

## Sample Size Needed to Detect Gene-Gene Interactions using Association Designs

Shuang Wang and Hongyu Zhao

From the Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT.

Received for publication August 2, 2002; accepted for publication May 8, 2003.

It is likely that many complex diseases result from interactions among several genes, as well as environmental factors. The presence of such interactions poses challenges to investigators in identifying susceptibility genes, understanding biologic pathways, and predicting and controlling disease risks. Recently, Gauderman (*Am J Epidemiol* 2002;155:478–84) reported results from the first systematic analysis of the statistical power needed to detect gene-gene interactions in association studies. However, Gauderman used different statistical models to model disease risks for different study designs, and he assumed a very low disease prevalence to make different models more comparable. In this article, assuming a logistic model for disease risk for different study designs, the authors investigate the power of population-based and family-based association designs to detect gene-gene interactions for common diseases. The results indicate that population-based designs are more powerful than family-based designs for detecting gene-gene interactions when disease prevalence in the study population is moderate.

genetic predisposition to disease; genetics; interaction; research design; sample size

The term “epistasis” was first introduced by William Bateson (1) almost a century ago. He described a “masking” effect in which one gene interferes with the phenotypic gene such that the phenotype is determined by the former, which prevents the phenotypic gene from manifesting its effect. However, a broader definition of epistasis has been widely accepted. Under this broader definition, epistasis is defined as a situation in which differential phenotypic expressions of a genotype at one locus depend on the genotype at another locus. Many studies have already demonstrated both the scientific and the public health importance of gene-gene interaction (2–5).

Gene-gene interaction can be studied in both linkage studies and association studies. Linkage studies include model-based methods in which a detailed model for the disease mode of inheritance is specified and model-free methods in which no details such as allele frequencies and modes of inheritance are specified. For example, Cordell et al. (6, 7) investigated multilocus linkage tests of joint genetic effects using affected relative pairs and statistical modeling of interlocus interactions. Mitchell et al. (8) studied epistasis using variance component linkage analysis. Because association studies are likely to be more powerful than linkage studies for identifying genes with small-to-moderate effects

in humans, in this article we focus our discussion on association designs.

Association designs can be broadly categorized into family-based designs and population-based designs. The family-based study design (9, 10) has received great attention in the last decade, both because of its robustness to population stratification and because of its power to identify genes with small-to-moderate effects (11). This design compares alleles transmitted to the affected children with those not transmitted. The population-based design has been criticized for possibly inducing spurious association due to population stratification, but it may be easier and less expensive to collect DNA samples from unrelated persons in the general population for certain diseases, and previous studies have shown that population-based studies such as traditional case-control studies can be more powerful than family-based studies in identifying disease genes, both for qualitative traits (12, 13) and for quantitative traits (14). Moreover, genomic markers can be used to control for population stratification in population-based association studies (15–18).

In a recent paper, Gauderman (19) discussed sample size requirements for detecting gene-gene interaction using four different study designs: the matched-case-control design, the case-sibling design, the case-parent design, and the case-only design. He used different statistical models for different

designs, and he assumed a very low disease prevalence rate in order to make the parameter estimates have comparable meanings. In this article, using the same logistic regression model for disease risks across different study designs, we calculate the sample sizes needed to detect gene-gene interaction with the case-parent design, the matched case-control design, and the unmatched case-control design. We make comparisons for different levels of gene-gene interaction under three genetic models: the additive model, the dominant model, and the recessive model. We find that the unmatched case-control design is more powerful than both the matched case-control design and the case-parent design, whereas the matched case-control design is more powerful than the case-parent design when the disease prevalence is moderate (10 percent) and less powerful when the disease prevalence is low (1 percent).

## METHODS

### Case-parent design

With the case-parent study design, affected persons and their parents are sampled. Throughout this article, we assume that the disease status of the offspring depends only on his/her genotype, that is,  $p(D|G_o, G_p) = p(D|G_o)$ , where  $G_o$  and  $G_p$  denote the genotype of the diseased child and the genotypes of the parents and  $D$  denotes that the child has the disease. The value of  $p(D|G_o)$  is known as penetrance. In epidemiologic studies, rather than direct modeling of the penetrance, a more common measure is the natural logarithm of the odds,  $\log[p/(1-p)]$ , which does not have the constraint of falling into the interval 0–1. Therefore, we use the following logistic model for gene-gene interactions:

$$\log[p(D|G_o)/(1-p(D|G_o))] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2,$$

where  $X_1$  and  $X_2$  are the codings for the genotypes at two candidate genes and the codings depend on the specific genetic model being studied. We assume that we will study each candidate gene at a polymorphic site with two allelic variants, a high-risk allele (denoted by capital letters,  $A$  and  $B$ ) and a low-risk allele (denoted by small letters,  $a$  and  $b$ ). Therefore, for each individual, there are nine possible genotype combinations at these two marker loci. We consider three genetic models—the additive model, the dominant model, and the recessive model—by coding genotypes differently as described in table 1. For example, under the additive model,  $X_1$  and  $X_2$  take the value 2 for the genotypes  $AA$  and  $BB$ , 1 for the genotypes  $Aa$  and  $Bb$ , and 0 for the genotypes  $aa$  and  $bb$ . We use the same coding system for both family-based designs and population-based designs. Note that Schaid (20) has proposed the logistic regression model for assessment of gene-environment interaction, whereas Gauderman (19) investigated gene-gene interaction using the conditional logistic regression model. In our model, the parameters to be estimated are  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$ , where  $\beta_0$  corresponds to the intercept,  $\beta_1$  and  $\beta_2$  correspond to the main effects at two candidate genes (denoted by  $X_1$  and  $X_2$ ), and  $\beta_3$  corresponds to the interaction effect (denoted by  $X_1 X_2$ , which is the product of  $X_1$  and  $X_2$ ). When there is gene-

**TABLE 1. Coding genotypes for the additive model, the dominant model, and the recessive model**

Offspring genotype at two loci ( $X_1, X_2$ )	Model ( $X_1, X_2, X_1 X_2$ )		
	Additive	Dominant	Recessive
$AA, BB$	2, 2, 4	1, 1, 1	1, 1, 1
$AA, Bb$	2, 1, 2	1, 1, 1	1, 0, 0
$AA, bb$	2, 0, 0	1, 0, 0	1, 0, 0
$Aa, BB$	1, 2, 2	1, 1, 1	0, 1, 0
$Aa, Bb$	1, 1, 1	1, 1, 1	0, 0, 0
$Aa, bb$	1, 0, 0	1, 0, 0	0, 0, 0
$aa, BB$	0, 2, 0	0, 1, 0	0, 1, 0
$aa, Bb$	0, 1, 0	0, 1, 0	0, 0, 0
$aa, bb$	0, 0, 0	0, 0, 0	0, 0, 0

gene interaction, that is, when  $\beta_3$  is not equal to 0, the effect of one gene varies over the levels of the other gene.

We assume that the frequencies of the two alleles at the first candidate gene,  $A$  and  $a$ , are  $p_A$  and  $p_a$  and the frequencies of the two alleles at the second candidate gene,  $B$  and  $b$ , are  $p_B$  and  $p_b$ . The genotype frequencies of nine possible genotypes for each individual can be calculated under the assumptions of Hardy-Weinberg equilibrium and linkage equilibrium between these two genes, as summarized in table 2. For parental mating types at these two genes, there are 81 possible combinations, if we distinguish two parents. We assume that parents are unrelated and the matings are random, so the probability of having a certain mating type is simply the product of the two parental genotype frequencies.

For the case-parent design, we consider two ways to form the likelihood for the observed genotypes of the affected children and their parents by either conditioning on the parental mating types or not conditioning on the parental mating types. In the conditional likelihood formulation, let  $p_{G_o|G_p, D}$  denote the probability that an affected offspring has genotype  $G_o$ , conditional on his/her parental mating type  $G_p$  and the fact that he/she is affected. The conditional likelihood for a set of independent family trios is

$$\prod_{G_{o_i}, G_{p_j}}^{a_{G_{o_i}, G_{p_j}, D}} p_{G_{o_i}|G_{p_j}, D}, i = 1, \dots, 9, j = 1, \dots, 81,$$

with the log-likelihood being

$$\ln L_c = \sum_{j=1}^{81} \sum_{i=1}^9 a_{G_{o_i}, G_{p_j}, D} \log(p_{G_{o_i}|G_{p_j}, D}),$$

**TABLE 2. Genotype frequencies for persons with different genotype combinations at two loci**

	$BB, p(G)$	$Bb, p(G)$	$bb, p(G)$
$AA$	$AABB, p_A^2 p_B^2$	$AABb, 2p_A^2 p_B p_b$	$AAbb, p_A^2 p_b^2$
$Aa$	$AaBB, 2p_A p_a p_B^2$	$AaBb, 4p_A p_a p_B p_b$	$Aabb, 2p_A p_a p_b^2$
$aa$	$aaBB, p_a^2 p_B^2$	$aaBb, 2p_a^2 p_B p_b$	$aabb, p_a^2 p_b^2$

where  $a_{G_o, G_p, D}$  is the number of families whose diseased child has genotype  $G_o$  and whose parents have genotypes  $G_p$ . The  $p_{G_o|G_p, D}$  can be calculated through the Bayes rule as

$$\begin{aligned} p_{G_o|G_p, D} &= \frac{p(G_o, G_p, D)}{p(G_p, D)} = \frac{p(D|G_o, G_p)p(G_o, G_p)}{\sum_{G_o} p(G_o, G_p, D)} \\ &= \frac{p(D|G_o)p(G_o|G_p)p(G_p)}{\sum_{G_o} p(D|G_o)p(G_o|G_p)p(G_p)} \\ &= \frac{p(D|G_o)p(G_o|G_p)}{\sum_{G_o} p(D|G_o)p(G_o|G_p)}, \end{aligned}$$

where  $G_o$ ,  $G_p$ , and  $D$  are defined as above. The conditional analysis is robust to population stratification in the testing of linkage or association between a candidate gene and disease using family trios. However, although conditional analysis is robust to population stratification for detecting main genetic effects, it is no longer robust for the detection of gene-gene interactions (more details are provided in the Discussion section). Therefore, we also consider unconditional analysis below as an alternative approach to studying gene-gene interactions.

For the unconditional formulation, let  $p_{G_o, G_p|D}$  be the probability that an affected offspring has genotype  $G_o$  and his/her parents have mating type  $G_p$ , conditional on the child's being affected. The unconditional likelihood for a set of independent family trios is

$$\prod_{G_o, G_p} p_{G_o, G_p|D}^{a_{G_o, G_p, D}}, i = 1, \dots, 9, j = 1, \dots, 81$$

with the log-likelihood being

$$\ln L_u = \sum_{j=1}^{81} \sum_{i=1}^9 a_{G_o, G_p, D} \log(p_{G_o, G_p|D}),$$

where  $a_{G_o, G_p, D}$  is the number of families in which the diseased child has genotype  $G_o$  and the parents have genotypes  $G_p$ . The  $p_{G_o, G_p|D}$  can be evaluated as

$$\begin{aligned} p_{G_o, G_p|D} &= \frac{p(G_o, G_p, D)}{p(D)} = \frac{p(D|G_o, G_p)p(G_o, G_p)}{\sum_{G_o} \sum_{G_p} p(G_o, G_p, D)} \\ &= \frac{p(D|G_o)p(G_o|G_p)p(G_p)}{\sum_{G_o} \sum_{G_p} p(D|G_o)p(G_o|G_p)p(G_p)} \\ &= \frac{p(G_p)p(D|G_o)p(G_o|G_p)}{\sum_{G_p} p(G_p) \sum_{G_o} p(D|G_o)p(G_o|G_p)}. \end{aligned}$$

To determine the sample sizes required to detect gene-gene interactions, we use the noncentral chi-squared distribution to approximate the distribution of the likelihood ratio statistics. To derive the noncentrality parameter, we need to maximize the expected log-likelihood for a set of independent families with the expected number of each type of family

$$\ln L_c^E = \sum_{j=1}^{81} \sum_{i=1}^9 a_{G_o, G_p, D}^* \log p_{G_o, G_p|D}$$

and

$$\ln L_u^E = \sum_{j=1}^{81} \sum_{i=1}^9 a_{G_o, G_p, D}^* \log p_{G_o, G_p|D},$$

where  $a_{G_o, G_p, D}^*$  is the expected number of families whose diseased child has genotype  $G_o$  and whose parents have genotypes  $G_p$ . It is easy to see that

$$\begin{aligned} a_{G_o, G_p, D}^* &= N \times p(G_o, G_p|D) = N \times \frac{p(G_o, G_p, D)}{p(D)} \\ &= N \times \frac{p(D|G_o, G_p)p(G_o, G_p)}{\sum_{j=1}^{81} \sum_{i=1}^9 p(D, G_o, G_p)} \\ &= N \times \frac{p(D|G_o)p(G_o|G_p)p(G_p)}{\sum_{j=1}^{81} \sum_{i=1}^9 p(D|G_o)p(G_o|G_p)p(G_p)} \\ &= N \times \frac{p(G_p)p(D|G_o)p(G_o|G_p)}{\sum_{j=1}^{81} p(G_p) \sum_{i=1}^9 p(D|G_o)p(G_o|G_p)}, \end{aligned}$$

where  $N$  is the number of trios. The total number of persons in the sample is  $3N$ . For sample size calculation, we consider various genetic models with different interaction effects, and the null hypothesis assumes a genetic model with no gene-gene interactions. The likelihood ratio statistic has an approximate noncentral chi-squared distribution with 1 degree of freedom and noncentrality parameter  $\lambda = N\delta^2 = 2(\ln \hat{L}^{E1} - \ln \hat{L}^{E0})$ , which is the expected log-likelihood ratio test statistic,  $\ln \hat{L}^{E1}$  is the expected log-likelihood allowing the presence of interaction, and  $\ln \hat{L}^{E0}$  is that without interaction. We maximize the likelihood by means of the simplex method (21). For a prespecified power—for example,  $b = 80$  percent—and a prespecified significance level—for example,  $a = 5$  percent—the sample size  $N$  can be calculated as  $(z_{a/2} + z_{1-b})/\delta^2$ , where  $z_a$  is the  $(1 - a)$ th percentile of the standard normal distribution.

In the following discussion, we call the conditional analysis using the case-parent design the “conditional case-

**TABLE 3. Sample size needed to detect gene-gene interaction for the family-based study design and the population-based study design with  $R_1 = 1.0$ ,  $R_2 = 1.0$ , and the population disease prevalence fixed at 10%**

Susceptible proportion (gene 1, gene 2) and model	Family-based design: $3N_{c-p}$						Population-based design: $2N_{c-c}$					
	Conditional on parental mating type: $p(G_o D, G_p)$			Unconditional on parental mating type: $p(G_o, G_p D)$			Matched case-control design			Unmatched case-control design		
	$R_{inter} = 2$	$R_{inter} = 4$	$R_{inter} = 6$	$R_{inter} = 2$	$R_{inter} = 4$	$R_{inter} = 6$	$R_{inter} = 2$	$R_{inter} = 4$	$R_{inter} = 6$	$R_{inter} = 2$	$R_{inter} = 4$	$R_{inter} = 6$
(0.1, 0.1)												
Additive	13,863	4,053	2,727	9,933	2,811	1,863	5,554	1,294	790	3,042	646	380
Dominant	15,708	4,347	2,922	11,424	3,054	2,019	6,690	1,504	884	3,742	770	434
Recessive	6,924	1,488	852	5,817	1,212	684	6,690	1,504	894	3,742	770	434
(0.1, 0.2)												
Additive	5,289	1,440	951	3,747	975	633	2,924	720	452	1,576	354	214
Dominant	6,594	1,677	1,047	4,797	1,170	714	3,882	902	538	2,172	460	262
Recessive	4,029	885	516	3,333	711	408	3,882	902	538	2,172	460	262
(0.1, 0.25)												
Additive	3,960	1,062	702	2,793	714	465	2,406	606	386	1,290	294	180
Dominant	5,208	1,290	810	3,795	897	549	3,360	794	476	1,872	404	230
Recessive	3,486	777	456	2,871	621	357	3,360	794	478	1,872	404	230
(0.2, 0.2)												
Additive	2,379	648	420	1,674	435	276	1,552	404	262	830	198	124
Dominant	3,291	810	495	2,406	564	336	2,250	540	328	1,266	278	162
Recessive	2,364	537	318	1,929	426	249	2,250	540	328	1,266	278	162
(0.2, 0.25)												
Additive	1,872	507	333	1,314	339	219	1,282	342	224	686	166	106
Dominant	2,700	666	405	1,980	468	276	1,948	474	290	1,096	246	144
Recessive	2,067	477	285	1,680	378	222	1,948	474	290	1,096	246	144
(0.25, 0.25)												
Additive	1,482	405	267	1,038	273	174	1,060	292	194	564	142	90
Dominant	2,265	555	339	1,668	390	234	1,688	418	258	956	218	128
Recessive	1,800	420	252	1,455	333	198	1,688	418	258	956	218	128

parent design” and the unconditional analysis the “unconditional case-parent design.”

### Case-control design

In the case-control study design, we consider both matched and unmatched case-control designs. For the unmatched case-control design, we assume that we sample  $N_D$  cases and  $N_{\bar{D}}$  controls, with  $N_{\bar{D}} = RN_D$ , where  $R$  can be any prespecified positive number. The total sample size is  $N = N_D + N_{\bar{D}} = (1 + R)N_D$ . Let  $P_{G_i|D}$  be the probability that the  $i$ th diseased individual has genotype  $G_i$  and  $p_{G_j|\bar{D}}$  be the probability that the  $j$ th normal individual has genotype  $G_j$ . The likelihood for the case-control data is

$$L = \prod_{i=1}^{N_D} P_{G_i|D} \prod_{j=1}^{N_{\bar{D}}} p_{G_j|\bar{D}},$$

where

$$P_{G_i|D} = \frac{p(G_i, D)}{p(D)} = \frac{p(D|G_i)p(G_i)}{\sum_{i=1}^9 p(D|G_i)p(G_i)},$$

and

$$p_{G_j|\bar{D}} = \frac{p(G_j, \bar{D})}{p(\bar{D})} = \frac{p(\bar{D}|G_j)p(G_j)}{\sum_{j=1}^9 p(\bar{D}|G_j)p(G_j)},$$

where  $p(G_i)$  is the genotype frequency summarized in table 2.

To determine the sample size, we need to maximize the expected log-likelihood for a sample with the expected numbers of cases and controls:

$$\ln L^E = \sum_{i=1}^9 (a_{iD}^* \log P_{G_i|D}) + \sum_{j=1}^9 (a_{j\bar{D}}^* \log p_{G_j|\bar{D}}),$$

**TABLE 4.** Sample size needed to detect gene-gene interaction for the family-based study design and the population-based study design with  $R_1 = 1.0$ ,  $R_2 = 1.0$ , and the population disease prevalence fixed at 1%

Susceptible proportion (gene 1, gene 2) and model	Family-based design: $3N_{c-p}$						Population-based design: $2N_{c-c}$					
	Conditional on parental mating type: $p(G_o D, G_p)$			Unconditional on parental mating type: $p(G_o, G_p D)$			Matched case-control design			Unmatched case-control design		
	$R_{inter} = 2$	$R_{inter} = 4$	$R_{inter} = 6$	$R_{inter} = 2$	$R_{inter} = 4$	$R_{inter} = 6$	$R_{inter} = 2$	$R_{inter} = 4$	$R_{inter} = 6$	$R_{inter} = 2$	$R_{inter} = 4$	$R_{inter} = 6$
(0.1, 0.1)												
Additive	4,986	945	516	3,444	588	303	5,030	1,004	562	2,108	324	158
Dominant	6,267	1,221	654	4,455	801	408	6,284	1,292	712	2,786	484	240
Recessive	5,010	909	465	4,188	729	363	6,284	1,292	712	2,786	484	240
(0.1, 0.2)												
Additive	2,466	474	267	1,686	288	153	2,636	560	330	1,068	168	86
Dominant	3,537	714	390	2,532	474	246	3,682	800	458	1,630	300	156
Recessive	2,970	570	303	2,448	450	234	3,682	800	458	1,630	300	156
(0.1, 0.25)												
Additive	1,980	384	219	1,350	231	123	2,164	474	286	860	138	74
Dominant	3,036	624	345	2,181	417	219	3,204	714	414	1,418	268	142
Recessive	2,595	510	276	2,127	402	213	3,204	714	414	1,418	268	142
(0.2, 0.2)												
Additive	1,257	261	159	852	156	90	1,394	324	208	548	94	54
Dominant	2,046	438	252	1,479	297	162	2,156	494	294	966	192	106
Recessive	1,782	366	207	1,449	288	159	2,156	494	294	966	192	106
(0.2, 0.25)												
Additive	1,020	219	138	690	129	78	1,150	280	184	446	80	48
Dominant	1,773	390	231	1,287	267	150	1,876	442	266	844	174	98
Recessive	1,566	333	192	1,269	261	147	1,876	442	266	844	174	98
(0.25, 0.25)												
Additive	831	186	123	561	111	69	950	242	168	364	68	44
Dominant	1,542	351	213	1,125	243	141	1,634	394	242	742	160	92
Recessive	1,380	303	180	1,113	240	138	1,634	394	242	742	160	92

where  $a_{iD}^*$  is the expected number of cases with genotype  $G_i$  and  $a_{j\bar{D}}^*$  is the expected number of controls with genotype  $G_j$ . It is easy to see that

$$a_{iD}^* = N_D \times p(G_i|D) = N_D \times \frac{p(D|G_i)p(G_i)}{\sum_{i=1}^9 p(D|G_i)p(G_i)},$$

$$a_{j\bar{D}}^* = N_{\bar{D}} \times p(G_j|\bar{D}) = R \times N_D \times \frac{p(\bar{D}|G_j)p(G_j)}{\sum_{j=1}^9 p(\bar{D}|G_j)p(G_j)}.$$

The likelihood ratio statistic has an approximate noncentral chi-squared distribution with 1 degree of freedom and noncentrality parameter  $\lambda = N_D \delta^2 = 2(\ln L^{E1} - \ln L^{E0})$ , where  $\ln L^{E1}$  is the expected log-likelihood under a model that allows interactions and  $\ln L^{E0}$  is that without interactions. The number of required samples is  $(1 + R)N_D$ . In this article, we assume an equal number of cases and controls, that is,  $R = 1$ .

For the matched case-control design, we consider the 1:1 matching situation. Let  $p_{G_{iD}, G_{j\bar{D}}}$  denote the probability that the diseased individual has genotype  $G_i$  and the normal individual has genotype  $G_j$  in a matched case-control pair. The conditional likelihood for  $N$  sets of independent matched case-control pairs is

$$L = \prod_{k=1}^N p_{k(G_{iD}, G_{j\bar{D}})} = \prod_{G_{iD}} \prod_{G_{j\bar{D}}} p_{G_{iD}, G_{j\bar{D}}}^{a_{G_{iD}, G_{j\bar{D}}}}, i = 1, \dots, 9, j = 1, \dots, 9,$$

where

$$p_{G_{iD}, G_{j\bar{D}}} = p(G_{iD}, G_{j\bar{D}}|G_{iD}, G_{j\bar{D}} \text{ or } G_{jD}, G_{i\bar{D}})$$

$$= \frac{p(D|G_i)p(\bar{D}|G_j)}{p(D|G_i)p(\bar{D}|G_j) + p(D|G_j)p(\bar{D}|G_i)}$$

and  $a_{G_{iD}, G_{j\bar{D}}}$  is the number of pairs in which the case has genotype  $G_i$  and the control has genotype  $G_j$ .

**TABLE 5. Sample size needed to detect gene-gene interaction for the family-based study design and the population-based study design with  $R_1 = 3.0$ ,  $R_2 = 3.0$ , and the population disease prevalence fixed at 10%**

Susceptible proportion (gene 1, gene 2) and model	Family-based design: $3N_{c,p}$						Population-based design: $2N_{c,c}$					
	Conditional on parental mating type: $p(G_o D, G_p)$			Unconditional on parental mating type: $p(G_o, G_p D)$			Matched case-control design			Unmatched case-control design		
	$R_{inter} = 2$	$R_{inter} = 4$	$R_{inter} = 6$	$R_{inter} = 2$	$R_{inter} = 4$	$R_{inter} = 6$	$R_{inter} = 2$	$R_{inter} = 4$	$R_{inter} = 6$	$R_{inter} = 2$	$R_{inter} = 4$	$R_{inter} = 6$
(0.1, 0.1)												
Additive	32,304	11,877	8,583	19,131	7,065	5,013	8,022	2,436	1,686	2,936	870	612
Dominant	33,903	11,937	9,102	20,241	6,984	5,370	8,282	2,474	1,698	2,978	890	622
Recessive	7,026	2,316	1,605	5,043	1,644	1,116	8,282	2,474	1,698	2,978	990	622
(0.1, 0.2)												
Additive	8,238	2,739	2,031	4,824	1,560	1,155	4,276	1,312	898	1,572	468	318
Dominant	8,466	2,811	2,100	4,947	1,590	1,176	4,620	1,362	922	1,720	496	334
Recessive	3,735	1,143	819	2,721	804	573	4,620	1,362	922	1,720	496	334
(0.1, 0.25)												
Additive	5,715	1,809	1,320	3,300	1,017	750	3,564	1,094	750	1,312	386	262
Dominant	6,021	1,959	1,395	3,549	1,116	783	3,950	1,156	782	1,484	418	282
Recessive	3,276	990	663	2,370	699	462	3,950	1,156	782	1,484	418	282
(0.2, 0.2)												
Additive	3,294	999	693	1,893	555	381	2,340	726	496	852	252	166
Dominant	3,687	1,095	765	2,154	621	426	2,680	792	536	1,012	288	190
Recessive	2,268	648	441	1,644	459	309	2,690	792	536	1,012	288	190
(0.2, 0.25)												
Additive	2,535	768	537	1,458	429	294	1,980	620	426	728	212	140
Dominant	2,943	870	594	1,764	501	333	2,348	696	470	922	254	166
Recessive	1,977	564	387	1,437	402	273	2,348	696	470	922	254	166
(0.25, 0.25)												
Additive	2,037	609	414	1,176	339	225	1,678	534	368	618	182	118
Dominant	2,487	720	486	1,506	420	279	2,076	622	424	832	234	154
Recessive	1,803	507	339	1,323	366	243	2,076	622	424	832	234	154

To determine the sample size, we need to maximize the expected log-likelihood for a sample with the expected number of pairs with a specific case genotype and control genotype combination:

$$\ln L^E = \sum_{i=1}^9 \sum_{j=1}^9 a_{G_{iD}, G_{j\bar{D}}}^* \log p_{G_{iD}, G_{j\bar{D}}},$$

where  $a_{G_{iD}, G_{j\bar{D}}}^*$  is the expected number of matched pairs with case genotype  $G_i$  and control genotype  $G_j$ . It is easy to see that

$$\begin{aligned} a_{G_{iD}, G_{j\bar{D}}}^* &= N \times p(G_i|D)p(G_j|\bar{D}) \\ &= N \times \frac{p(D|G_i)p(\bar{D}|G_j)p(G_i)p(G_j)}{\sum_{i=1}^9 p(D|G_i)p(G_i) \sum_{j=1}^9 p(\bar{D}|G_j)p(G_j)}, \end{aligned}$$

where  $p(G_i)$  is the genotype frequency summarized in table 2. The likelihood ratio statistic has an approximate noncentral chi-squared distribution with 1 degree of freedom and noncentral parameter  $\lambda = N\delta^2 = 2(\ln L^{E1} - \ln L^{E0})$ , where  $\ln L^{E1}$  is the maximum expected log-likelihood under a model that allows interactions and  $\ln L^{E0}$  is that without interactions.

The total numbers of subjects required under different gene-gene interaction alternatives are compared by calculating the relative efficiencies, defined as the required sample size under one scenario divided by the smallest required sample size among all possible scenarios (three genetic models: the additive model, the dominant model, and the recessive model; four study designs: the conditional case-parent design, the unconditional case-parent design, the matched case-control design, and the unmatched case-control design; and the number of alternative hypotheses of gene-gene interaction). The relative efficiency of 1 indicates the most efficient design and model setup, and relative efficiencies greater than 1 mean that more samples are needed to achieve the same statistical power as the most efficient method. Similar asymptotic relative efficiency was used

**TABLE 6.** Sample size needed to detect gene-gene interaction for the family-based study design and the population-based study design with  $R_1 = 3.0$ ,  $R_2 = 3.0$ , and the population disease prevalence fixed at 1%

Susceptible proportion (gene 1, gene 2) and model	Family-based design: $3N_{c-p}$						Population-based design: $2N_{c-c}$					
	Conditional on parental mating type: $p(G_o D, G_p)$			Unconditional on parental mating type: $p(G_o, G_p D)$			Matched case-control design			Unmatched case-control design		
	$R_{inter} = 2$	$R_{inter} = 4$	$R_{inter} = 6$	$R_{inter} = 2$	$R_{inter} = 4$	$R_{inter} = 6$	$R_{inter} = 2$	$R_{inter} = 4$	$R_{inter} = 6$	$R_{inter} = 2$	$R_{inter} = 4$	$R_{inter} = 6$
(0.1, 0.1)												
Additive	2,502	654	414	1,122	264	159	4,450	1,158	734	594	142	86
Dominant	3,015	750	462	1,431	321	186	5,054	1,248	772	840	184	106
Recessive	1,887	444	267	1,263	279	162	5,054	1,248	772	840	184	106
(0.1, 0.2)												
Additive	1,422	411	279	648	171	108	2,662	772	526	386	102	68
Dominant	2,070	546	351	1,050	252	153	3,252	866	564	668	160	100
Recessive	1,458	372	237	1,002	243	150	3,252	866	564	668	160	100
(0.1, 0.25)												
Additive	1,227	369	258	570	156	102	2,326	710	500	348	96	64
Dominant	1,950	528	351	1,029	255	162	2,958	814	540	668	166	106
Recessive	1,437	375	246	1,002	249	159	2,958	814	540	668	166	106
(0.2, 0.2)												
Additive	903	300	225	429	129	90	1,672	564	424	270	82	58
Dominant	1,569	471	336	849	237	162	2,176	640	444	558	156	106
Recessive	1,185	342	240	837	234	159	2,176	640	444	558	156	106
(0.2, 0.25)												
Additive	813	288	225	393	126	93	1,496	542	430	252	82	60
Dominant	1,530	483	354	858	252	180	2,012	614	440	568	168	118
Recessive	1,182	360	258	852	252	177	2,012	614	440	568	168	118
(0.25, 0.25)												
Additive	753	288	237	369	129	99	1,352	532	446	240	84	66
Dominant	1,509	501	381	876	273	201	1,872	598	438	580	182	134
Recessive	1,191	381	285	870	273	201	1,872	598	438	580	182	134

both by Schaid (20) in gene-environment interaction studies and by Gauderman (19) in gene-gene interaction studies.

We fix the population prevalence in the parameter estimation. Our experience is that unless population prevalence is fixed throughout the estimation, the estimate of  $\beta_0$  is very unstable; this may lead to an unreasonably high or low population disease prevalence and cause biologic meaning to be lost.

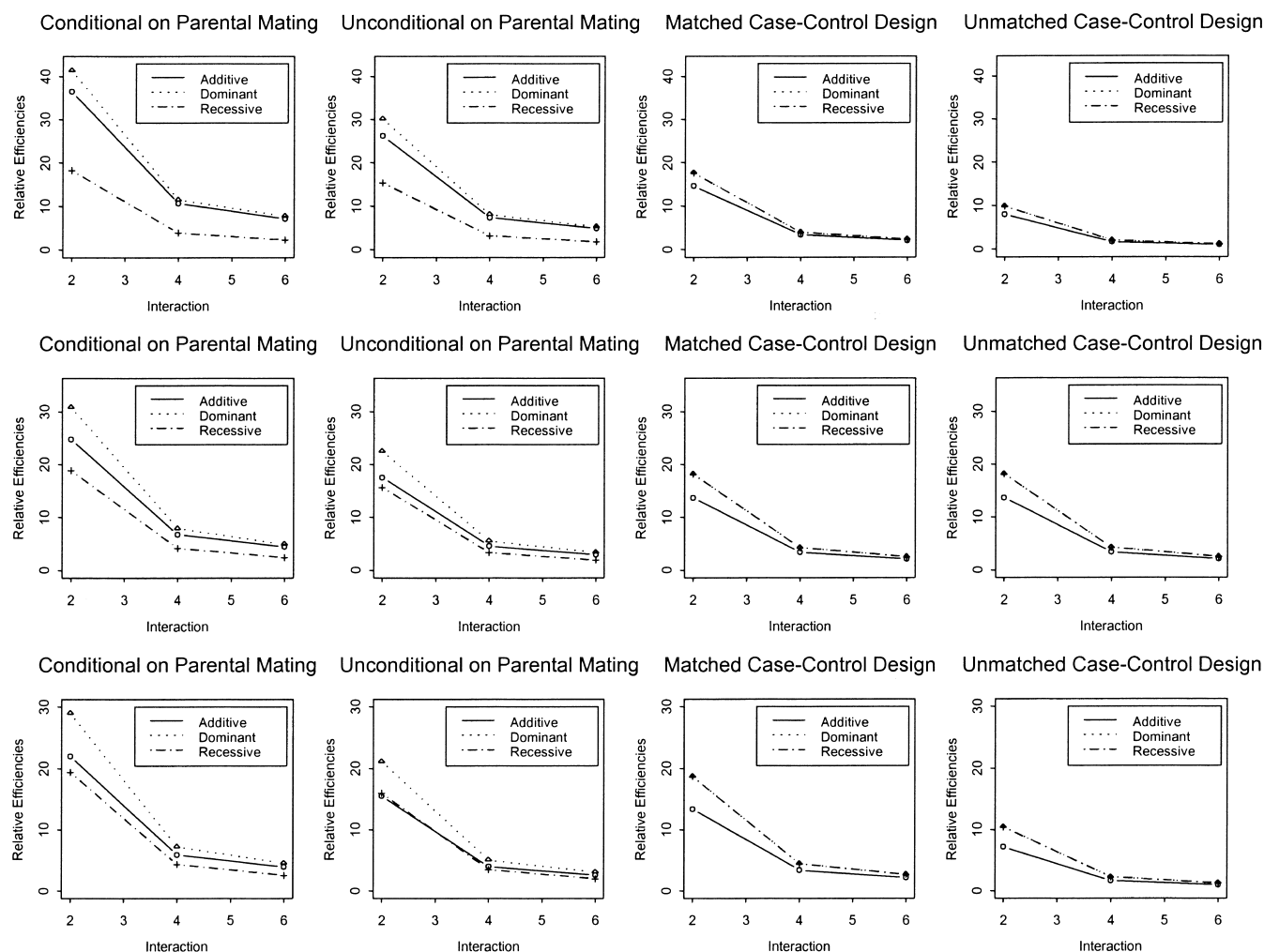
## RESULTS

The sample sizes needed to detect various levels of gene-gene interaction with 80 percent power at the 5 percent significance level are calculated under three genetic models: the additive model, the dominant model, and the recessive model. We consider three levels for the proportion of susceptible persons  $\phi$ , defined as the proportion of persons having the susceptible genotype at a given gene: 0.1, 0.2, and 0.25. From each proportion, we can obtain the underlying high-risk allele frequency for the dominant and additive models

through  $p_A^2 + 2p_A(1 - p_A) = \phi$  and for the recessive model through  $p_A^2 = \phi$ .

We consider both pure interaction models without main effects, that is,  $\beta_1 = 0$  and  $\beta_2 = 0$ , and interaction models with main effects, in which we fix the main effects of the candidate genes at  $R_1 = 3$  and  $R_2 = 3$ . The magnitude of the interaction is varied at  $R_{inter} = 2$ ,  $R_{inter} = 4$ , and  $R_{inter} = 6$ , with  $\beta_3$  equal to  $\log(2)$ ,  $\log(4)$ , and  $\log(6)$ . The population prevalence of the disease is varied at 10 percent and 1 percent, which correspond to a common disease and a disease with relatively low prevalence. The sample sizes for different models and designs are summarized in tables 3, 4, 5, and 6. Each table lists the total number of subjects required. For the family-based design, this number is  $3N_{c-p}$ , where  $N_{c-p}$  is the number of case-parent trios; for the population-based design, it is  $2N_{c-c}$ , where  $N_{c-c}$  is the number of case-control pairs.

The results shown in these tables suggest that the unmatched case-control design is more efficient than the matched case-control design and the case-parent designs (both conditional and unconditional) under all scenarios. Between the conditional and unconditional case-parent



**FIGURE 1.** Relative efficiencies (compared with the smallest sample size) to detect gene-gene interaction for the family-based and population-based study designs when  $R_1 = 1$ ,  $R_2 = 1$ , population prevalence is 10%, and susceptible proportions of the two loci  $\phi$  are small to moderate. From top row to bottom row: three combinations of proportions of susceptible persons, (0.1, 0.1), (0.1, 0.2), and (0.1, 0.25). From left column to right column: the conditional case-parent design, the unconditional case-parent design, the matched case-control design, and the unmatched case-control design. Relative efficiency is defined as the required sample size under one scenario divided by the smallest required sample size among all possible scenarios.

designs, we need approximately 1.5 times as many samples for the analysis conditional on parental mating types as the analysis not conditional on parental mating types to achieve the same power.

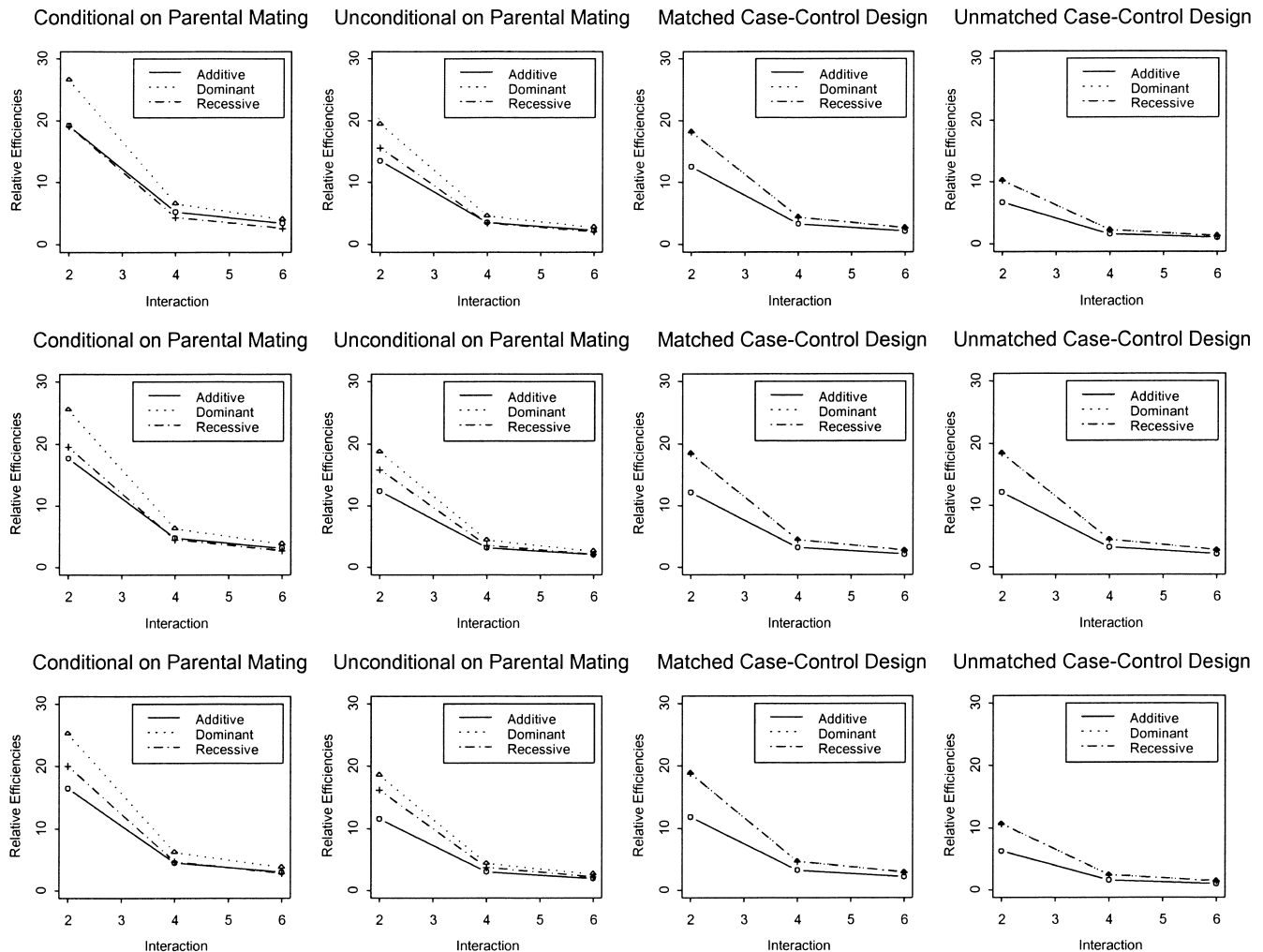
With regard to comparison between the matched case-control design and the conditional case-parent design, the sample size requirement depends on both the population prevalence and the interaction models. For pure interaction models without main effects, the conditional case-parent design is more efficient than the matched case-control design when the population prevalence is 1 percent. This result is consistent with Gauderman's finding for a rare disease, even though he compared sample units rather than total sample sizes. On the other hand, the matched case-control design is more efficient than the conditional case-parent design when

the disease prevalence is 10 percent. For interaction models with main effects, when the disease has a population prevalence of 10 percent, the matched case-control design is more efficient than the conditional case-parent design under the additive and dominant models, but it is slightly less efficient under the recessive models. When the disease prevalence is 1 percent, the conditional case-parent design is more efficient than the matched case-control design.

For each susceptible proportion combination between the two genes, we observe a decreasing trend in the sample size requirement for the three genetic models as the magnitude of the gene-gene interaction increases from  $R_{\text{inter}} = 2$  to  $R_{\text{inter}} = 6$ .

Regarding the sample sizes needed to detect gene-gene interactions under different genetic models, there is no clear pattern for family-based designs, but there is a clear and

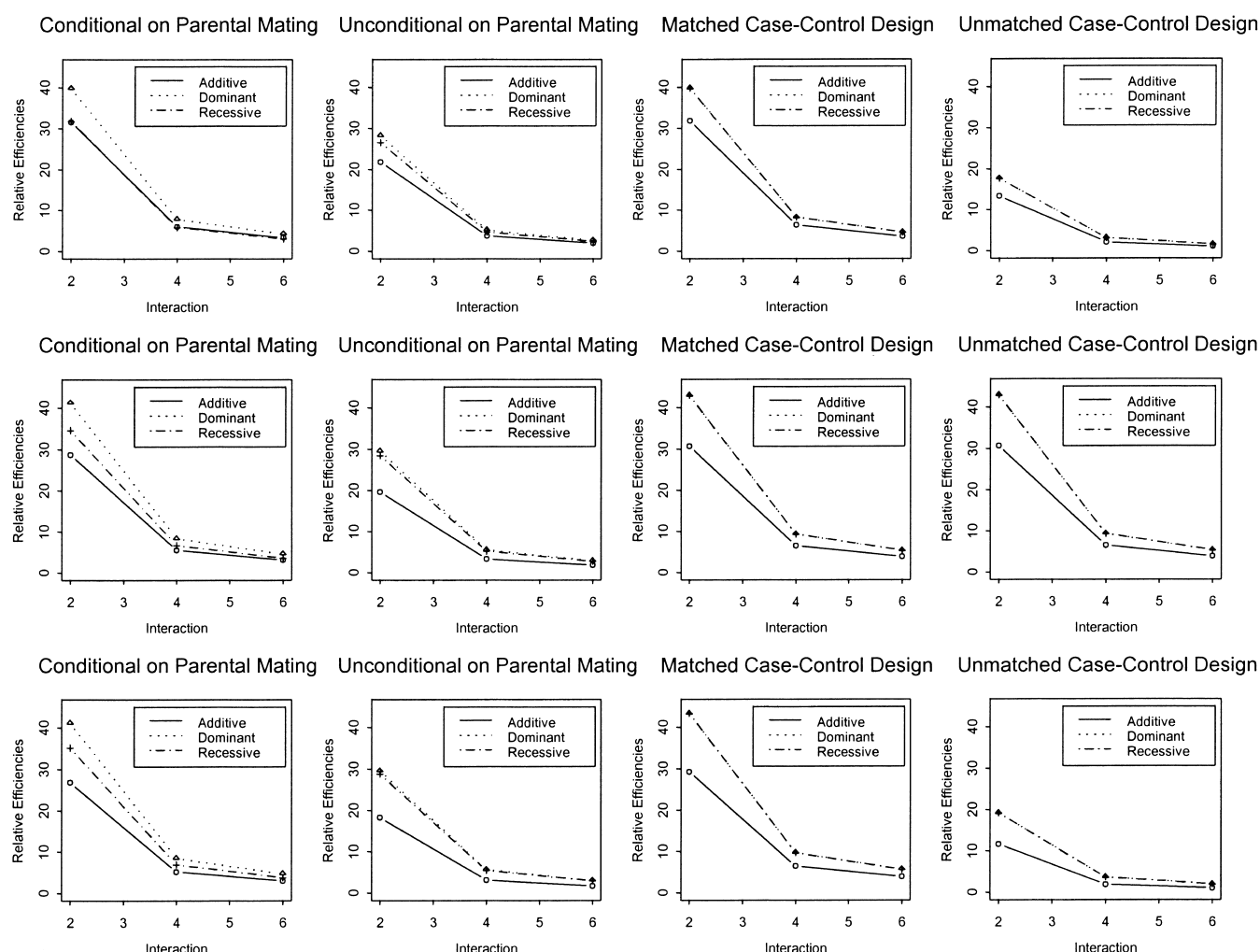




**FIGURE 2.** Relative efficiencies (compared with the smallest sample size) to detect gene-gene interaction for the family-based and population-based study designs when  $R_1 = 1$ ,  $R_2 = 1$ , population prevalence is 10%, and susceptible proportions of the two loci  $\phi$  are moderate to high. From top row to bottom row: three combinations of proportions of susceptible persons, (0.2, 0.2), (0.2, 0.25), and (0.25, 0.25). From left column to right column: the conditional case-parent design, the unconditional case-parent design, the matched case-control design, and the unmatched case-control design. Relative efficiency is defined as the required sample size under one scenario divided by the smallest required sample size among all possible scenarios.

consistent pattern for population-based designs. For population-based designs, we have higher power under the additive model and the same power for the dominant and recessive models across both population prevalence cases and different interaction models. For the family-based design, we first examine the situation where the population prevalence is 10 percent. In this case, we have the highest power to detect gene-gene interactions under a recessive model across different interaction models and susceptible proportions at two loci for the conditional case-parent design. For the unconditional case-parent design, we have the highest power to detect gene-gene interaction under a recessive model when the proportion of susceptible persons at one gene is 0.2 or less. When the susceptible proportion is 0.25 at

both genes, we have higher power to detect interactions under the additive model than under the recessive model. When the population prevalence is 1 percent, the power is higher under the additive model for most susceptible proportion combinations. These comparisons can be better visualized in figures 1, 2, 3, 4, 5, 6, 7, and 8, where the total sample sizes ( $3N_{c-p}$  for the case-parent designs and  $2N_{c-c}$  for the case-control designs) required under three gene-gene interaction alternatives are compared according to the relative efficiencies. We note that the relative efficiencies among different models are similar between pure interaction models and interaction models with main effects, and more samples are needed under interaction models with main effects.



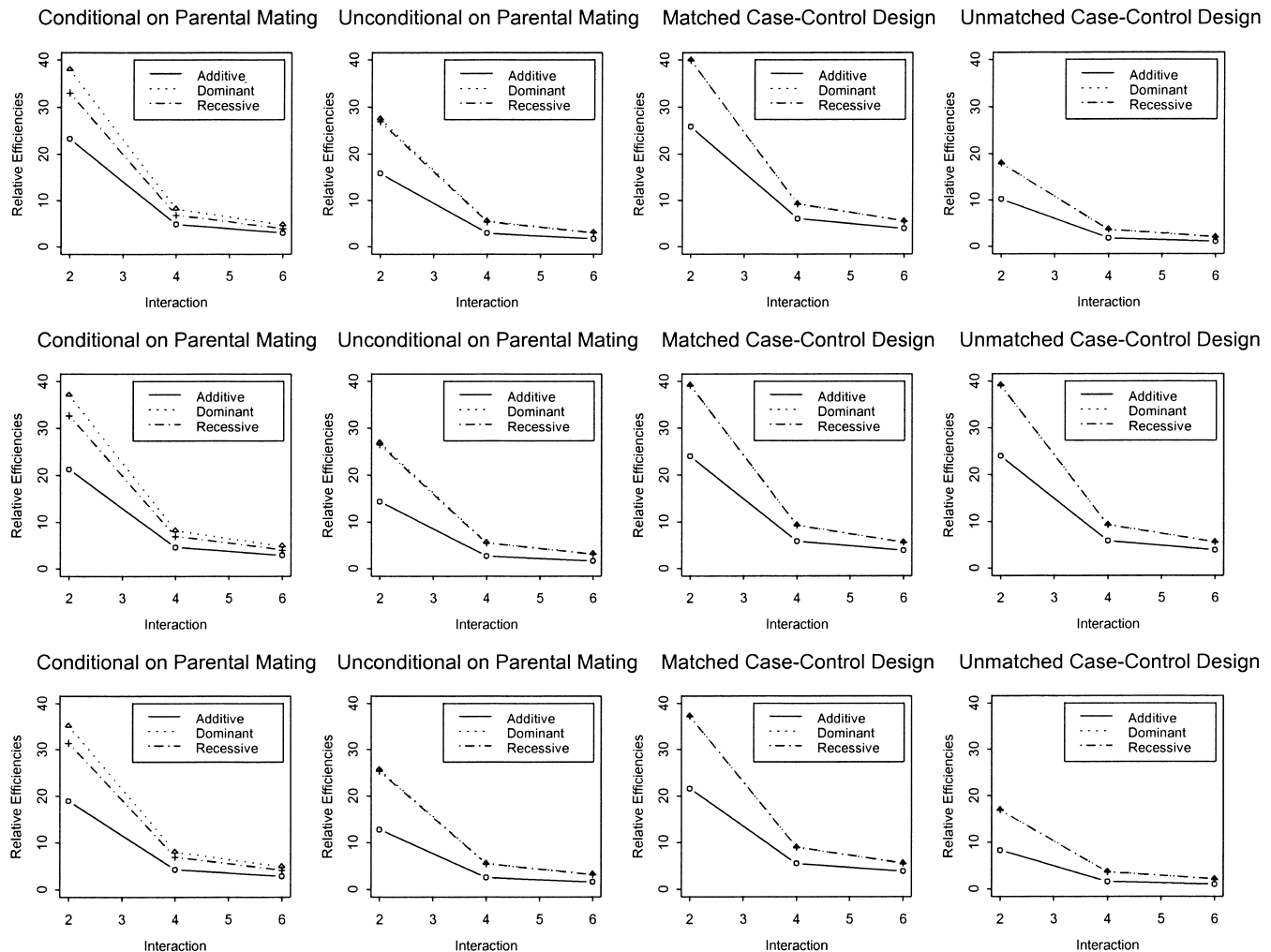
**FIGURE 3.** Relative efficiencies (compared with the smallest sample size) to detect gene-gene interaction for the family-based and population-based study designs when  $R_1 = 1$ ,  $R_2 = 1$ , population prevalence is 1%, and susceptible proportions of the two loci  $\phi$  are small to moderate. From top row to bottom row: three combinations of proportions of susceptible persons, (0.1, 0.1), (0.1, 0.2), and (0.1, 0.25). From left column to right column: the conditional case-parent design, the unconditional case-parent design, the matched case-control design, and the unmatched case-control design. Relative efficiency is defined as the required sample size under one scenario divided by the smallest required sample size among all possible scenarios.

## DISCUSSION

With the rapid progress being made in the identification of polymorphic markers, exploring the role of gene-gene interaction in determining traits of interest has become more and more relevant. In this article, we calculate and compare the sample sizes needed to detect gene-gene interactions for the family-based designs and the population-based designs. Our results suggest that the population-based designs are more powerful than the family-based designs for detecting gene-gene interactions when the disease prevalence is moderate (10 percent). When the disease prevalence is low (1 percent), the unmatched case-control design is still the most powerful design, whereas the matched case-control design is less powerful than both the unconditional case-parent design and

the conditional case-parent design. For both prevalence levels, the unmatched case-control design is more powerful than the matched case-control design, and the unconditional case-parent design is more powerful than the conditional case-parent design.

For simplicity, we have considered candidate genes with biallelic markers. Although it is possible to extend our models to incorporate multiple alleles at a marker, biallelic markers are much more common than other types of markers, and it is possible to group multiple alleles into two groups in modeling interactions. We have also focused only on gene-gene interactions between two markers. A complete analysis of interaction among all important markers is often not feasible either, because of the rapid increase in the

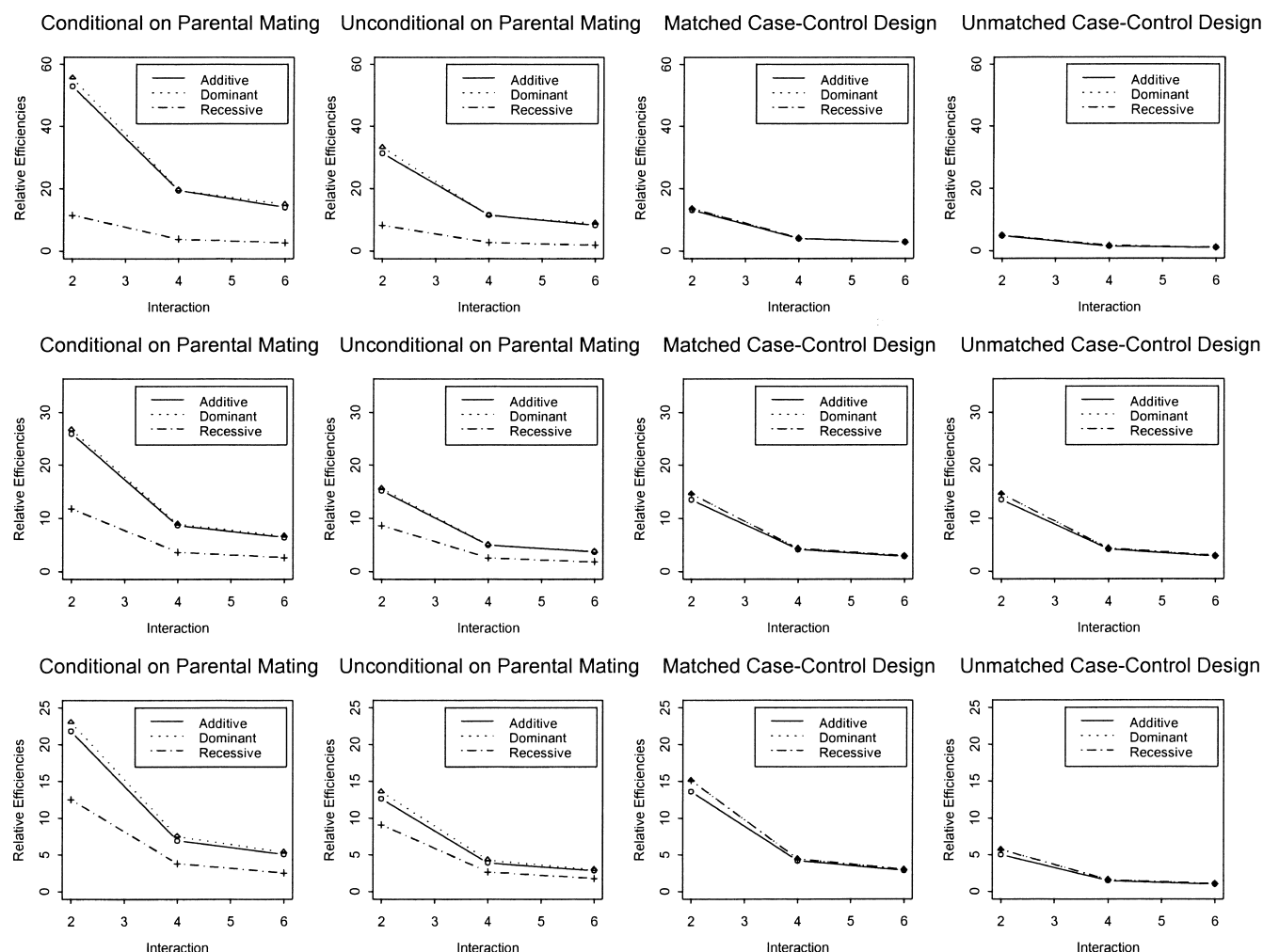


**FIGURE 4.** Relative efficiencies (compared with the smallest sample size) to detect gene-gene interaction for the family-based and population-based study designs when  $R_1 = 1$ ,  $R_2 = 1$ , population prevalence is 1%, and susceptible proportions of the two loci  $\phi$  are moderate to high. From top row to bottom row: three combinations of proportions of susceptible persons, (0.2, 0.2), (0.2, 0.25), and (0.25, 0.25). From left column to right column: the conditional case-parent design, the unconditional case-parent design, the matched case-control design, and the unmatched case-control design. Relative efficiency is defined as the required sample size under one scenario divided by the smallest required sample size among all possible scenarios.

number of analyses required and its consequential cost in type I error. We have used 0.05 as the type I error rate. This is valid if the two testing marker loci are known to be associated with the disease of interest from the previous studies. However, for a more general study, when it is not known whether the marker loci are associated with the disease, a more stringent criterion is needed.

In his paper, Gauderman (19) focused on a disease with a population prevalence of 0.01 percent. For a population-based association study, he investigated the matched case-control design instead of the case-control design to avoid population stratification plus some other confounding biases. Under these scenarios, it was found that the conditional case-parent design is more efficient than the matched case-control design. We were able to reproduce Gauder-

man's results when fixing the population prevalence at this low level and a more moderate level of 1 percent. However, when we considered a more common disease with a population prevalence of 10 percent, we found that the matched case-control design is more powerful than the conditional case-parent design. Our results also showed greater efficiency for the matched case-control design than for the unconditional case-parent design when the susceptible proportion of one locus is 10 percent. Compared with the results for the matched case-control design, the sample sizes required by the unmatched case-control design are systematically lower, which suggests that more efficiency can be obtained through the use of the unmatched case-control design. Consistent with another finding made by Gauderman when the population prevalence is very low, when the popu-

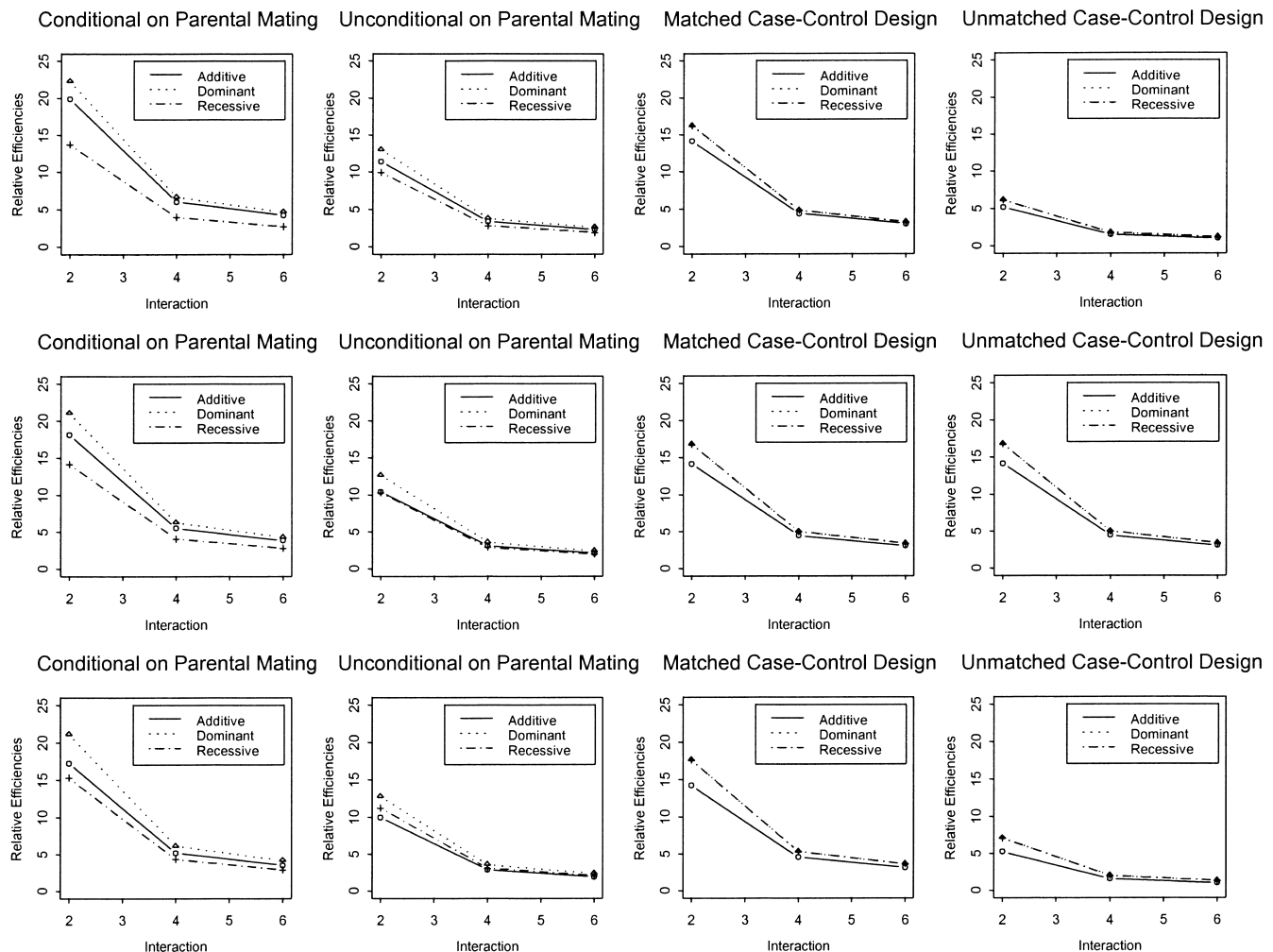


**FIGURE 5.** Relative efficiencies (compared with the smallest sample size) to detect gene-gene interaction for the family-based and population-based study designs when  $R_1 = 3$ ,  $R_2 = 3$ , population prevalence is 10%, and susceptible proportions of the two loci  $\phi$  are small to moderate. From top row to bottom row: three combinations of proportions of susceptible persons, (0.1, 0.1), (0.1, 0.2), and (0.1, 0.25). From left column to right column: the conditional case-parent design, the unconditional case-parent design, the matched case-control design, and the unmatched case-control design. Relative efficiency is defined as the required sample size under one scenario divided by the smallest required sample size among all possible scenarios.

lation prevalence is 10 percent, our results suggest that there is higher power to detect interactions under recessive models than under dominant models for the family-based designs and the same power under the recessive and dominant models for the population-based designs. For family-based designs, the relative sample sizes needed under additive models as compared with the other two genetic models change with different main effects, different population prevalences, and different susceptible proportions at two candidate loci.

We note that when disease prevalence is moderate (10 percent) and when both the susceptible proportions of the two loci and the interaction effect are moderate ( $\phi \geq 0.2$  and  $R_{\text{inter}} \geq 4$ ), we need at most several hundred subjects to detect an effect of gene-gene interaction for the family-based

designs across the three genetic models. For the population-based designs, we need fewer than 250 people. When both the susceptible proportions of the two loci and the interaction effect are very small ( $\phi \leq 0.1$  and  $R_{\text{inter}} \leq 2$ ), the sample size requirements for the three genetic models for the family-based design and the matched case-control design are unrealistically large, requiring several thousand subjects or more to achieve reasonable power. For the unmatched case-control design, the required sample sizes are possible to achieve but still relatively large for genetic studies. When the disease prevalence is low (1 percent), the sample size requirements for the unmatched case-control design are reasonable for all of the parameter combinations considered, but the required sample sizes are too large for the matched case-



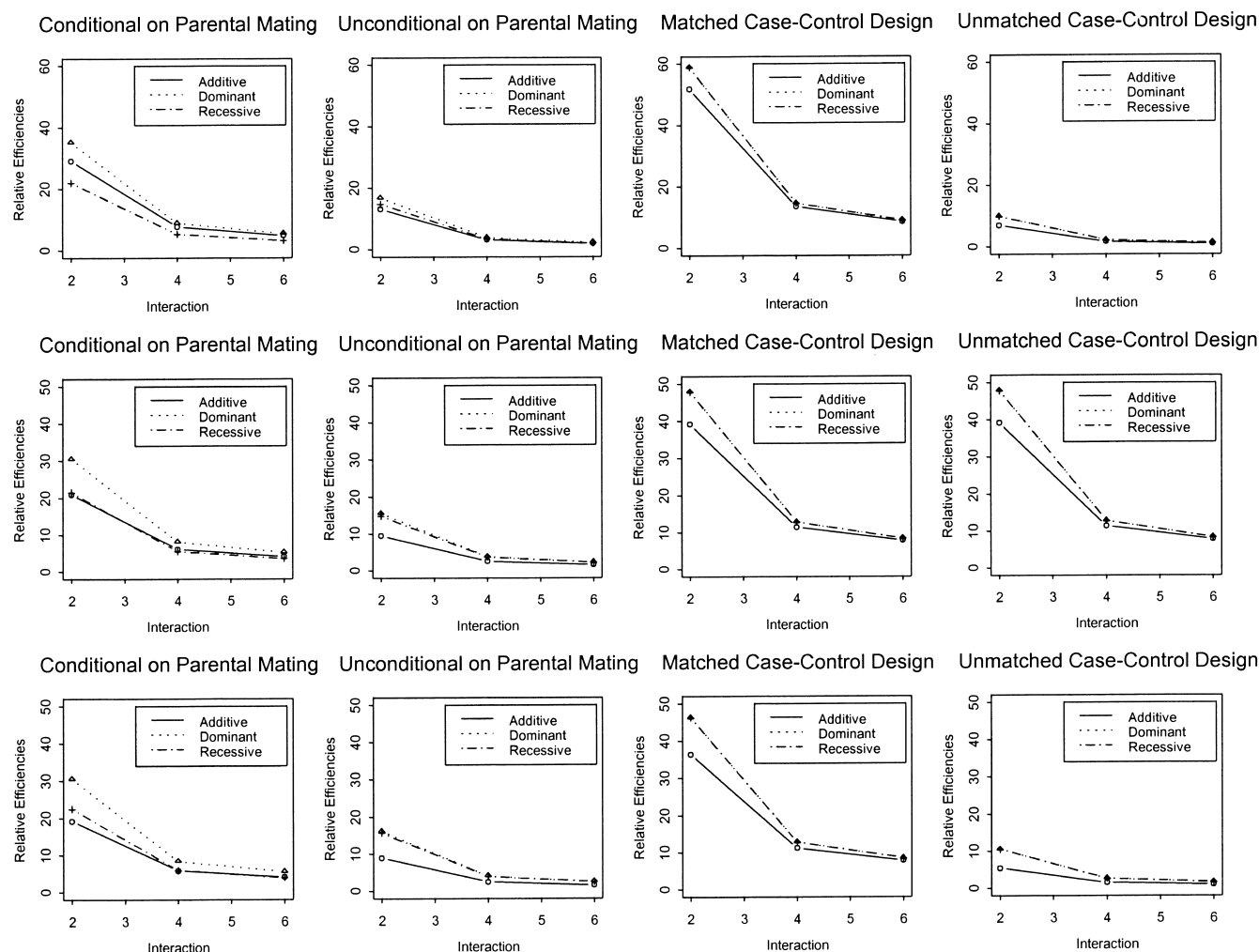
**FIGURE 6.** Relative efficiencies (compared with the smallest sample size) to detect gene-gene interaction for the family-based and population-based study designs when  $R_1 = 3$ ,  $R_2 = 3$ , population prevalence is 10%, and susceptible proportions of the two loci  $\phi$  are moderate to high. From top row to bottom row: three combinations of proportions of susceptible persons, (0.2, 0.2), (0.2, 0.25), and (0.25, 0.25). From left column to right column: the conditional case-parent design, the unconditional case-parent design, the matched case-control design, and the unmatched case-control design. Relative efficiency is defined as the required sample size under one scenario divided by the smallest required sample size among all possible scenarios.

control design and family-based designs when the gene-gene interaction effect is small.

Note that in the matched case-control design, we assume the same genotype probabilities for all matched case-control pairs. This assumption may not be realistic for the matched case-control design. For example, if we consider age as the matching variable, in an age-matched study of candidate disease loci and cardiovascular disease, the independence assumption would imply that cardiovascular disease incidence among persons with genotype  $aa$  does not vary with age. In this case, the sample size calculated above may seriously underestimate the actual sample size needed for the required power. Many research groups have addressed this issue in the literature. The method proposed by Dupon (22) can be used to obtain an accurate power calculation by

modeling the dependence between genotypes of cases and genotypes of controls within matched case-control pairs using the correlation coefficient for genotypes. The procedure of Fleiss and Levin (23) can be performed in order to employ Schlesselman's calculation (24) first and then introduce odds ratios for genotypes for matched cases and controls. Although the implications of the departure from our simple assumption for the matched case-control design need further study, our general conclusion is nevertheless not affected, because our results indicate that the unmatched case-control design is more powerful than the matched case-control design even when the sample size is probably underestimated for the matched design under our assumption.

We note that even though the biologic principle behind epistasis is intuitive, phenotypes represent unpredictable

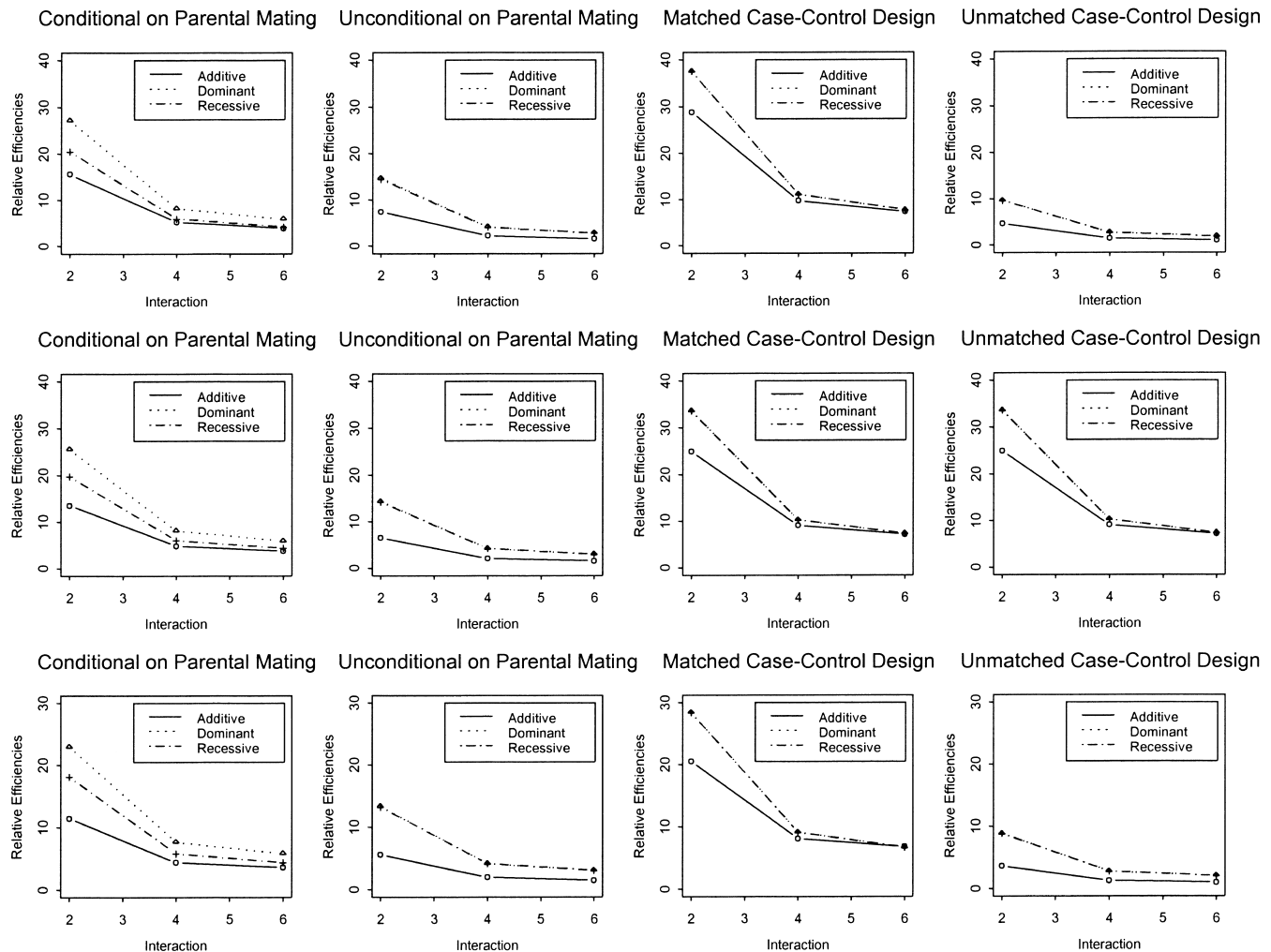


**FIGURE 7.** Relative efficiencies (compared with the smallest sample size) to detect gene-gene interaction for the family-based and population-based study designs when  $R_1 = 3$ ,  $R_2 = 3$ , population prevalence is 1%, and susceptible proportions of the two loci  $\phi$  are small to moderate. From top row to bottom row: three combinations of proportions of susceptible persons, (0.1, 0.1), (0.1, 0.2), and (0.1, 0.25). From left column to right column: the conditional case-parent design, the unconditional case-parent design, the matched case-control design, and the unmatched case-control design. Relative efficiency is defined as the required sample size under one scenario divided by the smallest required sample size among all possible scenarios.

results from disease determinants. Therefore, the detection of a statistical interaction does not necessarily imply interaction on the biologic level, and it may be problematic to interpret the interaction biologically. Moreover, there are many ways in which genes can interact. The scale of the measurement is another problem we may face when modeling gene-gene interaction. It is well known that certain measurement scales of phenotype can give the impression that the interaction exists when it is actually an artifact of the scale used. Since the presence of a statistical interaction depends on the measurement scale, that is, on whether we are modeling the penetrance or the log odds, it is even harder to interpret its biologic meaning. Clearly, issues of scale and interpretation

of underlying biologic meaning need additional investigation.

We have assumed a homogeneous population in this article. However, there may be potential population heterogeneity. The unmatched case-control design and the unconditional case-parent design are valid only when the study population is homogeneous, and these designs may be biased in the presence of population stratification. The major advantage of the case-parent design in comparison with the case-control design is its robustness against population stratification in the detection of genes underlying traits of interest. However, this robustness for gene identification no longer holds for the detection of gene-gene interaction. To demonstrate this, we have conducted simulations by consid-



**FIGURE 8.** Relative efficiencies (compared with the smallest sample size) to detect gene-gene interaction for the family-based and population-based study designs when  $R_1 = 3$ ,  $R_2 = 3$ , population prevalence is 1%, and susceptible proportions of the two loci  $\phi$  are moderate to high. From top row to bottom row: three combinations of proportions of susceptible persons, (0.2, 0.2), (0.2, 0.25), and (0.25, 0.25). From left column to right column: the conditional case-parent design, the unconditional case-parent design, the matched case-control design, and the unmatched case-control design. Relative efficiency is defined as the required sample size under one scenario divided by the smallest required sample size among all possible scenarios.

ering two populations with different disease prevalences and different allele frequencies at each of the two candidate genes. We have found that even when these two candidate genes have no interaction effects on the disease risk under the logistic model in either population, an interaction effect may be identified when a single logistic model is fitted to the case-parent samples from the combined population. Such interaction effects can be substantial, especially when the main effects of the same allele are opposite in the two populations—for example, when allele *A* increases disease risk in the first population but reduces disease risk in the second population.

Therefore, all of the designs considered in this article are potentially subject to bias due to population stratification. Conditional case-parent analysis is no longer superior to

unconditional case-parent analysis in terms of population stratification. Approaches other than the case-parent design are needed to ensure that the identification of gene-gene interaction is not subject to bias caused by population stratification. It is likely that genomic control methods can be used as a possible correction (14). Such methods use genomic markers believed to be independent of the disease and the candidate genes to estimate background association due to population stratification. If a positive association of disease with the genomic marker is detected, that indicates the existence of population structure. The candidate gene can then be adjusted for population stratification. However, the magnitude of gene-gene interaction between two adjusted candidate markers requires further investigation.

## ACKNOWLEDGMENTS

This work was supported in part by grant GM59507 from the National Institutes of Health.

The authors are grateful to Dr. Shuanglin Zhang for helpful discussions.

## REFERENCES

1. Bateson W. Facts limiting the theory of heredity. *Science* 1907; 26:649–60.
2. Andersson C, Hamsten A, Karpe F. Gene-gene interaction between apoB and apoE in determining plasma levels and heterogeneity of apoB-containing lipoproteins. (Abstract 2806). *Circulation* 1998;17(suppl 1):534.
3. Cox NJ, Frigge M, Nicolae DL, et al. Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nature Genet* 1999;21:213–15.
4. Barnes KC, Mathias RA, Nickel R, et al. Testing for gene-gene interaction controlling total IgE in families from Barbados: evidence of sensitivity regarding linkage heterogeneity among families. *Genomics* 2001;71:246–51.
5. Dong CH, Wang S, Li WD, et al. Interacting genetic loci on chromosomes 20 and 10 influence extreme human obesity. *Am J Hum Genet* 2003;72:115–24.
6. Cordell HJ, Wedig GC, Jacobs KB, et al. Multilocus linkage tests based on affected relative pairs. *Am J Hum Genet* 2000; 66:1273–86.
7. Cordell HJ, Todd JA, Hill NJ, et al. Statistical modeling of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type I diabetes. *Genetics* 2001;158:357–67.
8. Mitchell BD, Ghosh S, Schneider JL, et al. Power of variance component linkage analysis to detect epistasis. *Genet Epidemiol* 1997;14:1017–22.
9. Falk CT, Rubinstein P. Haplotype relative risks: an easy, reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet* 1987;51:227–33.
10. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993; 52:506–15.
11. Risch N, Merikangas K. The future of genetic studies of complex human disease. *Science* 1996;273:1516–19.
12. Risch N, Teng J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. I. DNA pooling. *Genome Res* 1998;12:1273–88.
13. Risch N. Searching for genetic determinants in the new millennium. *Nature* 2000;405:847–56.
14. Van den Oord EJ. A comparison between different designs and tests to detect QTLs in association studies. *Behav Genet* 1999; 29:245–56.
15. Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999;55:997–1004.
16. Bacanu SA, Devlin B, Roeder K. The power of genomic control. *Am J Hum Genet* 2000;60:676–90.
17. Pritchard JK, Stephens M, Rosenberg NA, et al. Association mapping in structured population. *Am J Hum Genet* 2000;67: 170–81.
18. Zhang SL, Zhao HY. Quantitative similarity-based association test using population samples. *Am J Hum Genet* 2001;69:601–14.
19. Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. *Am J Epidemiol* 2002;155:478–84.
20. Schaid DJ. Case-parents design for gene-environment interaction. *Genet Epidemiol* 1999;16:261–73.
21. Press WH, Teukolsky SA, Vetterling WT, et al. Numerical recipes in C. New York, NY: Cambridge University Press, 1992.
22. Dupon WD. Power calculation for matched case-control studies. *Biometrics* 1988;44:1157–68.
23. Fleiss JL, Levin B. Sample size determination in studies with matched pairs. *J Clin Epidemiol* 1988;47:727–30.
24. Schlesselman JJ. Case-control studies: design, conduct, analysis. New York, NY: Oxford University Press, 1982.