

Implications of the Codominance Model for Hardy-Weinberg Testing in Genetic Association Studies

Stefan Wellek^{a,b} Martina Müller-Nurasyid^{b,c} Konstantin Strauch^b

^aDepartment of Biostatistics, CIMH Mannheim, Mannheim Medical School of the University of Heidelberg, Mannheim, Germany; ^bInstitute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center, Johannes Gutenberg University, Mainz, Germany; ^cInstitute for Medical Information Processing, Biometry and Epidemiology, Faculty of Medicine, LMU Munich, Munich, Germany

Keywords

Equivalence test · Genome-wide association study · Goodness-of-fit test · Intersection-union test · Population stratification

Abstract

Introduction: The standard way of using tests for compatibility of genetic markers with the Hardy-Weinberg equilibrium (HWE) assumption as a means of quality control in genetic association studies (GAS) is to carry out this step of preliminary data analysis with the sample of non-diseased individuals only. We show that this strategy has no rational basis whenever the genotype-phenotype relation for a marker under consideration satisfies the assumption of codominance. **Methods:** The justification of this statement is obtained through rigorous analytical arguments. **Results:** The key result of the paper is that there holds the following proposition: under the codominance model, the genotype distribution of a diallelic marker is in HWE among the controls if and only if the same is true for the cases. **Conclusion:** The major practical consequence of that theoretical result is that under the codominance model, testing for HWE should be done both for cases and controls aiming to

establish the combined (intersection) hypothesis of compatibility of both underlying genotype distributions with the HWE assumption. A particularly useful procedure serving this purpose is obtained through applying the confidence-interval inclusion rule derived by Wellek, Goddard, and Ziegler (Biom J. 2010; 52:253–270) to both samples separately and combining these two tests by means of the intersection-union principle.

© 2024 The Author(s).
Published by S. Karger AG, Basel

1 Introduction

Testing the genotype frequencies observed for the SNPs involved for compatibility with Hardy-Weinberg equilibrium (HWE) condition is considered an important tool for the detection of technical issues and errors in genotyping. Accordingly, it has become one of the basic steps of quality control to be taken in the statistical analysis of any genetic association study (GAS), with genome-wide association study (GWAS) as a particularly important special case. Especially, in meta-analysis, one aims at controlling for deviations from HWE in the single studies to avoid systematic bias (see, e.g., Salanti

et al. [1]). When focusing on case-control GAS, various recommendations for appropriate testing can be found in the literature. Since random mating is a mandatory assumption, most of them emphasize the necessity of HWE testing within the control group. A hybrid approach tailored for mixed populations comprising both cases and controls has been recommended by Wang and Shete [2]: These authors argue that testing for compatibility with HWE has in principle to be done for the total population which might include a substantial proportion of individuals being affected by the disease when the prevalence of the latter is large. In such a setting, testing for HWE with unaffected controls alone would exclude a relevant proportion of the total population, yielding potentially biased estimates of the proportions under evaluation. Wang and Shete [3] also address other potential sources of systematic deviations from HWE in the control group alone which are not considered technical issues. In GWAS, current recommendations focus on more stringent testing within the control group and suggest non-rigorous testing in cases (Turner et al. [4]; Marees et al. [5]). The tool EasyQC (Winkler et al. [6]) has become popular for genome-wide meta-analysis. It is to be used for filtering HWE violations in controls only, as suggested previously by Anderson et al. [7].

An issue with respect to which the practice of genetic epidemiology exhibits remarkable uniformity is the numerical coding of genotypes as determined in terms of single SNPs. In the vast majority of GAS, this is done through counting the allele hypothesized to be a potential risk factor for the disease under consideration (the allele of minor frequency in most cases). The corresponding genetic model, which is often called “codominance model,” reflects the notion that its impact on the (binary) phenotype is, on the usual log-odds scale, representable as a point on the straight line joining the points corresponding to the two homozygous genotypes. Thereby, it differs from the classical dominant and recessive models under which the effect of the heterozygous genotype equals that of one of the two homozygous genotypes. As is well known, the assumption of codominance is a presupposition for both of the most commonly used tests for association between a SNP and the disease under study (Cochran-Armitage test, logistic regression analysis with the allele count as the covariable of interest; for a comparison of both methods, see Wellek and Ziegler [8]).

The aim of this contribution is to show that the frequently recommended strategy of confining HWE testing to the controls has no rational basis whenever the genotype-phenotype relation for a marker under con-

Table 1. Contingency table making up the frame of reference for the SNP-wise analysis of a genetic association study concerning a binary phenotype

Phenotype	Genotype			Σ
	aa	aA	AA	
Control	X_0	X_1	X_2	N_0
Case	Y_0	Y_1	Y_2	N_1
Σ	S_0	S_1	S_2	N

sideration satisfies the codominance model. This will be done through proving in a mathematically rigorous way the following logical equivalence relation: Under codominance, the distribution of a SNP-determined genotype is in HWE either in both underlying populations (i.e., cases and controls) or in none of them. In section 2, we start with introducing some notation and formulate the codominance model for an association study using cohort sampling. Afterward, we show how the model translates to the case-control setting which we are actually interested in. The main result is stated and proved in subsection 2.4 and section 3. In section 4, we outline and apply to some real examples a strategy of testing.

HWE compatibility in case-control studies on autosomal diallelic markers which is in line with the theoretical result stated in section 3 and also with the following crucial fact pointed out repeatedly in previous publications [9–12]: A statistical testing procedure which is really suited for justifying the inclusion of a SNP in the final evaluation of the respective marker as a risk factor for the disease under study should have the form of an equivalence test tailored for establishing goodness rather than lack of fit of a marker with HWE. In the Discussion, it is pointed out that a qualitative difference between the results of the two tests for goodness-of-fit with HWE to be performed for each SNP in a GAS can be taken as a reason against basing the test for possible association of the SNP with the disease status on the customary codominance model.

2 Materials and Methods

2.1 Notation and Models

The data which are used in the statistical analysis of a genetic association study devoted to a binary phenotype and a diallelic marker (SNP) are the values observed for or previously assigned to the random variables appearing in a 2×3 contingency table of the

Table 2. Cell probabilities associated with the contingency table shown as Table 1

Phenotype	Genotype			Σ
	aa	aA	AA	
Control	$\pi_{00} (1 - \tau)$	$\pi_{01} (1 - \tau)$	$\pi_{02} (1 - \tau)$	$1 - \tau$
Case	$\pi_{10}\tau$	$\pi_{11}\tau$	$\pi_{12}\tau$	τ
Σ	ξ_0	ξ_1	ξ_2	1.00

generic form shown in Table 1. In order to derive the theoretical result making up the core of this contribution, it is convenient to represent the probabilities corresponding to the different cells of this contingency table as functions of the following population parameters:

τ = probability that an individual randomly chosen from the total population is affected by the disease under study (disease prevalence);

π_{dj} = conditional probability of observing a genotype containing j A-alleles given the genotyped individual has been sampled from subpopulation $d \in \{0, 1\}$, with $d = 0$ and $d = 1$ indexing healthy and affected subjects, respectively, and j ranging over $\{0, 1, 2\}$;

ξ_j = frequency of genotype $j \in \{0, 1, 2\}$ in the total population.

The relationship between these parameters and the cell probabilities associated with the basic contingency table scheme displayed as Table 1 is made explicit in Table 2.

2.2 Codominance Model for Studies Using Cohort Sampling

By definition, under the cohort design, the following basic facts can be stated:

1. The values s_0, s_1, s_2 of the column totals S_0, S_1, S_2 are considered as the sizes of three subsamples resulting from genotyping. The triplet (s_0, s_1, s_2) not necessarily represents the genotype distribution in the total population.
2. The Y_j are mutually independent binomial random variables with parameters (s_j, δ_j) where δ_j denotes the penetrance of genotype j , i.e., the disease risk of individuals carrying j A-alleles ($j = 0, 1, 2$).
3. The codominance model assumes that, as measured on the log-odds scale, the increase in risk between “adjacent” genotypes remains constant. This is the case if and only if there exist constants α and β such that

$$\log\left(\frac{\delta_j}{1 - \delta_j}\right) = \alpha + \beta j \quad \forall j \in \{0, 1, 2\} \quad (1)$$

Equivalent to this assumption is that there exists some positive constant ρ such that there holds:

$$\frac{\delta_{j+1} / (1 - \delta_{j+1})}{\delta_j / (1 - \delta_j)} = \rho \quad \text{for } j = 0, 1 \quad (2)$$

where $\rho = e^\beta$.

2.3 Transforming the Codominance Model into a Model for Case-Control Studies

In a case-control study, the sample size made available for the two phenotypes has fixed values n_0 and n_1 , implying that (X_0, X_1, X_2) and (Y_0, Y_1, Y_2) are trinomially distributed random vectors being independent of each other. By the notational conventions introduced before (recall Table 2), the parameters of these multinomial distributions are given by $(\pi_{00}, \pi_{01}, \pi_{02})$ and $(\pi_{10}, \pi_{11}, \pi_{12})$ for (X_0, X_1, X_2) and (Y_0, Y_1, Y_2) , respectively. Since cases and controls are assumed to make up random samples from the subpopulations of affected and nonaffected individuals, respectively, the probabilities π_{dj} are left unchanged when instead of the total population, one considers the population underlying the case-control study. Recalling in addition the definitions of the parameters τ, ξ_j and δ_j , it is clear that for each $j \in \{0, 1, 2\}$, $\pi_{1j}\tau$, and $\delta_j\xi_j$ are different expressions for the same quantity, namely, the unconditional probability of finding in the total population a diseased individual carrying the j th genotype. Similarly, for unaffected individuals, we have $\pi_{0j}(1 - \tau) = P[J = j, D = 0] = (1 - \delta_j)\xi_j$, with J and D denoting the genotype and the disease status of an individual randomly selected from the total population (comprising both affected and nonaffected individuals). Accordingly, we can write:

$$\pi_{0j} = (1 - \delta_j)\xi_j / (1 - \tau), \quad \pi_{1j} = \delta_j\xi_j / \tau, \quad j = 0, 1, 2,$$

which, after a bit of elementary algebra, implies that there holds:

$$\frac{\pi_{11} / \pi_{10}}{\pi_{01} / \pi_{00}} = \frac{\delta_1 / (1 - \delta_1)}{\delta_0 / (1 - \delta_0)}, \quad \frac{\pi_{12} / \pi_{11}}{\pi_{02} / \pi_{01}} = \frac{\delta_2 / (1 - \delta_2)}{\delta_1 / (1 - \delta_1)} \quad (3)$$

Equations (2) and (3) and the fact that the π_{dj} are the same in the case-control study population as the total population justify the following conclusion: the codominance model, in its primary formulation referring to GAS using the cohort design, is satisfied if and only if the same holds true for the trinomial distributions to be analyzed in the corresponding case-control study.

2.4 Establishing Logical Equivalence between HWE Compatibility for Cases and Controls

A first step toward establishing our main result consists of parametrizing the model for a case-control association study in terms of ratios of genotype frequencies. Recalling that the trinomial parameters π_{dj} satisfy the boundary condition $\sum_{j=0}^2 \pi_{dj} = 1$ both for $d = 0$ and $d = 1$, elementary algebra suffices to verify that the following identities hold true:

$$\pi_{d0} = 1 / [1 + \eta_d (1 + \zeta_d)] \quad (4.a)$$

$$\pi_{d1} = \eta_d / [1 + \eta_d (1 + \zeta_d)] \quad (4.b)$$

$$\pi_{d2} = \eta_d \zeta_d / [1 + \eta_d (1 + \zeta_d)] \quad (4.c)$$

with $\eta_{d^*} = \pi_{d1} / \pi_{d0}$, $\zeta_{d^*} = \pi_{d2} / \pi_{d1}$.

Let us now denote by HWE_0 and HWE_1 the proposition stating that the Hardy-Weinberg condition is satisfied by the population of controls and cases, respectively, and assume that HWE_0 holds true. Using the symbol p_0 to denote the frequency of the risk-allele A in the subpopulation of controls, the validity of HWE_0 implies that the following equations hold true:

$$\eta_0 = \frac{\pi_{01}}{\pi_{00}} = 2p_0 \frac{(1-p_0)}{(1-p_0)^2} = \frac{2p_0}{(1-p_0)} \quad (5.a)$$

$$\zeta_0 = \frac{\pi_{02}}{\pi_{01}} = p_0^2 \frac{1}{2p_0(1-p_0)} = \frac{1}{2} \frac{p_0}{1-p_0} \quad (5.b)$$

By (2), (3), and the definition of η_d, ζ_d , (5.a, b) imply that

$$\eta_1 = 2\rho \frac{p_0}{1-p_0}, \quad \zeta_1 = \frac{\rho}{2} \frac{p_0}{1-p_0} \quad (6)$$

for some common factor $\rho > 0$. Plugging-in these expressions for η_1 and ζ_1 in formulas (4.a, b, c) yields:

$$\begin{aligned} \pi_{10} &= \frac{1}{1 + \left(\frac{2\rho p_0}{(1-p_0)}\right) \left(1 + \frac{\rho}{2} \frac{p_0}{(1-p_0)}\right)} \\ &= \frac{1}{1 + 2 \frac{\rho p_0}{1-p_0} + \left(\frac{\rho p_0}{1-p_0}\right)^2} \\ &= \left(1 + \frac{\rho p_0}{1-p_0}\right)^{-2} = \left(\frac{1-p_0}{1-p_0 + \rho p_0}\right)^2 = \left(1 - \frac{\rho p_0}{1-p_0 + \rho p_0}\right)^2; \\ \pi_{11} &= \frac{2\rho p_0 / (1-p_0)}{1 + (2\rho p_0 / (1-p_0)) \left(1 + \frac{\rho}{2} \frac{p_0}{(1-p_0)}\right)} = \frac{2\rho p_0 / (1-p_0)}{\left(1 + \frac{\rho p_0}{1-p_0}\right)^2} \\ &= \frac{2\rho p_0 (1-p_0)}{(1-p_0 + \rho p_0)^2} \\ &= 2 \cdot \frac{\rho p_0}{1-p_0 + \rho p_0} \cdot \left(1 - \frac{\rho p_0}{1-p_0 + \rho p_0}\right). \end{aligned}$$

The major conclusion from these expressions for the genotype frequencies π_{10} and π_{11} holding for the population of cases is as follows. If the parameter ($\pi_{00}, \pi_{01}, \pi_{02}$) of the genotype distribution among the controls satisfy the HWE condition with risk-allele frequency p_0 and the codominance model applies, then the population of cases is likewise in HWE with risk-allele frequency $p_1 = \frac{\rho p_0}{1-p_0 + \rho p_0}$.

The proof of the reverse implication, stating that under codominance, the validity of HWE_1 ensures that of proportion HWE_0 , is fully analogous. It turns out that for a given ($\pi_{10}, \pi_{11}, \pi_{12}$) satisfying HWE with risk-allele frequency p_1 , it follows that there holds for the following:

$$\begin{aligned} \pi_{00} &= \left(1 - \frac{\rho^{-1} p_1}{1-p_1 + \rho^{-1} p_1}\right)^2, \\ \pi_{01} &= 2 \frac{\rho^{-1} p_1}{1-p_1 + \rho^{-1} p_1} \left(1 - \frac{\rho^{-1} p_1}{1-p_1 + \rho^{-1} p_1}\right). \end{aligned}$$

3 Results

The arguments and considerations made explicit in section 2 yield a rigorous proof of the fact that under the codominance model, in a genetic association study, the subpopulations to be compared with respect to the dis-

tribution of a diallelic marker are either both in HWE or in HW disequilibrium. Precisely speaking, the following statement turns out to hold true:

In the codominance case, the distribution of the diallelic marker under study satisfies the HWE condition in the population of cases if and only if the same holds true for the population of controls. Using the short-hand notation introduced at the beginning of subsection 2.4, this means that in the codominance case, we can write:

$$HWE_0 \Leftrightarrow HWE_1.$$

Remark, an alternative proof of the above result would be to derive it from Sasieni's [13] Theorem 1.

4 Consequences for Preliminary HWE Testing in Genetic Association Studies

4.1 A Testing Strategy Taking the Result Derived in § 3 into Account

In the genetic epidemiology literature, the question of which of the two samples made available in a case-control association study should be checked by means of a preliminary statistical test for compatibility with HWE of the SNPs under evaluation has been a topic of controversial discussion (in addition to the references cited in the Introduction, see Wittke-Thompson et al. [14]; Ziegler [15], § 4.3.2). Essentially, the following two strategies are advocated:

- (I) Before testing a diallelic marker for association with the phenotype, it should be made sure that it exhibits neither among the controls nor the cases a relevant amount of HW disequilibrium.
- (II) Compatibility with HWE is a necessary requirement only for the population of controls. For cases, no test tailored for establishing the hypothesis of goodness-of-fit to HWE should be performed, since HW disequilibrium might indicate the existence of a true association being not just a spurious effect of population stratification or genotyping errors.

The result derived in this contribution admits the conclusion that strategy (II) is not really in line with the well-established practice of genetic epidemiology according to which genetic association studies are analyzed by fitting logistic regression models representing genotypes by single discrete covariates taking values 0, 1, or 2 rather than couples of binary dummy

Genotype	II	ID	DD	Σ
X_{0j} (Controls)	193	292	215	801 (n_0)
X_{1j} (Cases)	184	460	217	861 (n_1)

	$C_l(1 - \alpha)$	$\hat{\omega}$	$C_u(1 - \alpha)$
Controls	0.8587	0.9646	1.0837
Cases	1.0285	1.1510	1.2882

covariates. Our result implies that adopting this methodology and keeping it fully consistent requires that only SNPs which pass a preliminary test for goodness-of-fit with HWE both for controls and cases should be included in the confirmatory analysis of a GAS.

A recommendable, theoretically well-grounded strategy of testing the null hypothesis that a SNP exhibits relevant HW disequilibrium in at least one of the two subpopulations involved, against the alternative that the HWE condition is at least approximately satisfied both for cases and controls, proceeds in two steps.

Step 1: carrying out separately with each of the two samples a test at significance level α for goodness rather than lack of fit of the genotype distribution to the HWE model (for the theory and practical implementation of the exact version of this test, see [9, 10]). An asymptotic version based on confidence intervals which is particularly easy to implement, has been provided by Wellek, Goddard, and Ziegler [16]).

Step 2: acceptance of the SNP under consideration for evaluation as a potential genetic risk factor by means of logistic regression analysis if and only if each of the two individual tests performed in step 1 leads to rejecting the null hypothesis of relevant deviations from HWE.

4.2 Real Data Examples Illustrating the Use of the Proposed Strategy

In the following examples, the proposed two-step testing procedure is carried out in the version based on asymptotic confidence intervals for the parameter

$$\omega = \frac{\pi_2}{2\sqrt{\pi_1\pi_3}} \quad (8)$$

termed relative excess heterozygosity by Wellek, Goddard, and Ziegler [16]. In this formula, the π_j are generic symbols standing for the genotype frequencies in either of the two subpopulations, i.e., cases or controls. Following the proposal made in the same source, the equivalence margin for ω whose value under perfect HWE equals unity is set equal to 0.4, implying that the confidence interval has to be checked for inclusion in the acceptance interval $(1/(1 + 0.4), 1 + 0.4) = (0.714,$

$1.400)$. In other words, the null hypothesis of relevant deviations from HWE has to be rejected if and only if it turns out that there holds both $C_l(1 - \alpha) > 0.714$ and $C_u(1 - \alpha) < 1.400$, where $C_l(1 - \alpha)$ and $C_u(1 - \alpha)$ denote the lower and upper confidence bound to ω at 1-sided confidence level $1 - \alpha$, respectively. As usual, the latter is chosen to be 95% throughout. All necessary computations are performed using the SAS/IML script `ExactConfBound_RelHetZyg.sas` to be found in [16] as accompanying material.

1. An insertion/deletion polymorphism (INDEL) passing the preliminary goodness-of-fit test for both subpopulations.

The following data (absolute genotype frequencies; I: insertion, D: deletion) were obtained by Kee et al. [17] for a polymorphism of the angiotensin-converting enzyme (ACE) gene in an association study of risk factors for myocardial infarction:

Computing with these data the point estimates and asymptotic 95% confidence bounds for the parameter ω yields:

Obviously, both confidence intervals are contained in the interval $(0.714, 1.400)$ so that the test for goodness-of-fit to the HW model rejects the null hypothesis of relevant deviations from HWE both for the controls and the cases. Thus, the INDEL under investigation satisfies the proposed criterion for being included in a subsequent test for association with the disease status.

2. An SNP passing the preliminary goodness-of-fit test only for controls.

In an association study of genes predisposing to colorectal cancer, Shannon et al. [18] found the following absolute frequencies of genotypes of a SNP encoding in the methylenetetrahydrofolate reductase (MTHFR) enzyme:

With these data, the point estimates and asymptotic 95% confidence bounds for the parameter ω are computed to be:

The corresponding interval with endpoints $C_l(1 - \alpha)$ and $C_u(1 - \alpha)$ is included in the interval $(1/1.40, 1.40)$ for controls but not for cases.

3. A SNP passing the preliminary goodness-of-fit test only for cases.

The data for our next example are taken from a GAS of Alzheimer's disease published by Bhojak et al. [19]. In the

Genotype	TT	CT	CC	Σ
X_{0j} (Controls)	114	560	533	1,207 (n_0)
X_{1j} (Cases)	28	64	94	186 (n_1)

	$C_l(1 - \alpha)$	$\hat{\omega}$	$C_u(1 - \alpha)$
Controls	1.0179	1.1359	1.2676
Cases	0.4755	0.6237	0.8182

Genotype	GG	GC	CC	Σ
X_{0j} (Controls)	34	22	12	68 (n_0)
X_{1j} (Cases)	110	148	44	302 (n_1)

	$C_l(1 - \alpha)$	$\hat{\omega}$	$C_u(1 - \alpha)$
Controls	0.3485	0.5446	0.8510
Cases	0.8713	1.0637	1.2985

subgroup of APOE*4 carriers, these authors found the following absolute genotype frequencies of an IL-6 promoter polymorphism:

As point estimates and asymptotic 95% confidence bounds for ω , one obtains:

Clearly, only the second of the corresponding intervals ($C_l(1 - \alpha)$, $C_u(1 - \alpha)$) satisfies the condition ($C_l(1 - \alpha)$, $C_u(1 - \alpha)$) \subseteq (1/1.40, 1.40) so that approximate compatibility of the observed genotype distribution with HWE can be declared established at significance level $\alpha = 0.05$ only for cases.

4. A SNP passing the preliminary goodness-of-fit test for neither subpopulation.

Our final example is taken from a study of possible associations between polymorphisms of the cystatin C gene and Alzheimer's disease in the Japanese population (Maruyama et al. [20]). The absolute frequencies found in this study for the CST3+73 G/A genotype were as follows:

As point estimates and asymptotic 95% confidence bounds for the parameter ω , one obtains with these data:

Since there holds neither $(0.3551, 0.7823) \subseteq (1/1.40, 1.40)$ nor $(0.3404, 0.7747) \subseteq (1/1.40, 1.40)$, the equivalence test rejects HW disequilibrium for neither subpopulation.

Genotype	GG	GA	AA	Σ
X_{0j} (Controls)	180	40	8	228 (n_0)
X_{1j} (Cases)	137	34	8	179 (n_1)

	$C_l(1 - \alpha)$	$\hat{\omega}$	$C_u(1 - \alpha)$
Controls	0.3551	0.5270	0.7823
Cases	0.3404	0.5135	0.7747

5 Discussion

In two out of the four examples considered in §4.2, the statistical analysis yields results which according to our basic theoretical result are inconsistent with the codominance model, due to the fact that the significance statements to be made in the test for goodness-of-fit with HWE for cases and controls are qualitatively different. Whenever this occurs, relying on the codominance assumption in evaluating the association of the genetic marker under consideration with the disease status can hardly be justified, and the recommendation to apply the intersection-union procedure described before loses its rational basis. The problem of how to combine the information on possible deviations from HWE obtained for both subpopulations has to be reconsidered, then. Developing a strategy of HWE testing for settings involving SNPs which do not satisfy the codominance model might be a worthwhile topic for future research.

Irrespective of whether or not the codominance model holds true, the combined testing procedure described in §4.1 maintains the predefined level α of significance as is guaranteed by the so-called principle of intersection-union (cf. R.L. Berger [21]; Wellek [11], §7.1). By this general result, it can be taken for granted that the risk of declaring some given SNP as evaluable with respect to its contribution to the risk of presenting the disease under investigation, although it fails to be in (approximate) HWE in at least one of both subpopulations, keeps below α (e.g., $\alpha = 0.05$).

Researchers who do not see a problem in carrying out tests for lack of fit when establishing goodness of fit is obviously the aim, might be tempted to adopt the strategy outlined in section 4.1 with the traditional χ^2 instead of the equivalence test as building block. This would not affect the justification for applying the recommended rule for combining the evidence in favor or against HWE contained in the data from the two arms of a GAS but would entail the risk of failing to keep the type-1 error risk of the combined testing procedure, at the

target significance level. Actually, the risk of inferring approximate compatibility with HWE from a χ^2 -based p value being larger than the threshold $\alpha = 0.05$ might grossly deviate from α to both directions. An instance where it is lower than required for making the decision based on the lack-of-fit procedure consistent with that taken in the goodness-of-fit test is the distribution obtained for the sample of cases of Example 4.2 (i). With $X_{00} = 184$, $X_{01} = 460$, $X_{02} = 217$, the traditional χ^2 test for lack of fit yields a p value being as small as 0.0397, although the test for goodness-of-fit with the recommended equivalence interval (1/1.4, 1.4) decides in favor of approximate compatibility with HWE (at the same nominal significance level $\alpha = 0.05$). Nevertheless, as a test for equivalence, the traditional χ^2 -based procedure will in general be anti-conservative, in the sense that the risk of declaring compatibility with HWE, although the true value of ω fails to fall between the limits 1/1.4 and 1.4, exceeds α . As has been shown by Wellek, Goddard, and Ziegler [16], using the equivalence interval (1/1.4, 1.4) leads to reasonable power to conclude compatibility with HWE with sample sizes of the order of magnitude available in standard GWAS.

Statement of Ethics

An ethics statement is not applicable because this study is based exclusively on analytical considerations and published literature.

References

- Salanti G, Higgins JP, Trikalinos TA, Ioannidis JP. Bayesian meta-analysis and meta-regression for gene-disease associations and deviations from Hardy-Weinberg equilibrium. *Stat Med*. 2007;26(3):553–67. doi: [10.1002/sim.2575](https://doi.org/10.1002/sim.2575).
- Wang J, Shete S. Using both cases and controls for testing Hardy-Weinberg proportions in a genetic association study. *Hum Hered*. 2010;69:212–8.
- Wang J, Shete S. Testing departure from hardy-weinberg proportions. *Methods Mol Biol*. 2017;1666:83–115. doi: [10.1007/978-1-4939-7274-6_6](https://doi.org/10.1007/978-1-4939-7274-6_6).
- Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, et al. Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet*. 2011;68:1. doi: [10.1002/0471142905.hg0119s68](https://doi.org/10.1002/0471142905.hg0119s68).
- Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int J Methods Psychiatr Res*. 2018;27(2):e1608. doi: [10.1002/mpr.1608](https://doi.org/10.1002/mpr.1608).
- Winkler TW, Day FR, Croteau-Chonka DC, Wood AR, Locke AE, Mägi R, et al. Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc*. 2014;9(5):1192–212. doi: [10.1038/nprot.2014.071](https://doi.org/10.1038/nprot.2014.071).
- Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nat Protoc*. 2010;5(9):1564–73. doi: [10.1038/nprot.2010.116](https://doi.org/10.1038/nprot.2010.116).
- Wellek S, Ziegler A. Cochran-Armitage test versus logistic regression in the analysis of genetic association studies. *Hum Hered*. 2012;73(1):14–7. doi: [10.1159/000334085](https://doi.org/10.1159/000334085).
- Wellek S. Tests for establishing compatibility of an observed genotype distribution with Hardy-Weinberg equilibrium in the case of a biallelic locus. *Biometrics*. 2004;60(3):694–703. doi: [10.1111/j.0006-341X.2004.00219.x](https://doi.org/10.1111/j.0006-341X.2004.00219.x).
- Goddard KAB, Ziegler A, Wellek S. Adapting the logical basis of tests for Hardy-Weinberg equilibrium to the real needs of association studies in human and medical genetics. *Genet Epidemiol*. 2009;33(7):569–80. doi: [10.1002/gepi.20409](https://doi.org/10.1002/gepi.20409).
- Wellek S. Testing statistical hypotheses of equivalence and noninferiority. Second edition. Boca Raton: Chapman & Hall/CRC; 2010.
- Wellek S, Ziegler A. Testing for goodness rather than lack of Testing for goodness rather than lack of fit of an X-chromosomal SNP to the Hardy-Weinberg model of a X-chromosomal SNP to the Hardy-Weinberg model. *PLoS One*. 2019;14(2):e0212344. doi: [10.1371/journal.pone.0212344](https://doi.org/10.1371/journal.pone.0212344).
- Sasieni P. From genotypes to genes: doubling the sample size. *Biometrics*. 1997;53(4):1253–61. doi: [10.2307/2533494](https://doi.org/10.2307/2533494).
- Wittke-Thompson JK, Pluzhnikov A, Cox NJ. Rational inferences about departures from Hardy-Weinberg equilibrium. *Am J Hum Genet*. 2005;76(6):967–86. doi: [10.1086/430507](https://doi.org/10.1086/430507).
- Ziegler A, König IR. A statistical approach to genetic epidemiology: concepts and applications. Second edition Weinheim: Wiley-VCH; 2010.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

There are no funding sources to acknowledge.

Author Contributions

S.W. formulated the problem addressed in this manuscript, developed the theory leading to its solution, worked out the proofs, and drafted an initial version of the manuscript. M.M.-N. and K.S. elaborated on the review of the existing literature on Hardy-Weinberg testing in genetic association studies and made many valuable suggestions for clarifying the conclusions to be drawn from the presented theoretical result.

Data Availability Statement

The data analyzed in this paper are publicly available and do not contain any information about individual subjects. The sources from which the data are taken are the references listed in the bibliography as items [17–20]. Further inquiries can be directed to the corresponding author.

- 16 Wellek S, Goddard KAB, Ziegler A. A confidence-limit-based approach to the assessment of Hardy-Weinberg equilibrium: a confidence-limit-based approach to the assessment of Hardy-Weinberg equilibrium. *Biom J.* 2010;52(2):253–70. doi: [10.1002/bimj.200900249](https://doi.org/10.1002/bimj.200900249).
- 17 Kee F, Morrison C, Poirier O, McCrum E, Mallet C, Nicaud V, et al. Angiotensin II type-I receptor and ACE polymorphisms and risk of myocardial infarction in men and women. *Eur J Clin Invest.* 2000;30(12):1076–82. doi: [10.1046/j.1365-2362.2000.00741.x](https://doi.org/10.1046/j.1365-2362.2000.00741.x).
- 18 Shannon B, Gnanasampanthan S, Beilby J, Iacopetta B. A polymorphism in the methylenetetrahydrofolate reductase gene predisposes to colorectal cancers with microsatellite instability. *Gut.* 2002;50(4):520–4. doi: [10.1136/gut.50.4.520](https://doi.org/10.1136/gut.50.4.520).
- 19 Bhojak TJ, DeKosky ST, Ganguli M, Kamboh MI. Genetic polymorphisms in the catecholamine D and interleukin-6 genes and the risk of Alzheimer's disease. *Neurosci Lett.* 2000;288(1):21–4. doi: [10.1016/S0304-3940\(00\)01185-X](https://doi.org/10.1016/S0304-3940(00)01185-X).
- 20 Maruyama H, Izumi Y, Oda M, Torii T, Morino H, Toji H, et al. Lack of an association between cystatin C gene polymorphisms in Japanese patients with Alzheimer's disease. *Neurology.* 2001;57(2):337–9. doi: [10.1212/wnl.57.2.337](https://doi.org/10.1212/wnl.57.2.337).
- 21 Berger RL. Multiparameter hypothesis testing and acceptance sampling. *Technometrics.* 1982;24(4):295–300. doi: [10.2307/1267823](https://doi.org/10.2307/1267823).