternative hypothesis when q is not estimated for $q = 0.2$ and $n = 300$.


b00 H0

b01 H0

b02 H0

b00 H1

b01 H1

b02 H1

b10 H1

b11 H1

b12 H1

Figure H.8: Q-Q plots using the $N(0,1)$ distribution for the estimates of intercepts under the null and alternative hypothesis when q is not estimated for $q = 0.2$ and $n = 1000$.

Appendix I

Copyright License Agreement

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS

Apr 11, 2010

This is a License Agreement between Vaneeta K Grover ("You") and John Wiley and Sons ("John Wiley and Sons") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by John Wiley and Sons, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	2403740984610
License date	Apr 07, 2010
Licensed content publisher	John Wiley and Sons
Licensed content publication	Annals of Human Genetics
Licensed content title	Attributing Hardy-Weinberg Disequilibrium to Population Stratification and Genetic Association in Case-Control Studies
Licensed content author	Grover Vaneeta K., Cole David E. C., Hamilton David C.
Licensed content date	Nov 20, 2009
Start page	77
End page	87
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Order reference number	
Total	0.00 USD
Terms and Conditions	

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one if its group companies (each a "Wiley Company") or a society for whom a Wiley Company has exclusive publishing rights in relation to a particular journal (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your Rightslink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

1. The materials you have requested permission to reproduce (the "Materials") are protected by copyright.

2. You are hereby granted a personal, non-exclusive, non-sublicensable, non-transferable, worldwide, limited license to reproduce the Materials for the purpose specified in the licensing process. This license is for a one-time use only with a maximum distribution equal to the number that you identified in the licensing process. Any form of republication granted by this licence must be completed within two years of the date of the grant of this licence (although copies prepared before may be distributed thereafter). Any electronic posting of the Materials is limited to one year from the date permission is granted and is on the condition that a link is placed to the journal homepage on Wiley's online journals publication platform at www.interscience.wiley.com. The Materials shall not be used in any other manner or for any other purpose. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher and on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Material. Any third party material is expressly excluded from this permission.

3. With respect to the Materials, all rights are reserved. No part of the Materials may be copied, modified, adapted, translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Materials without the prior permission of the respective copyright owner. You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Materials, or any of the rights granted to you hereunder to any other person.

4. The Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc or one of its related companies (WILEY) or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto.

5. WILEY DOES NOT MAKE ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND WAIVED BY YOU.

6. WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

7. You shall indemnify, defend and hold harmless WILEY, its directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

8. IN NO EVENT SHALL WILEY BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

9. Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

10. The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

11. This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

12. These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in a writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

13. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

14. WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

15. This Agreement shall be governed by and construed in accordance with the laws of England and you agree to submit to the exclusive jurisdiction of the English courts.

BY CLICKING ON THE "I ACCEPT" BUTTON, YOU ACKNOWLEDGE THAT YOU HAVE READ AND FULLY UNDERSTAND EACH OF THE SECTIONS OF AND PROVISIONS SET FORTH IN THIS AGREEMENT AND THAT YOU ARE IN AGREEMENT WITH AND ARE WILLING TO ACCEPT ALL OF YOUR OBLIGATIONS AS SET FORTH IN THIS AGREEMENT.

V1.2

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK10764214. Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

If you find copyrighted material related to this license will not be used and wish to cancel, please contact us referencing this license number 2403740984610 and noting the reason for cancellation.

Questions? customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.