Genetic model selection in two-phase analysis for case–control association studies

GANG ZHENG

*Office of Biostatistics Research, National Heart, Lung and Blood Institute,
6701 Rockledge Drive, Bethesda, MD 20892-7931, USA*

HON KEUNG TONY NG*

*Department of Statistical Science, Southern Methodist University,
3225 Daniel Avenue, PO Box 750332,
Dallas, TX 75275-0332, USA
ngh@mail.smu.edu*

SUMMARY

The Cochran–Armitage trend test (CATT) is well suited for testing association between a marker and a disease in case–control studies. When the underlying genetic model for the disease is known, the CATT optimal for the genetic model is used. For complex diseases, however, the genetic models of the true disease loci are unknown. In this situation, robust tests are preferable. We propose a two-phase analysis with model selection for the case–control design. In the first phase, we use the difference of Hardy–Weinberg disequilibrium coefficients between the cases and the controls for model selection. Then, an optimal CATT corresponding to the selected model is used for testing association. The correlation of the statistics used for selection and the test for association is derived to adjust the two-phase analysis with control of the Type-I error rate. The simulation studies show that this new approach has greater efficiency robustness than the existing methods.

Keywords: Cochran–Armitage trend test; Disease risk; Efficiency robustness; Hardy–Weinberg disequilibrium; SNP.

1. INTRODUCTION

The case–control design is often employed for testing association between a marker and a disease. A single-nucleotide polymorphism (SNP) is a marker with 2 alleles, denoted as A and B , where A is the at-risk one. Its genotypes are denoted as AA , AB , or BB . Testing genetic association is equivalent to finding any general association between the outcome and the observed genotypes in a 2×3 contingency table, for which Pearson's chi-squared test can be used. It has an asymptotic chi-squared distribution with 2 degrees of freedom (df) under the null hypothesis. However, as A is the risk allele, an individual with genotype AA is more likely to have disease than an AB individual, who in turn is more likely to have disease than a BB individual. The Cochran–Armitage trend test (CATT) (Cochran, 1954;

*To whom correspondence should be addressed.

Armitage, 1955), which utilizes this risk model, is usually more powerful than Pearson's chi-squared test with 2 df (Zheng *and others*, 2006). CATT has been suggested for genetic association in the case-control design (Sasieni, 1997). It is known that the statistical properties of CATT depend on the choice of scores and that the best (or optimal) set of scores which maximizes the power of the test depends on the genetic model (Sasieni, 1997; Slager and Schaid, 2001; Zheng *and others*, 2003). For complex diseases, however, the genetic models of the disease loci are unknown. Therefore, analysis using the CATT based on one score is not robust (Graubard and Korn, 1987; Freidlin *and others*, 2002). In this situation, one usually uses the CATT optimal for the additive model (Sasieni, 1997). On the other hand, when the true model is unknown, robust tests are preferable. The maximin efficiency robust test (MERT) (Gastwirth, 1966, 1985) and the maximum of the 3 CATTs (MAX) were studied in Freidlin *and others* (2002). Here, we follow the criterion of efficiency robustness in the previous articles and say that one test has greater efficiency robustness across a set of plausible models than another test when the minimum power of the first test is higher than that of the second test. Wang and Sheffield (2005) introduced a restricted likelihood ratio test for the case-control data and showed that it had comparable power with MAX. Empirical results also demonstrate that MAX has greater efficiency robustness than MERT (Freidlin *and others*, 2002). For comparison with our proposed method, we only consider CATT, MAX, and MERT.

Witke-Thompson *and others* (2005) studied the directions (signs) of the Hardy-Weinberg disequilibrium (HWD) coefficients when Hardy-Weinberg equilibrium (HWE) holds in the population and used these to confirm the underlying genetic model. We further show that HWD coefficients can be used to divide the parameter space into 4 unique regions, from which genetic models can be selected. The selection of genetic models based on the above theory is, however, robust to departure from HWE in the population. Next, we propose a two-phase analysis for genetic association with model selection. In the first phase, we apply the difference of HWD coefficients between the cases and the controls to classify the underlying genetic model into 3 categories: the recessive region, additive/multiplicative region, or dominant region. In the second phase, we apply the appropriate CATT, optimal for the selected model, to test genetic association. Such two-phase selection-testing analysis has been studied by Hogg (1974) and extensively studied in clinical trials (e.g. Thall *and others*, 1988), but not in genetic analysis. We further derive and adjust the asymptotic correlation between the 2 analyses so that the two-phase analysis with model selection has an approximately correct size. Extensive simulation studies are also conducted to compare this new approach with the existing approaches. Finally, for illustration, we apply the two-phase analysis with model selection to real data.

2. NOTATION AND MODEL

Denote the allele frequencies in cases and controls as $p = \Pr(A|\text{case})$ and $q = \Pr(A|\text{control})$ and the genotypes as $g_0 = BB$, $g_1 = AB$, and $g_2 = AA$. Assume that r cases and s controls are independently sampled. The observed counts for genotypes (g_0, g_1, g_2) are (r_0, r_1, r_2) in cases and (s_0, s_1, s_2) in controls. The observed genotype counts (r_0, r_1, r_2) and (s_0, s_1, s_2) follow multinomial distributions $\text{mul}(r; p_0, p_1, p_2)$ and $\text{mul}(s; q_0, q_1, q_2)$, respectively, where $p_i = \Pr(g_i|\text{case})$ and $q_i = \Pr(g_i|\text{control})$ for $i = 0, 1, 2$. The null hypothesis of no association can be stated as $H_0: p_i = q_i$ for $i = 0, 1, 2$. Denote the disease prevalence as $k = \Pr(\text{case})$. Then, $p_i = \Pr(g_i)f_i/k$ and $q_i = \Pr(g_i)(1 - f_i)/(1 - k)$, where $f_i = \Pr(\text{case}|g_i) > 0$ is the penetrance and $k = \sum_{i=0}^2 \Pr(g_i)f_i$. The null hypothesis is equivalent to $H_0: f_0 = f_1 = f_2 = k$. A genetic model is referred to as a relationship between 2 genotype relative risks, denoted as $\lambda_1 = f_1/f_0$ and $\lambda_2 = f_2/f_0$. The null and alternative hypotheses can be equivalently stated as $H_0: \lambda_1 = \lambda_2 = 1$ against $H_1: \lambda_2 \geq \lambda_1 \geq 1$ with at least one inequality. Then, under H_0 , $p = q$. The recessive (REC), additive (ADD), multiplicative (MUL), and dominant (DOM) models refer to $\lambda_1 = 1$, $\lambda_1 = (\lambda_2 + 1)/2$, $\lambda_2 = \lambda_1^2$, and $\lambda_1 = \lambda_2$, respectively.

CATT models association between the outcome as a linear trend in genotypes. In a logistic regression model for case-control data with a single covariate (genotype) coded with scores (0, x , 1) for (g_0 , g_1 , g_2), where $0 \leq x \leq 1$, the score statistic is equivalent to the CATT statistic (Sasieni, 1997)

$$Z_x = \frac{n^{1/2} \sum_{i=0}^2 x_i (sr_i - rs_i)}{[rsn\{n \sum_{i=0}^2 x_i^2 n_i - (\sum_{i=0}^2 x_i n_i)^2\}]^{1/2}}, \quad (2.1)$$

where $n_i = r_i + s_i$ for $i = 0, 1, 2$, $n = r + s$, and $(x_0, x_1, x_2) = (0, x, 1)$. Under H_0 , for a given $x \in [0, 1]$, Z_x asymptotically follows a standard normal distribution $N(0, 1)$. Sasieni (1997), Slager and Schaid (2001), and Zheng *and others* (2003) showed that the optimal choices of x for the REC, ADD (MUL), and DOM models are $x = 0, 1/2$, and 1, respectively.

In genetic association studies, departure from HWE in cases has also been used to test genetic association in the case-control design (Nielsen *and others*, 1998). However, using departure from HWE in cases as a test statistic has lower power for the additive model and no power at all for the multiplicative model (Nielsen *and others*, 1998; Deng *and others*, 2000; Song and Elston, 2003, 2006). In general, $\Pr(g_0) = \{\Pr(B)\}^2 + F\Pr(A)\Pr(B)$, $\Pr(g_1) = 2\Pr(A)\Pr(B)(1 - F)$, and $\Pr(g_2) = \{\Pr(A)\}^2 + F\Pr(A)\Pr(B)$, where F is the Wright coefficient of inbreeding and HWE holds in the population if and only if $F = 0$.

3. TWO-PHASE ANALYSIS WITH GENETIC MODEL SELECTION

3.1 HWD coefficients and genetic models

HWE is usually tested using the HWD coefficients (Weir, 1996), denoted as $\Delta = \Pr(AA) - \{\Pr(AA) + \Pr(AB)/2\}^2$. When HWE holds in the population, $\Delta = 0$, under which Wittke-Thompson *and others* (2005) studied the directions of Δ in cases ($\Delta_p = p_2 - (p_2 + p_1/2)^2$) and controls ($\Delta_q = q_2 - (q_2 + q_1/2)^2$) and their relations with the underlying genetic model. They considered an additive model as $\lambda_1 = \lambda_2/2$ regardless of the reference penetrance f_0 , but here we use the definition $\lambda_1 = (\lambda_2 + 1)/2$.

Additional relations among Δ , Δ_p , and Δ_q are as follows: (i) $\Delta = 0$ does not imply that $\Delta_p = 0$ or $\Delta_q = 0$ and vice versa; (ii) under the null hypothesis of no association, it follows from $p_i = q_i = \Pr(g_i)$ that $\Delta = \Delta_p = \Delta_q$. This implies that, when HWE holds in the population ($\Delta = 0$), genetic association can be tested under the null hypothesis $H_0: \Delta_p = \Delta_q$. On the other hand, under the alternative hypothesis of association, whether or not HWE holds in the population cannot be tested using the retrospective case-control data. Song and Elston (2006) considered a Hardy-Weinberg disequilibrium trend test (HWDTT) based on $H_0: \Delta_p - \Delta_q = 0$. Here, following Wittke-Thompson *and others* (2005), we assume that HWE holds in the population and use the HWDTT for genetic model selection. The sensitivity of departure from HWE is examined empirically in Section B of the supplementary material available at *Biostatistics* online (<http://www.biostatistics.oxfordjournals.org>). Substituting $p_i = \Pr(g_i)f_i/k$ and $q_i = \Pr(g_i)(1 - f_i)/(1 - k)$ into Δ_p and Δ_q with $\Delta = 0$, one obtains the following: $\Delta_p = f_0^2 p^2 q^2 (\lambda_2 - \lambda_1^2)/k^2$ and $\Delta_q = f_0 p^2 q^2 (2\lambda_1 - 1 - \lambda_2 - f_0 \lambda_1^2 + f_0 \lambda_2)/(1 - k)^2$. Wittke-Thompson *and others* (2005) proved that $\Delta_p > 0$ and $\Delta_q < 0$ under the REC model ($\lambda_1 = 1$), $\Delta_p < 0$ and $\Delta_q > 0$ under the DOM model ($\lambda_1 = \lambda_2$), and $\Delta_p = 0$ and $\Delta_q < 0$ under the MUL model ($\lambda_2 = \lambda_1^2$). Under the ADD model ($f_1 = (f_0 + f_2)/2$), it can be shown similarly that both Δ_p and Δ_q are negative. The 4 genetic models with the directions of (Δ_p, Δ_q) are plotted in Figure 1. The open triangle, formed by REC and DOM in Figure 1, indicates the null and alternative hypotheses. The vertex $C(1, 1)$ corresponds to the null hypothesis. The signs of (Δ_p, Δ_q) for the 4 models are indicated with the labels of the genetic models. The curve CD corresponds to $\lambda_2 = (-f_0 \lambda_1^2 + 2\lambda_1 - 1)/(1 - f_0)$, the condition used to determine

Fig. 1. Signs of HWD coefficients and genetic models.

the sign of Δ_q . It follows from Figure 1 that the 4 genetic models plus the curve CD divide the alternative hypothesis into 4 mutually exclusive open regions: R_1 , R_2 , R_3 , and R_4 (boundaries are excluded). The signs of (Δ_p, Δ_q) in the 4 open regions can be uniquely determined when HWE holds in the population (see Section A of the supplementary material available at *Biostatistics* online). Based on the above analysis, the difference of HWD coefficients between cases and controls can be used to classify the REC and DOM models. For example, the REC and DOM models imply that $\Delta_p - \Delta_q > 0$ and $\Delta_p - \Delta_q < 0$, respectively.

3.2 Two-phase analysis with model selection

When the genetic model is unknown, $Z_{1/2}$ has the greatest efficiency robustness among the family of statistics $\{Z_x: x \in [0, 1]\}$ against model misspecification and is often used in applications (Sasieni, 1997; Slager and Schaid, 2001; Freidlin *and others*, 2002; Zheng *and others*, 2005). Other robust tests considered by Freidlin *and others* (2002) are $\text{MAX} = \max(|Z_0|, |Z_{1/2}|, |Z_1|)$ and the MERT given by $Z_{\text{MERT}} = (Z_0 + Z_1)/\{2(1 + \hat{\rho}_{01})\}^{1/2}$, where $\hat{\rho}_{01} = n_0 n_2 / \{n_0 n_2 (n_0 + n_1)(n_1 + n_2)\}^{1/2}$, is a consistent estimate of the asymptotic null correlation between Z_0 and Z_1 . The MERT and MAX have been shown to have high-efficiency robustness in hypothesis testing in survival and ordered categorical data, and genetic data analyses when the model underlying the data is uncertain, but a family of scientifically plausible models and their normally distributed tests are available (Gastwirth, 1966, 1985; Davies, 1977; Zucker and Lakatos, 1990; Whittemore and Tu, 1998; Freidlin *and others*, 1999; Gastwirth and Freidlin, 2000; Shih and Whittemore, 2001).

We propose a two-phase analysis where we classify the underlying genetic model in the first phase followed by testing association using the optimal CATT for the selected model in the second phase. Song and Elston (2006) considered testing association using the test statistic

$$Z_{\text{HWDTT}} = \frac{(rs/n)^{1/2}(\hat{\Delta}_p - \hat{\Delta}_q)}{\{1 - n_2/n - n_1/(2n)\}\{n_2/n + n_1/(2n)\}},$$

where $\hat{\Delta}_p = \hat{p}_2 - (\hat{p}_2 + \hat{p}_1/2)^2$ and $\hat{\Delta}_q = \hat{q}_2 - (\hat{q}_2 + \hat{q}_1/2)^2$ are the estimates of Δ_p and Δ_q , respectively, with $\hat{p}_i = r_i/r$ and $\hat{q}_i = s_i/s$.

From Figure 1 and the signs of (Δ_p, Δ_q) for the 4 genetic models and regions, it seems that the ADD and MUL models cannot be distinguished by the signs as (Δ_p, Δ_q) have the same sign $(-, -)$ under the 2 models. However, $Z_{1/2}$ is asymptotically optimal for both ADD and MUL models under local alternatives. Thus, it is not necessary to distinguish them. On the other hand, the ADD (MUL) model may be distinguished from the REC and DOM models by the signs of (Δ_p, Δ_q) . Thus, in the first phase, we classify the genetic model in the recessive region (R_1) if $Z_{\text{HWDTT}} > c$, the dominant region (R_4) if $Z_{\text{HWDTT}} < -c$, and in the middle region (R_2 and R_3) containing the ADD or MUL models otherwise, where $c > 0$ is a prespecified real number. In the second phase, we choose Z_{model} as a test statistic for association, given by $Z_{\text{model}} = Z_0$ if $Z_{\text{HWDTT}} > c$, Z_1 if $Z_{\text{HWDTT}} < -c$, and $Z_{1/2}$ if otherwise, where we choose $c = \Phi^{-1}(0.95) = 1.645$. The performance of the above model selection is examined and the results are reported in Section B of the supplementary material available at *Biostatistics* online. When $p \geq 0.3$, the probabilities to select correct REC/DOM and ADD/MUL models are greater than 65% and 85% and are robust to departure from HWE.

3.3 Adjustment of the sizes of the two-phase analysis with model selection

To correct the size of Z_{model} analytically, the correlation between the two phases of the analysis needs to be derived. In Section C of the supplementary material available at *Biostatistics* online, we obtain $\text{corr}_{H_0}(Z_{\text{HWDTT}}, Z_0) = \{(1 - p)/(1 + p)\}^{1/2} + O(n^{-1})$, $\text{corr}_{H_0}(Z_{\text{HWDTT}}, Z_{1/2}) = O(n^{-1})$, and $\text{corr}_{H_0}(Z_{\text{HWDTT}}, Z_1) = -\{p/(2 - p)\}^{1/2} + O(n^{-1})$, where $p = \Pr(A|\text{case}) = \Pr(A|\text{control})$ under H_0 . Note that Z_{HWDTT} and $Z_{1/2}$ are asymptotically independent.

Suppose in the second phase the null hypothesis corresponding to the j th SNP ($j = 1, \dots, M$) is rejected at the level α^*/M using the normal distribution. Then, to control the overall level at the given α in the two-phase analysis with model selection for M SNPs, we have $\alpha/M = \Pr_{H_0}(|Z_{1/2}| > z(1 - \alpha^*/(2M)), |Z_{\text{HWDTT}}| < c) + \Pr_{H_0}(|Z_0| > z(1 - \alpha^*/(2M)), Z_{\text{HWDTT}} > c) + \Pr_{H_0}(|Z_1| > z(1 - \alpha^*/(2M)), Z_{\text{HWDTT}} < -c) + \Pr_{H_0}(|Z_{1/2}| > z(1 - \alpha^*/(2M)), |Z_{\text{HWDTT}}| < c) = 0.9\alpha^*/M$, where $z(1 - \alpha)$ is the upper α th percentile of $N(0, 1)$. Using the bivariate normal distribution of (Z_x, Z_{HWDTT}) with correlation ρ_x , α^* satisfies

$$\frac{\alpha - 0.9\alpha^*}{M} = \sum_{x=0}^1 \int_{\Omega_x} \left\{ \Phi\left(\frac{-z(1 - \frac{\alpha^*}{2M}) - \rho_x u}{(1 - \rho_x^2)^{1/2}}\right) + \Phi\left(\frac{-z(1 - \frac{\alpha^*}{2M}) + \rho_x u}{(1 - \rho_x^2)^{1/2}}\right) \right\} d\Phi(u), \quad (3.1)$$

where $\Omega_0 = \{u: u > c\}$, $\Omega_1 = \{u: u < -c\}$, and $\rho_x = \text{corr}_{H_0}(Z_{\text{HWDTT}}, Z_x)$. Given α and M and under HWE in the population, this shows that α^* only asymptotically depends on p ($=q$, under H_0), which can be estimated consistently using the observed data (Weir, 1996) by $\hat{p} = \hat{q} = (2n_2 + n_1)/(2n)$. The supplementary material available at *Biostatistics* online (Section E) includes a SAS IML macro to calculate α^* .

Table 1. Type-I error rates of the robust tests based on 10 000 replications with the nominal level 0.05 using r cases and s controls and Wright's coefficient of inbreeding F . The α is unadjusted and α^* is analytically adjusted levels

F	r	s	p	k	$Z_{1/2}$	MAX	Z_{MERT}	Z_{model}	
								α	α^*
0	500	500	0.1	0.01	0.0486	0.0497	0.0395	0.0621	0.0423
0	500	500	0.1	0.1	0.0477	0.0506	0.0436	0.0606	0.0415
0	250	250	0.3	0.01	0.0482	0.0510	0.0477	0.0607	0.0456
0	250	250	0.3	0.1	0.0525	0.0509	0.0524	0.0634	0.0506
0	250	250	0.5	0.01	0.0476	0.0472	0.0471	0.0590	0.0455
0	250	250	0.5	0.1	0.0511	0.0514	0.0512	0.0640	0.0506
0	400	600	0.1	0.01	0.0481	0.0503	0.0427	0.0703	0.0505
0	400	600	0.1	0.1	0.0468	0.0448	0.0409	0.0644	0.0459
0	200	300	0.3	0.01	0.0472	0.0498	0.0454	0.0628	0.0487
0	200	300	0.3	0.1	0.0469	0.0468	0.0455	0.0630	0.0481
0	200	300	0.5	0.01	0.0452	0.0446	0.0452	0.0597	0.0470
0	200	300	0.5	0.1	0.0522	0.0539	0.0523	0.0654	0.0531
0.05	500	500	0.1	0.01	0.0490	0.0562	0.0477	0.0648	0.0475
0.05	500	500	0.1	0.1	0.0438	0.0476	0.0443	0.0580	0.0411
0.05	250	250	0.3	0.01	0.0485	0.0486	0.0500	0.0625	0.0472
0.05	250	250	0.3	0.1	0.0466	0.0468	0.0467	0.0612	0.0462
0.05	250	250	0.5	0.01	0.0491	0.0519	0.0490	0.0627	0.0494
0.05	250	250	0.5	0.1	0.0503	0.0514	0.0504	0.0627	0.0489

Table 1 (the last column) reports the simulated levels of significance using the adjusted significance level (α^*) for Z_{model} . The adjusted level α^* was calculated for each simulated data set. The sizes of the analytically adjusted level (α^*) based on (3.1) are not significantly different from the nominal level. For rare allele frequency ($p = 0.1$), the level α^* leads to a conservative test. The last part of Table 1 also shows that departure from HWE does not inflate the Type-I error rates of the 4 genotype-based test statistics (see comments of Sasieni, 1997).

4. NUMERICAL RESULTS

In Section D of the supplementary material available at *Biostatistics* online, we reported simulation studies to compare power for various approaches. Our results show that the two-phase analysis with model selection (Z_{model}) has greater efficiency robustness than MAX, although their power difference is about 2–3%. The Z_{model} has at least 95% efficiency under ADD/MUL and 100–225% (100–123%) under REC (DOM) relative to the CATT.

To illustrate the use of the two-phase model selection, we consider the top 2 SNPs in a genome-wide case-control study for age-related macular degeneration (AMD) presented in Klein *and others* (2005). After the quality control of genotyping, there were 103 611 SNPs left for testing association. Each SNP was tested at the level $\alpha = 0.05/103\,611 = 4.82 \times 10^{-7}$. The top 2 SNPs (rs380390 and rs1329428) with smallest p -values were found in a close region and were further studied by biology and haplotype analysis. The results show that this region is associated with AMD. For both SNPs, the underlying genetic models are uncertain.

The data set for the top SNP (rs380390) was given by $(r_0, r_1, r_2) = (50, 35, 11)$ for cases and $(s_0, s_1, s_2) = (6, 25, 19)$ for controls. The CATT with the additive score is $Z_{1/2} = 5.1171$ with unadjusted

p -value $3.1 \times 10^{-7} < \alpha$. The MERT is $Z_{\text{MERT}} = 5.0735$ with unadjusted p -value $3.91 \times 10^{-7} < \alpha$. Further, we have $\text{MAX} = Z_{1/2} = 5.1171$ and its simulated p -value is 5.9×10^{-7} based on 10^8 replicates, which indicates that the association is marginally significant. For the model selection, using (3.1), we obtained the adjusted significance level $\alpha^* = 0.0387$ and $Z_{\text{HWDTT}} = 1.0283$. This suggests that we use $Z_{\text{model}} = Z_{1/2}$ with p -value 3.1×10^{-7} . After Bonferroni correction, the adjusted p -value is $0.032 < \alpha^*$. Thus, the SNP is significantly associated with the AMD after Bonferroni correction. Note that the two-phase analysis selected the largest CATT in the first phase.

The second top SNP (rs1329428) was given by $(r_0, r_1, r_2) = (2, 24, 68)$ for cases and $(s_0, s_1, s_2) = (5, 29, 14)$ for controls. Using (3.1), the adjusted significance level is $\alpha^* = 0.0374$. The CATT $Z_{1/2} = 4.9189$ with unadjusted p -value $8.7 \times 10^{-7} > \alpha$. On the other hand, $Z_{\text{MERT}} = 4.4523$ with unadjusted p -value $8.5 \times 10^{-6} > \alpha$. The MAX is equal to $Z_0 = 4.9268$ and its simulated p -value is 6.6×10^{-7} . Since $Z_{\text{HWDTT}} = 1.930$, we select the CATT with the recessive score in the second phase and $Z_{\text{model}} = Z_0 = 4.9268$ with unadjusted p -value 8.36×10^{-7} , which is not significant after Bonferroni correction. In fact, none of the 4 nonadaptive robust tests reached statistical significance; however, the two-phase test selected the appropriate CATT.

5. DISCUSSION

In testing genetic association between a SNP and a disease using a case-control design, we studied a new two-phase analysis, where an underlying genetic model is selected in the first phase based on the difference of HWD coefficients between cases and controls. Then, in the second phase, we use the CATT for association with the score optimal for the selected model. Since two-phase analyses are correlated, we derived their asymptotic null correlations and used them to adjust the level in the two-phase analysis to control the Type-I error rate.

Our simulation results demonstrate that this new approach compares favorably with other existing approaches, including the CATT with the additive score (Sasieni, 1997) and the robust tests (MERT and MAX) of Freidlin *and others* (2002). When the underlying genetic model is uncertain, it has the greatest efficiency robustness among all the tests considered here. Thus, it is useful for genetic data analysis using case-control data. Since it has been shown that MAX and likelihood ratio tests have comparable power, we did not compare our approach to the likelihood ratio test.

One assumption in this article is that HWE holds in the population. Although model selection can be uniquely determined when HWE holds in the population, our empirical results showed that a moderate departure from HWE had little effect on the power performance of the model selection under various genetic models. Nor did departure of HWE inflate Type-I error rates. It is known that case-control designs for genetic association may be affected by population stratifications and/or cryptic relatedness (Devlin and Roeder, 1999; Pritchard and Rosenberg, 1999; Satten *and others*, 2001; Zheng *and others*, 2005; Gorroochurn *and others*, 2006). This needs to be examined in the future. In addition, we chose a fixed threshold value as the upper 95th normal percentile in the model selection phase. Data-driven threshold levels may further improve the power. This also needs future research.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the editor, the associate editor, and the 2 anonymous referees for their constructive comments. The authors would also like to thank Prof. William R. Schucany and Ms Julia Kozlitina for their suggestions which resulted in a much improved version of the manuscript. *Conflict of Interest:* None declared.

REFERENCES

- ARMITAGE, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375–386.
- COCHRAN, W. G. (1954). Some methods for strengthening the common chi-square tests. *Biometrics* **10**, 417–451.
- DAVIES, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64**, 247–254.
- DENG, H. W., CHEN, W. M. AND RECKER, R. R. (2000). QTL fine mapping by measuring and testing for Hardy-Weinberg and linkage disequilibrium at a series of linked marker loci in extreme samples of population. *American Journal of Human Genetics* **66**, 1027–1045.
- DEVLIN, B. AND ROEDER, K. (1999). Genomic control for association studies. *Biometrics* **55**, 997–1004.
- FREIDLIN, B., PODGOR, M. J. AND GASTWIRTH, J. L. (1999). Efficiency robust tests for survival or ordered categorical data. *Biometrics* **55**, 883–886.
- FREIDLIN, B., ZHENG, G., LI, Z. AND GASTWIRTH, J. L. (2002). Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Human Heredity* **53**, 146–152.
- GASTWIRTH, J. L. (1966). On robust procedures. *Journal of the American Statistical Association* **61**, 929–948.
- GASTWIRTH, J. L. (1985). The use of maximin efficiency robust tests in combining contingency tables and survival analysis. *Journal of the American Statistical Association* **80**, 380–384.
- GASTWIRTH, J. L. AND FREIDLIN, B. (2000). On power and efficiency robust linkage tests for affected sibs. *Annals of Human Genetics* **64**, 443–453.
- GORROOCHURN, P., HEIMAN, G. A., HODGE, S. E. AND GREENBERG, D. A. (2006). Centralizing the non-central chi-square: a new method to correct for population stratification in genetic case-control association studies. *Genetic Epidemiology* **30**, 277–289.
- GRAUBARD, B. I. AND KORN, E. L. (1987). Choice of column scores for testing independence in ordered $2 \times K$ contingency tables. *Biometrics* **43**, 471–476.
- HOGG, R. V. (1974). Adaptive robust procedures—partial review and some suggestions for future applications and theory. *Journal of the American Statistical Association* **69**, 909–923.
- KLEIN, R. J., ZEISS, C., CHEW, E. Y., TSAI, J. Y., SACKLER, R. S., HAYNES, C., HENNING, A. K., SANGIOVANNI, J. P., MANE, S. M., MAYNE, S. T. and others (2005). Complement factor H polymorphism in aged-related macular degeneration. *Science* **308**, 385–389.
- NIELSEN, D. M., EHM, M. G. AND WEIR, B. S. (1998). Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. *American Journal of Human Genetics* **63**, 1531–1540.
- PRITCHARD, M. J. AND ROSENBERG, N. A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics* **65**, 220–228.
- SASIENI, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics* **53**, 1253–1261.
- SATTEN, G. A., FLANDERS, W. D. AND YANG, Q. (2001). Account for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *American Journal of Human Genetics* **68**, 466–477.
- SHIH, M. AND WHITEMORE, A. S. (2001). Allele-sharing among affected relatives: nonparametric methods for identifying genes. *Statistical Methods in Medical Research* **10**, 27–55.
- SLAGER, S. L. AND SCHAID, D. J. (2001). Case-control studies of genetic markers: power and sample size approximations for Armitage's test for trend. *Human Heredity* **52**, 149–153.
- SONG, K. AND ELSTON, R. C. (2003). The Hardy-Weinberg disequilibrium (HWD) measure and test statistics for a disease-susceptibility locus with multiple alleles allowing for an inbreeding coefficient (F). *Genetica* **119**, 269–293.

- SONG, K. AND ELSTON, R. C. (2006). A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. *Statistics in Medicine* **25**, 105–126.
- THALL, P. F., SIMON, R. AND ELLENBERG, S. S. (1988). Two-stage selection and testing designs for comparative clinical trials. *Biometrika* **75**, 303–310.
- WANG, K. AND SHEFFIELD, V. C. (2005). A constrained-likelihood approach to marker-trait association studies. *American Journal of Human Genetics* **77**, 768–780.
- WEIR, B. S. (1996). *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sunderland, MA: Sinauer Associations Inc.
- WHITTEMORE, A. S. AND TU, I.-P. (1998). Simple, robust linkage tests for affected sibs. *American Journal of Human Genetics* **62**, 1228–1242.
- WITTKÉ-THOMPSON, J. K., PLUZHNIKOV, A. AND COX, N. J. (2005). Rational inferences about departure from Hardy-Weinberg equilibrium. *American Journal of Human Genetics* **76**, 967–986.
- ZHENG, G., FREIDLIN, B. AND GASTWIRTH, J. L. (2006). Comparison of robust tests for genetic association using case-control studies (2nd special issue in honor of E. L. Lehmann). *IMS Lecture Notes—Monograph Series*, **49**, 253–265.
- ZHENG, G., FREIDLIN, B., LI, Z. AND GASTWIRTH, J. L. (2003). Choice of scores in trend tests for case-control studies of candidate-gene associations. *Biometrical Journal* **45**, 335–348.
- ZHENG, G., FREIDLIN, B., LI, Z. AND GASTWIRTH, J. L. (2005). Genomic control for association studies under various genetic models. *Biometrics* **61**, 186–192.
- ZUCKER, D. M. AND LAKATOS, E. (1990). Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika* **77**, 853–864.

[Received April 16, 2007; first revision July 25, 2007; second revision October 1, 2007;
accepted for publication October 12, 2007]