 Int J Hum Genet, 16(1,2): 16-21 (2016)

Testing the Trend for Genotype Distribution of Hypertension Patients in Case-Control Studies

Özge Karadağ[1] and Serpil Aktaş[2]

Department of Statistics, Faculty of Science, Hacettepe University, Beytepe, Ankara, 06800, Turkey
E-mail: [1]<ozgekaradag@hacettepe.edu.tr>, [2]<spxl@hacettepe.edu.tr>

KEYWORDS 2x*K* Contingency Tables. Blood Pressure. Cochran Armitage Test. Genetic Analysis. Mode of Inheritance

ABSTRACT This study uses data from independent samples taken from two populations, each with the *K* distinct categories, which are the 2x*K* contingency tables. The goal here is to test whether there is a difference between two multinomial populations. When the categories are nominal, Pearson's chi squared test statistic is the most widely used method, but when the categories are ordered, it is not appropriate because of the ordinal nature. Cochran Armitage test is used for assessing the presence of an association for 2x*K* tables for specific scores, which use the ordinal information by assigning different weights. In the genetic studies, if the inheritance model is unknown, the weights are assigned according to the suspected model. The effect of the weights is discussed on *Genetic Analysis Workshop 18* data. It is concluded that test scores have a significant effect on the test results depending on the genetic model.

INTRODUCTION

In genetic analysis, the case-control studies are commonly used for testing the association between the phenotype of interest and the genetic variant. In case-control study design, affected individuals and non-effected controls are compared to specify factors that may cause the disease. In most of the association studies, genetic component of the disease can be considered as genotype. Genotype is a combination of alleles, which describe the variation among the genes. Due to bi-allelic structure of markers, for the gene status allele there are two possibilities, one is dominant and the other is recessive denoted by *A* and *a,* respectively. Human organisms have diploid cells, as each individual inherits two alleles for each gene. Hence, the genotype can be characterized by three different possible categories, *aa, aA* and *AA*. The genotype distribution in each group for a case-control study can be summarized by a 2x3 contingency table as represented in Table 1.

Table 1: Genotype distribution for case-control studies

	aa	*aA*	*AA*	*Total*
Case	n_{10}	n_{11}	n_{12}	n_1
Control	n_{20}	n_{21}	n_{22}	n_2
Total	$n_{.0}$	$n_{.1}$	$n_{.2}$	N

In Table 1, n_{10}, n_{11}, n_{12} and n_{20}, n_{21}, n_{22} are the genotype frequencies and distributed as multinominals for genotypes *aa, aA* and *AA* with probabilities p_0, p_1, p_2 and q_0, q_1, q_2, respectively (Freidlin et al. 2002; Zheng et al. 2003).

Various approaches have been developed for investigating the association between the disease of interest and a SNP such as testing the trend between genotype categories. For testing the trend between the three possible levels of genotype, the Cochran Armitage (CA) trend test can be used (Armitage 1955). In the following section, CA trend test for a genotype based association study is mentioned briefly.

In literature, there are several solutions for testing the trend in a genotype based association study when the inheritance model is unknown. An unconditional test called Max test

Address for correspondence:
Serpil Aktas
Hacettepe University,
Faculty of Science, Department of Statistics,
Beytepe, Ankara, Turkey, 06800
Telephone: +90 312 297 79 30
Fax: +90 312 297 79 13
E-mail: spxl@hacettepe.edu.tr

was proposed by Freidlin et al. (2002), which successfully detects the largest of the three CA tests with the score vectors suggested by Sasieni (1997). Following Zheng et al. (2003), Zheng (2008) introduced a new parameter η, describing the scores as $(0, \eta, 1)$ and aimed at reaching optimal ς. A reformulating of Max test was followed by Hortorn et al. (2009). The authors also generalized the test for conditional inference procedures. All of these proposed methods put emphasis on determining the appropriate scores, which might have a substantial effect on the results. Beside these, the power function of CA trend test was examined in the presence of exposure misclassification problem by Buonaccorsi et al. (2014) and a simulation study on trend test was performed by using multiple genetic models by Talluri et al. (2014).

Objectives

This paper aims to test the trends in genotype distribution when the mode of inheritance is unknown. To evaluate the effects of scores under different genetic model assumptions, researchers analyzed two different blood pressure associated genes in 2x3 contingency table framework.

METHODOLOGY

In genetic association studies, it is aimed to evaluate the association between the genetic marker and the disease. For genetic disorders, the probability of the disease occurring is related to the genotype. The probability of occurring and the number of high-risk alleles are directly proportional. The probability increases with the amount of high-risk alleles. Hence, there is an ordering of the genotype levels. Assuming *A* as the high-risk allele, the individuals with the *aa* genotype are less likely to be at risk compared to individuals with the *AA* genotype.

In a genotype based case-control study design, depending on the ordering information the null hypothesis of no association and the one-sided alternative hypothesis can be written as:

$H_0: f_0 = f_1 = f_2$

$H_A: f_0 \leq f_1 \leq f_2$

Here f_0, f_1 and f_2 denote the disease penetrance for *aa, aA* and *AA* genotypes, respectively. In genetics, penetrance is the proportion of affected individuals with a particular genotype ($f_i = p(case/genotype=i)$, $i=0,1,2$).

The CA trend test investigates the genetic association using the case-control design regardless of Hardy Weinberg equilibrium. The CA trend test also evaluates the linear trend by using weights as a measure of exposure dosage, (Slager et al. 2001). For measuring the allele effects, the weights w_i can be assigned as the number of high risk alleles of an individual for a specific variant for $i=0, 1, 2$. CA trend test statistic is defined as follows:

$$Z = \frac{U}{\sqrt{\widehat{Var}(U)}} \qquad (1)$$

Where, $U = \sum_{i=0}^{2} w_i \left(\frac{n_1}{N} n_{1i} - \frac{n_2}{N} n_{2i}\right)$ and w_i are the weights. The choice of the weights in the test is related to the accepted genetic model.

Under the null hypothesis of no association, Z has a standard normal distribution, and Z^2 follows an asymptotic chi square distribution with one degree of freedom.

Genetic models can be categorized as additive, dominant and recessive models. For the additive model, the genotype based trend test investigates the ordered relationship between all three genotype groups. However, for the dominant model, the test compares *aa* to *aA* and to *AA* together and for the recessive model, the test compares *aa* and *aA* jointly to *AA*. For the known mode of inheritance, the CA trend test weights are assigned as [0,1,2]; [0,1,1] and [0,0,1] respectively, and shown as in Table 2.

Table 2: Scores for genetic models

Model	*aa*	*aA*	*AA*	H_A
Additive	0	1	2	$f_0 \leq f_1 \leq f_2$
Dominant	0	1	1	$f_0 \leq f_1 = f_2$
Recessive	0	0	1	$f_0 = f_1 \leq f_2$

CA trend test is a function of the weights but it is unvarying under a linear conversion of the weights (Zheng et al. 2003). It is easy to assign these scores when the underlying genetic model is known. In genetic studies, the genetic model is rarely known and in such a case the weights can be assigned according to the suspected genetic model. The choice of weights in the test might have a substantial effect on the results. Misspecification of the scores would lead to a loss of test power. In this paper, three different weights are considered as a solution for the association testing when the mode of inheritance is unknown.

Longitudinal Blood Pressure Measurements

Blood pressure is a hereditary disease, which measures the pressure during beating and relaxing of heart. Diastolic blood pressure (DBP) is measured at the moment when the heart is relaxing, whereas systolic blood pressure (SBP) is measured when the heart is beating. Over one billion people worldwide have hypertension (SBP$\geq$140 mm Hg or DBP$\geq$90 mm Hg) (Kearney et al. 2005). It has been identified as a risk factor for cardiovascular events such as kidney failure and heart attack. For high blood pressure, a genetic component has been discovered by several genome wide association studies (The Wellcome Trust Case-Control Consortium 2007; Levy et al. 2009). The International Consortium for Blood Pressure Genome Wide Association Studies (2011) published 16 loci that have been associated with hypertension.

The present genotype-based association analysis considered two candidate markers that have been associated with blood pressure: in Chromosome 15, variant *rs1378942* in *CYP1A1-ULK3* gene (*hg19 position 75077367*), and in Chromosome 5, variant *rs1173771* in *NPR3-C5orf23* gene *(hg19 position 32815028)*. The p-values for the association are $1x10^{-8}$ and $3.2x10^{-10}$, respectively. For *rs1378942*, the odds ratio per risk allele is 1.075 and for *rs1173771*, the odds ratio per risk allele is 1.063.

Genetic Analysis Workshop 18 (GAW 18) data set contains both entire genome sequence genotype data and longitudinal phenotype data. Blood pressure measurements (DBP and SBP) were taken at four different time intervals. First measurement was taken between 1991 and 1996, the second between 1997 and 2000, the third between 1998 and 2006, and the last one between 2009 and 2011 (Almasy et al. 2014).

RESULTS

In this paper, DBP is used as an identifier of high blood pressure and hypertension (HTN) is defined as DBP$\geq$90. First, three measurements are analyzed due to the missing values in the last replication. There are 467 individuals with available genotype information for *CYP1A1-ULK3* gene and 445 individuals for *NPR3-C5orf23* gene. The genotype distributions for both specified variants are presented in Table 3 and Table 4.

Based on Table 3 and Table 4, the distributions of genotype levels differ from each other. *CYP1A1-ULK3* data has the largest amount of individuals with a homozygote genotype (GG) compared to *NPR3-C5orf23* data. For *NPR3-C5orf23* gene, heterozygous genotype level (CT) has 229 individuals.

CA trend test results are summarized for both variants in Table 5. The test statistics and the p-

Table 3: Variant rs1378942 in CYP1A1-ULK3 gene

Replication	*Status*	*GG*	*GT*	*TT*	*Total*
1	*Case*	31	37	8	76
	(Row %)	(40.8%)	(48.7%)	(10.5%)	(16.3%)
	(Column %)	(12.9%)	(20.2%)	(17.8%)	
	Control	208	146	37	391
	(Row %)	(53.2%)	(37.3%)	(9.5%)	(83.7%)
	(Column %)	(87.1%)	(79.8%)	(82.2%)	
2	*Case*	65	60	12	137
	(Row %)	(47.4%)	(43.8%)	(8.8%)	(29.3%)
	(Column %)	(27.2%)	(32.8%)	(26.7%)	
	Control	174	123	33	330
	(Row %)	(52.7%)	(37.3%)	(10.0%)	(70.7%)
	(Column %)	(72.8%)	(67.2%)	(73.3%)	
3	*Case*	87	71	17	175
	(Row %)	(49.7%)	(40.6%)	(9.7%)	(37.5%)
	(Column %)	(36.4%)	(38.8%)	(37.8%)	
	Control	152	112	28	292
	(Row %)	(52.1%)	(38.4%)	(9.5%)	(62.5%)
	(Column %)	(63.6%)	(61.2%)	(62.2%)	
	Total of Each Replication	239	183	45	467
	(Column %)	(51.2%)	(39.2%)	(9.6%)	(100.0%)

Table 4: Variant rs1173771 in NPR3-C5orf23 gene

Replication	*Status*	*CC*	*CT*	*TT*	*Total*
1	*Case*	19	39	15	73
	(Row %)	(26.0%)	(53.4%)	(20.6%)	(16.4%)
	(Column %)	(13.9%)	(17.0%)	(18.8%)	
	Control	117	190	65	372
	(Row %)	(31.4%)	(51.1%)	(17.5%)	(83.6%)
	(Column %)	(86.1%)	(83.0%)	(81.2%)	
2	*Case*	34	69	27	130
	(Row %)	(26.1%)	(53.1%)	(20.8%)	(29.3%)
	(Column %)	(25.0%)	(30.1%)	(33.8%)	
	Control	102	160	53	315
	(Row %)	(32.4%)	(50.8%)	(16.8%)	(70.7%)
	(Column %)	(75.0%)	(69.9%)	(66.2%)	
3	*Case*	42	92	31	165
	(Row %)	(25.4%)	(55.8%)	(18.8%)	(37.1%)
	(Column %)	(30.8%)	(40.2%)	(38.8%)	
	Control	94	137	49	280
	(Row %)	(33.6%)	(48.9%)	(17.5%)	(62.9%)
	(Column %)	(69.2%)	(59.8%)	(61.2%)	
	Total of Each Replication (Column %)	136 (30.6%)	229 (51.5%)	80 (17.9%)	445 (100.0%)

values are computed assuming three possible genetic models. In the computation process, the "coin" package is used from R software (Hothorn et al. 2008).

Based on Table 5, the association between the first variant *CYP1A1-ULK3* and HTN is statistically significant for the first replication. The results suggest that HTN follows a dominant model. However, for the sequent replications such a model does not fit the data. Depending on the determined scores, CA test indicates that having a *T* allele and HTN has a linear relationship, the probability of having HTN increases linearly with the presence of the *T* allele in the genotype. For the second variant *NPR3-C5orf23,* there is no significant association neither between three genetic models nor between replications. Due to the structure of the case-control design in GAW18, the existing association between HTN and *NPR3-C5orf23* may not be supported by the CA trend test. Pearson chi square test is also performed for the 2x3 contingency tables for testing the hypothesis of no association between genotype and HTN. The results for both genes are given in Table 6. Based on the results of Pearson chi square with two degree of freedom, no association is detected. Note that the Pearson chi square test for comparing the three genotype groups does not take into account the mode of inheritance, which is a crucial point in genetic case-control studies.

DISCUSSION

When it comes to comparing the results with other genetic case-control studies, there is a heterogeneous distribution between the levels of genotype in GAW data. As an alternative data

Table 5: One-sided CA trend test results

Marker	*Weights*	*Replication 1*		*Replication 2*		*Replication 3*	
		Chi-square	*p-val.*	*Chi-square*	*p-val.*	*Chi-square*	*p-val.*
CYP1A1-ULK3	Additive *(0,1,2)*	2.645	0.103	0.362	0.547	0.152	0.696
	Recessive *(0,0,1)*	0.082	0.774	0.170	0.679	0.002	0.964
	Dominant *(0,1,1)*	**3.912**	**0.048**	1.079	0.299	0.239	0.625
NPR3-C5orf23	Additive *(0,1,2)*	0.937	0.333	2.023	0.155	1.951	0.162
	Recessive *(0,0,1)*	0.390	0.532	0.968	0.325	0.116	0.733
	Dominant *(0,1,1)*	0.844	0.358	1.678	0.195	3.216	0.073

Table 6. Pearson Chi-square results

Marker	*Replication 1*		*Replication 2*		*Replication 3*	
	Chi-square	*p-val.*	*Chi-square*	*p-val.*	*Chi-square*	*p-val.*
CYP1A1-ULK3	4.078	0.130	1.733	0.420	0.255	0.879
NPR3-C5orf23	0.973	0.614	2.05	0.357	3.274	0.194

structure instance, Shahbazi et al. (2002) analyzed the associations between malignant melanoma phenotype and Epidermal Growth Factor (EGF) gene by using a case-control study design. For *EGF* gene, *G* allele is settled for risk allele and the *A* allele for alternative. The genotype total counts in the example data are 38, 55, and 30 for *AA*, *AG* and *GG*, respectively. The CA trend test is a sufficient method for testing the trend when the mode of inheritance is unknown in the presence of homogeneous distributed genotype levels rather than heterogeneous ones. Testing the trend of a heterogeneous genotype distribution is also considered in Lee (2016) by an optimal trend test over CA trend test, which reveals that in the presence of heterogeneous genetic effects, a more sensitive test is required.

CA trend test can be used to investigate the linear relationship between the probabilities of a disease for the genotype levels in genetic association studies. It is obvious that test scores have a substantial effect on the results. In this paper, the researchers consider the case where there is no exact information about the underlying genetic model of the disease. The researchers show how test results change depending on the assumed genetic model (additive, dominant or recessive) by analyzing the association between hypertension and two genetic markers *CYP1A1ULK3* and *NPR3-C5orf23* at three different time intervals. It is also shown that the association results may be influenced by the structure of the experimental design.

CONCLUSION

CA trend test, however, might cause deprivation of power when the inheritance model is unknown and unspecified. In the literature, a few methods have been proposed recently, for instance, the genetic model selection exclusion methods, which control the Type I error. In the presence of known inheritance model, the optimal CA trend test can be employed. On the other hand, the CA trend test is not robust when the scores are incorrectly determined. Even though there are several alternative tests to the CA trend test, the advantage of CA trend test is computational simplicity and flexibility.

REFERENCES

Almasy L, Dyer TD, Peralta JM, Jun G, Wood AR et al. 2014. Data for Genetic Analysis Workshop 18: Human whole genome sequence, blood pressure and simulated phenotypes in extended pedigrees. *BMC Proc,* 8(suppl 2): S2.

Armitage P 1955. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3): 375–386.

Buonaccorsi JP, Laake P, Veierod MB 2014. On the power of the Cochran-Armitage test for trend in the presence of misclassification. *Statistical Methods in Medical Research*, 23(3): 218-243.

Freidlin B, Zheng G, Li Z, Gastwirth JL 2002. Trend tests for case-control studies of genetic markers: Power, sample size and robustness. *Human Heredity*, 53: 146-152.

Hothorn LA, Hothorn T 2009. Order-restricted scores test for the evaluation of population-based case-control Studies when the genetic model is unknown. *Biom J*, 51: 659-669.

Hothorn T, Hornik K, van de Wiel MA, Zeileis A 2008. Implementing a class of permutation tests: The coin package. *Journal of Statistical Software*, 28(8): 1–23.

Kearney PM, Whelton M, Reynolds K, Munter P, Whelton PK et al. 2005. Global burden of hypertension: analysis of worldwide data. *Lancet*, 365: 217–223.

Lee, WC 2016. Optimal trend tests for genetic association studies of heterogeneous diseases. *Scientific Reports*, 6(27821): 1-7.

Levy D, Ehret GB, Rice K, Verwoert GC, Launer LJ 2009. Genome-wide association study of blood pressure and hypertension. *Nat Genet,* 41: 677-687.

Sasieni PD 1997. From genotypes to genes: Doubling the sample size. *Biometrics*, 53: 1253-1261.

Shahbazi M, Pravica P, Nasreen N, Fakhoury H, Fryer AA et al. 2002. Association between functional polymorphism in EGF gene and malignant melanoma. *Lancet*, 359: 397-401.

Slager SL, Schaid DJ 2001. Case-control studies of genetic markers: Power and sample size approxima-

tions for Armitage's test for trend. *Human Heredity*, 52(3): 149-153.

Talluri R, Wang J, Shete S 2014. Calculation of exact p-values when SNPs are tested using multiple genetic models. *BMC Genetics*, 15(75): 1-10.

The International Consortium for Blood Pressure Genome Wide Association Studies 2011. Genetic variants in a novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478: 103-105.

The Wellcome Trust Case-Control Consortium 2007. Genome-Wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447: 661-678.

Zheng G, Freidlin B, Li Z, Gastwirth JL 2003. Choice of scores in trend tests for case-control studies of candidate-gene associations. *Biometr J*, 45: 335-348.

Zheng G 2008. Analysis of ordered categorical data: Two score-independent approaches. *Biometrics*, 64: 1276-1279.

Paper received for publication on January 2016
Paper accepted for publication on May 2016