NON-REPLICATION IN ASSOCIATION STUDIES: 'PSEUDO-FAILURES' TO REPLICATE?

P GORROOCHURN PhD[1], S E HODGE D Sc[1,2], G A HEIMAN PhD[3], M DURNER MD[1] & D A GREENBERG PhD[1,2]

[1]Division of Statistical Genetics, Department of Biostatistics, Mailman School of Public Health, Columbia University. [2]Clinical-Genetic Epidemiology Unit, New York State Psychiatric Institute, New York, NY. [3]Department of Epidemiology, Mailman School of Public Health, Columbia University.

Running Title: **Non-replication in association studies**

Key Words: **complex diseases, replication fallacy, replication probability, effect size**

*Correspondence: P Gorroochurn – Columbia University, Department of Biostatistics, Rm 620, 722 W 168th Street, New York, NY 10032, ph: (212) 342-1263, fax: (212) 342-0484, e-mail: pg2113@columbia.edu.

CONFLICT OF INTEREST NOTIFICATION PAGE

No conflict of interest

This work was supported in part by NIH grants DK-31813, NS27941, DK31775.

ABSTRACT

Recently, serious doubts have been cast on the usefulness of association studies as a means to genetically 'dissect' complex diseases, because most initial findings fail to replicate in subsequent studies. The reasons usually invoked are population stratification, genetic heterogeneity and inflated Type I errors. In this article, we argue that, even when these problems are addressed, the scientific community usually has unreasonably high expectations on replication success, based on initial low *p*-values – a phenomenon known as the 'replication fallacy.' We present a modified formula which gives the replication power of a second association study based on the *p*-value of an initial study. When both studies have similar sample sizes, this formula shows that: (i) A *p*-value only slightly less than the nominal α results in only about 50% replication power, (ii) very low *p*-values are required in order to achieve a replication power of least 80% (e.g., at α = .05, a *p*-value of less than .005 is required). Because many initially significant findings have low replication power, replication failure should not be surprising or be interpreted as necessarily refuting the initial findings. We refer to replication failures for which the replication power is low as 'pseudo-failures.'

INTRODUCTION

Despite much initial optimism, genetic association studies have been far from entirely successful. A decade ago, Risch and Merikangas (1) argued that association analysis could provide a more statistically powerful framework to 'dissect' complex diseases than linkage analysis under certain circumstances. This fact, compounded by an avalanche of SNPs which later steadily became available, triggered great enthusiasm in association studies. Several association studies have been performed and published in journals spanning a wide range of interdisciplinary interests. However, the results have been less than impressive: most of these studies seem to have failed one of the benchmarks of scientific success, namely replicability (2, p. 14). Consider this: a review paper by Hirschhorn et al. (3) found that, of more than 600 associations previously reported, 166 were studied at least thrice, and of these, only six were consistently replicated. Thus, the reality today seems to be one of skepticism, if not downright pessimism (4-9). What has gone wrong? Should we eventually abandon association studies? More precisely, is a 3.6% replication rate really bad? In this article, we will show that the concept of replication has been largely misunderstood, and that the genetics community has unintentionally put overly high replication expectations on initially significant findings (even when these are quite significant). In brief, we shall contend that a 3.6% replication rate, when looked at in the right statistical light, is really not surprising and should not give way to disillusionment.

MAPPING BY POPULATION ASSOCIATION

A genetic association exists between a phenotype (e.g. disease) and a specific allele at a genetic marker if the allele occurs more often (than would be expected by chance) in a group of individuals with the disease (cases) compared to a group without the disease (controls). Whereas linkage analysis is concerned with the co-segregation of marker loci with disease *within families*, association analysis looks at the dependence between marker alleles and disease *at a population level* (see, for example, Hodge (10)).

For association mapping by linkage disequilibrium to be successful, two conditions are desirable (11). First, the disease allele must have arisen only once in the population so that there is complete linkage disequilibrium between the marker and disease alleles (hence small isolated populations are often preferred). Second, the marker and disease loci must be in very close physical proximity, so that the disequilibrium between the marker and disease alleles is due to tight linkage (typically the recombination fraction $\theta \approx 10^{-5}$). Problems in linkage disequilibrium mapping arise when the disequilibrium in question is instead due to population history, natural selection or population stratification. In all these cases, an association between marker allele and disease can be observed, but the disease locus is not necessarily close to the marker locus (in the cases of population history or selection) or is even nonexistent (in the case of population stratification).

Population stratification can be especially deleterious to association studies (e.g. 12;13), and much effort has recently been channeled to further understand and correct for this confounder (14-20). If population stratification was to be corrected for and other design issues were to be improved in future association studies, should we expect a

sudden leap in successful replication rates? We will explain why such an expectation will likely remain unfulfilled in years to come. The main problem lies in a common misunderstanding of the meaning of a replication and in how likely it is to occur. Furthermore, we will show that many of the apparent failures to replicate have in fact been 'pseudo-failures.' We now examine these critical issues.

REPLICATION FALLACY, REPLICATION PROBABILITY, AND *P*-VALUES

We shall use 'replication' to refer to a situation where 'a null hypothesis that has been rejected at time 1 is rejected again, and with the same direction of outcome, on the basis of a new study at time 2' (21, p. 542; see also Table 1 below).

insert Table 1 here

Now, suppose a test of association is rejected with $p = .01$: what is the probability the study will be replicated? Oakes (22, pp. 79-80) reports that, in a survey of 70 academic psychologists each with at least 2 years' research experience, 60% endorsed the statement that if an initial study is significant with $p = .01$, then 'You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great many times, you would obtain a significant result on 99% of occasions.' Nothing could be further from the truth. In point of fact, as we will soon show, when $\alpha = .05$, an initial study with $p = .05$ implies a replication probability of only about .5, one with $p = .02$ implies only about .64 replication probability, and one with $p = .01$ results in no more than about .73 probability of replication! These results: (i) hold irrespective of the sample

size used as long as the sample sizes in the initial and subsequent studies are equal (the case of unequal sample sizes will be dealt with later), (ii) assume a *z*-test (we discuss deviations from this assumption below), and (iii) are based on the concept of *power,* as we shall soon explain. The incorrect belief that $1 - p$ is the probability of replication of an initial study is known as the *replication fallacy* (23;24). The replication fallacy is a reason why the scientific community generally has overly high expectations that an initial study which yields a relatively low *p*-value should be replicated, and if that does not happen, then surely the initial finding must have been false. Nothing, again, could be further from the truth. The whole issue is succinctly described by Oakes (22, p. 18):

'We have seen that the power of a replication of an independent means *t*-test design when the first experiment has an associated probability of .02 is approximately .67…Suppose then, psychologist B, suspecting that A's results were an artefact…decided to perform an exact replication of A's study. Suppose B's results were in the same direction as A's but were not significant. **It would be folly surely for B to assert that A's findings were indeed artefactual.' [**Bolding is our own]

It is precisely this kind of apparent replication failure that we shall describe as a 'pseudo-failure,' because the probability of a successful replication is a priori quite low and a subsequent 'failure' ought to come neither as a surprise nor as a contradiction of the initial positive finding.

If the *p*-value is not a direct measure of the replicability of an initial study, is there anything at all it can tell about the likelihood of replication? The short answer is yes. Although neither the exact *p*-value nor its complement can be interpreted as the probability of a successful replication, it has been shown that, for a given test size α, 'the

replicability of null hypothesis rejection is a continuous, increasing function of the complement of its *p* value' (25, p. 181). Although this fact about the *p*-value has been well-known in the social and behavioral sciences (e.g. 22;24;25;26;27), it seems to have been overlooked by the genetics community, nor have its implications been explored as we do here.

Let us now consider an initial association study based on a simple chi-square test (see Table 2) which yields a *p*-value p_1 for a test size α, where $p_1 \leq \alpha$. Then, it can be shown that (see Appendix) the initial finding, assuming it is not a false-positive, will be replicated with probability

$$p_{\text{REP}} = \Phi\left\{\frac{z_1\sqrt{n_2/n_1} - z_{crit}}{1 - z_1^{\,2}/(2n_1)}\right\}, \qquad (1)$$

Insert Table 2 here

where n_1, n_2 are the sample sizes (i.e. the number of cases or controls) in the first and second studies, respectively, $z_{crit} = \Phi^{-1}(1-\alpha/2)$, $z_1 = \Phi^{-1}(1-p_1/2)$, and Φ, Φ^{-1} are the standard normal distribution function and inverse distribution function, respectively. Note that, in accordance with our definition of a replication, Eq. (1) will be used only when: (i) the initial test is significant (i.e. $p_1 \leq \alpha$), and (ii) the outcomes of both tests are in the same direction (i.e. $a_1 - c_1$ and $a_2 - c_2$ in Table 2 have the same signs). Moreover, although Eq. (1) assumes an equal number of cases and controls in the first study, and similarly equal numbers in the second study, situations with different numbers of cases and controls can be dealt through the use of *harmonic means*. More specifically, any study containing *r* cases and *s* controls (where $r \neq s$) can be treated as one with the same number of cases and controls *t,* where $t = 2rs/(r+s)$ (28, p. 286). Note that Eq. (1)

implicitly assumes the same test size α for both the first and subsequent studies: for unequal test sizes, say α* and α, respectively, the same equation can be used as long as $p_1 \leq \alpha^*$. Finally, we also note that the whole issue of replicability can also be approached from a Bayesian perspective (e.g. (29)).

The critical assumption made in Eq. (1) is that the expected effect size in the second study equals the observed effect size in the initial study (Greenwald et al. (25)). Therefore the replication probability calculated is, in effect, the conditional *power* of the second study to be statistically significant given this assumption about effect size (hence 'replication probability' and 'replication power' will be used interchangeably from now on). Regarding this assumption, Greenwald et al. (25, p. 180) state that it 'involves a step of inductive reasoning that is (a) well recognized to lack rigorous logical foundation but is (b) nevertheless essential to ordinary scientific activity.' Of more pertinence is the fact that, when calculating replication probabilities, the literature has usually assumed the second study is an *exact* repetition of the first study (e.g. 25;27), in the sense that both studies: (i) have the *same* sample size, and (ii) are performed using the *same* population. Clearly, this is the norm neither in association studies nor elsewhere, especially the second condition. Regarding the first condition, we point out that all replication calculations are based on the assumption of constancy of effect size, as we explained at the start of the paragraph. Now, since effect size is unaffected by sample size (24, p. 98), there is enough justification to use the effect size in a study with a different sample size. However, it is very desirable for the first study to have a reasonably large sample size so that the standard error of its observed effect size is relatively small. For examples on the use of different sample sizes in replication calculations, based on initial effect sizes, see

Tversky and Kahneman (30) and Heils (31). Regarding the second condition, Tan et al. (32, p. 1435) point out that: 'Estimates of effect size will tend to regress to the true effect size in subsequent [association] studies, which is usually less extreme.' The same feeling has been echoed elsewhere (7;8;33), and Göring et al. (5) have provided a theoretical justification. Thus, the expected effect size of the second study will, on average, be smaller than the observed effect size of the first study. What this means is that, even when a different population is used for the subsequent study, Eq. (1) can still be used, with the understanding that the replication probabilities calculated from that equation actually represent *upper bounds* for the true replication probabilities (25).

Let us now examine three of the more unexpected consequences of Eq. (1), when both the first and second studies have similar sample sizes (i.e. $n_2 \approx n_1$):

CONSEQUENCE 1: A *p*-value only slightly less than the nominal α in the first study (e.g. $p = .04$ at $\alpha = .05$) results in a replication power of only about 50% for the second study (see Fig. 1).

CONSEQUENCE 2: Reasonably low *p*-values in the first study do not necessarily result in high replication power of the second study (e.g. $p = .02$ at $\alpha = .05$ implies a replication power of no more than 64%) (see Fig. 1).

CONSEQUENCE 3: To achieve a replication power of 80% for the second study at $\alpha = .05$, a *p*-value of at most .005 must be obtained in the first study (see Fig. 1).

When n_1 and n_2 are allowed to be different, two more consequences are:

CONSEQUENCE 4: For reasonably large sample sizes, the replication power depends only on the *ratio* of the initial and subsequent sample sizes, not on their absolute values

(e.g. if p = .02 at α = .05, then $p_{\mathrm{REP}} = .727$ when n_1 =100 and n_2 = 120, and $p_{\mathrm{REP}} = .723$ when n_1 =500 and n_2 = 600).

CONSEQUENCE 5: For a given sample size of the first study, if the initial *p*-value is only slightly less than the nominal α, the sample size required for the second study must be much larger in order to achieve a replication power of 80%. Suppose an initial association study with n_1 cases and n_1 controls yields a *p*-value p_1. Then, the number of cases (or controls) required for the second study in order to achieve a replication power of p_{REP} follows directly from Eq. (1),

$$n_2 \approx n_1 \left\{ \frac{z_{crit} + \Phi^{-1}(p_{\mathrm{REP}})}{z_1} \right\}^2 , \qquad (2)$$

where, as before, $z_{crit} = \Phi^{-1}(1-\alpha/2)$, $z_1 = \Phi^{-1}(1-p_1/2)$, Φ^{-1} is the inverse standard normal distribution function, and we have assumed $z_1^2 << 2n_1$. Fig. 2 illustrates the variation of n_2/n_1 with p_1. For example, if the initial study has p = .05 at $\alpha = .05$, then the sample size of the second study must be about 1.86 times that of the first study for a replication power of 80%. (Since the expected effect size of the second study will, on average, be smaller than the observed effect size of the first study, as we explained earlier, the actual required relative sample size will be *greater* than 1.86.)

Insert Fig. 1 here

Insert Fig. 2 here

IMPLICATIONS FOR ASSOCIATION STUDIES AND DISCUSSION

Initial association studies that achieve borderline significance, and even those which result in relatively low *p*-values, result in low replication probability (or power) for a subsequent study. Therefore, failure to replicate such an initial finding in a subsequent study should neither come as a surprise nor be deemed 'troubling'; nor is this failure to replicate necessarily an outright refutation of the initial finding.

So what should one do with an initial study that yields a *p*-value of, say, .03? We will leave it to the reader to decide, but consider this: if roughly the same sample size is used for both the initial and subsequent studies, the replication power is less than 58%, and no clinical trial with a power of 58% would ever pass a review board. If one insists on replicating, one should do so with the understanding that there are big chances of not succeeding. There is no denying that consistent replication and subsequent biological confirmation should be the gold standard. However, replication of some initial findings might just not be within the realm of high probability.

On the other hand, we have seen that an association study which yields a *p*-value of .005 at $\alpha = .05$ implies that a subsequent study will have almost 80% probability of replication (assuming the expected effect size of the second study is the observed effect size of the first). Does this mean that *any time* one obtains an initial *p*-value of .005, one should expect 80% of future studies to indeed result in a confirmation of the initial finding? To answer this question, one must remember that the replication power in Eq. (1) is calculated on the assumption that the effect size observed in the initial study is the population effect size in the subsequent study. If this is indeed the case, one would obtain 80% replication rate in future findings. (Since the original effect size will more likely be

overestimated, the actual success rate will be slightly less than 80%, as explained previously). However, if the initial finding is a false-positive, which will occur at a rate of 5%, the replication rate will be much less than 80% (29).

The above point leads us to a vital consideration before Eq. (1) can be applied. It is extremely important for the association study to be well-designed so that sources of bias are either minimal or have been removed. Although this will not completely eliminate false-positives, it will keep them as low as possible. Otherwise, it will be less likely for the expected effect size in the subsequent study to be even close to the observed effect size of the first study. For example, suppose an investigator reports a *p*-value of .005, but the design used suffers from considerable population stratification. Then, this significant finding will more likely be a false-positive, the observed effect size will very likely be a gross overestimate of the actual effect size, and application of Eq. (1) will lead to exaggerated confidence in the replication power of a subsequent study (i.e. the true replication power will be much less than 80%).

A critical issue for genome screen association studies concerns corrections for multiple testing. The question then arises as to how to compute the replication power after observing one or more positive results in a genome screen. Which test size α should one use, and which *p*-value? The answer may vary, depending on the situation. The simplest case is one in which a single association peak from a genome scan is chosen to study for replication, i.e., at a locus at which a few apparently highly associated SNPs are tested for replication. It would then be misleading to use the test with the smallest *p*-value, while still assuming a test size α, to calculate the replication power. Since a genome-wide association study could potentially consist of an extremely large number of

tests, an FDR correction procedure (34) is appealing. The appropriate *p*-values then to use in the computation of replication power are those that are significant with the FDR procedure. In the special case when only a few tests are performed and a Bonferroni adjustment of the test size α is opted for, a *p*-value which is significant at α/*C* could be used to compute the replication power for the associated test.

In this paper, we have focused on the replication power for a second study based on the *p*-value of *only one first study*. This is mainly because, in many cases, only one or two replications have been attempted, as can be seen for example from the survey conducted by Hirschchorn (3). However, it also legitimate to ask how Eq. (1) could be used if we wished to compute the replication power based on the results of *several initial studies*, i.e. based on a meta-analytic approach (21;35). In this case, we propose computing the average of the effect sizes ($\overline{\phi_1}$) of the initial studies (see Appendix) and estimating the initial effective sample size (n_h) from the harmonic mean of the initial sample sizes. The value of z_1 can then be calculated by using $z_1 = \overline{\phi_1}\sqrt{2n_h}$, and Eq. (1) can be used to calculate the replication power of the second study, with $n_1 = n_h$. However, there are three points to note with such an approach: (a) *all* initial studies (whether significant or not) must be used in the calculation of z_1, (b) the value of z_1 must at least as large as z_{crit}, and (c) sources of bias, e.g. population stratification, in the initial studies must be corrected before $\overline{\phi_1}$ is calculated; otherwise the latter will be overestimated.

How robust is Eq. (1) to distributional and sample size assumptions? To be sure, it is important for the sample sizes of the second and especially the first studies to be

reasonably large for at least two reasons: (i) so that chi-square tests can be legitimately used, (ii) so that the effect size in the first experiment can be consistently estimated. For small sample sizes when testing for two proportions, replication power can still be calculated from a noncentral hypergeometric distribution, but will not be accurate. When comparing two means using the *t*-test, formulas for the replication power have been given by Greenwald et al. (25) and Posavac (27). Finally, Consequence 1 (i.e. $p_{\mathrm{REP}} \approx .5$ when an initial study has *p*-value only slightly less than α) always holds irrespective of the underlying distribution, as long as the latter is symmetric.

The rationale for assuming that an expected value of the effect size in a second study is equal to the observed value of the statistic in a first study, which is at the basis of Eq. (1), can be legitimately questioned. Referring more specifically to the odds ratio, Fleiss et al. (36, p.67) point out that, whereas differences in proportions would vary between studies, measures of effect size could remain constant from study to study. Murphy and Myors (37, p. 12) further state that the effect size '…provides a simple metric that allows for comparison of treatment effects from different studies, areas or research…' In fact, as we mentioned previously, it is more justifiable to assume that the expected effect size in the second study is, on average, slight less than the observed effect size in the initial study. Consequently, Eq. (1) really gives the replication power for the *best-case scenario*, so that the actual replication power of the second study will be slightly less than that given by Eq. (1).

In spite of the various deficiencies of the *p*-value, which have been discussed at great length elsewhere (26;38-40), we believe that, in addition to measures of effect size and confidence intervals, researchers should continue to report *p*-values 'because tests of

statistical significance provide information that effect sizes and confidence intervals do not' (27, p. 101). We advocate, as does Greenwald et al. (25), the reporting of exact *p*-values, rather than expressions such as $p < .01$ or $p > .05$. Whenever a subsequent study is planned, we also believe researchers should compute replication power, based on the initial *p*-value and on the sample sizes of the initial and subsequent studies.

The literature on genetic association studies is rife with admonitions and possible explanations for their non-replications. The reasons are usually one or more of the following (6;8;9): (i) population stratification, (ii) genetic heterogeneity, (iii) inflation in Type I error, and (iv) gene-environment interaction. We do not deny that these are important problems, and attempts should be made to correct for them. But even if these problems were to be remedied, trying to replicate many initial findings, even if they are quite significant, may be predisposed to failure and should not be interpreted as necessarily contradicting the initial association. To our knowledge, only one publication in the genetic-epidemiology literature (5) has acknowledged this fact. Moreover, only one publication (31) on association studies has actually reported calculations for the power of a replication (based on the initial observed effect size), but even that power calculation seems to be slightly inflated.

When we looked at five of the six studies surveyed by Hirschhorn et al. (3), for which *p*-values could be obtained and which had been consistently replicated, we found that all of them had $p < 10^{-3}$. Thus, subsequent studies of these with similar sample sizes had about 99.8% replication power! We also randomly sampled 50 of the 166 initial studies that had reported exact *p*-values. We found that: (i) 78% of them had *p*-values greater than .005, implying a subsequent study of similar sample size would be

underpowered (a replication power of less than 80%), (ii) 38% of them had *p*-values greater or equal to .02, implying a subsequent study of similar sample size would be seriously underpowered (a replication power of less than 64%). While it is true that many of the replication failures reported by Hirschhorn et al. (3) could be due to the several reasons mentioned by these authors (e.g. population stratification, heterogeneity, etc.), it is equally true that, even for the remaining cases where these sources of error were actually minimal, replication failures were *bound* to occur, owing to their low a priori replication power. Furthermore, even if sources of bias had actually been minimal in *most* cases, replication success would still be low. On a second look, therefore, Hirschhorn et al.'s (3) review should perhaps not look so depressing, since it contains many potential 'pseudo-failures,' that is, replications that were just not meant to be. We must stress, however, that our paper in no way attempts to challenge Hirschhorn et al.'s arguments. It acknowledges these, but it also adds a whole new perspective to the issue of replication: old problems still need to be grappled with, because they reduce the odds of false-positives, *but even if they are addressed low replication power will continue to deny high replication rates in future association studies.*

In conclusion, if the *p*-value in one's initial study is *very* small (e.g. $p = .005$ when $\alpha = .05$), then one can indeed anticipate a high replication probability. However, for more commonly observed *p*-values (e.g. $p = .02$ and even $p = .01$, when $\alpha = .05$), replication probabilities are notably lower than one might hope. In these situations, one should not be surprised by a 'failure to replicate.'

ACKNOWLEDGMENTS

This work was supported in part by NIH grants R01 AA013654, DK-31813, NS27941, DK3177. We wish to thank Dr Bruce Levin for helpful discussions and Dana Politis for her invaluable time. We also thank two referees whose comments greatly improved the paper.

APPENDIX: Proof of Eq. (1)

Let p_1, p_2 be the two population proportions being compared. Let $P_1^{(i)}, P_2^{(i)}$ be the corresponding sample proportions (based on n_i cases and n_i controls) in the i^{th} study (i = 1, 2). The chi-square test statistic, X_i^2, for the i^{th} study (see Table 2) can be converted into a z-test, Z_i, where $Z_i = \text{sign}(P_1^{(i)} - P_2^{(i)}) \times \sqrt{X_i^2}$. The observed effect size for the i^{th} study is $\hat{\phi}_i = \sqrt{X_i^2/(2n_i)} = Z_i/\sqrt{2n_1}$ (41). Let the population effect size for the i^{th} study be ϕ_i. If the first test has size α, then its critical value is $z_{crit} = \Phi^{-1}(1-\alpha/2)$, where Φ is the standard normal cumulative distribution function. Also, let the observed value of Z_1 be z_1.Under $H_0 : p_1 = p_2$, for the first study, $Z_1 \sim N(0,1)$. Conditional on $\phi_2 = \hat{\phi}_1$ and on n_1, n_2, $Z_2 \sim N(\mu_2, {\sigma_2}^2)$, but $\mu_2 \neq 0$ and ${\sigma_2}^2 \neq 1$. Indeed, using $\hat{\phi}_i = Z_i/\sqrt{2n_i}$, we have

$$\begin{aligned}\mu_2 &= \text{E}(Z_2 \mid \phi_2 = \hat{\phi}_1, n_1, n_2) \\ &= \phi_2\sqrt{2n_2} \\ &= \hat{\phi}_1\sqrt{2n_2} \\ &= z_1\sqrt{\tfrac{n_2}{n_1}},\end{aligned}$$

and, using $\text{var}\,\hat{\phi}_i = (1-\phi_i^2)^2/(2n_i - 2)$ (42, p. 238) and assuming n_2 is reasonably large, we have

$$
\begin{aligned}
\sigma_2^{\,2} &= \mathrm{var}(Z_2 \mid \phi_2 = \hat{\phi}_1, n_1, n_2) \\
&= 2n_2 \,\mathrm{var}\,\hat{\phi}_2 \\
&= 2n_2 \tfrac{(1-\phi_2^{\,2})^2}{2n_2 - 2} \\
&\approx (1-\phi_2^{\,2})^2 \\
&= (1-\hat{\phi}_1^{\,2})^2 \\
&= (1-\tfrac{z_1^{\,2}}{2n_1})^2.
\end{aligned}
$$

Therefore, the replication probability is

$$
\begin{aligned}
p_{\mathrm{REP}} &= \Pr\left\{ Z_2 \geq z_{crit} \mid Z_2 \sim N[z_1\sqrt{\tfrac{n_2}{n_1}}, (1-\tfrac{z_1^{\,2}}{2n_1})^2] \right\} \\
&= \Pr\left\{ Z \geq \tfrac{z_{crit} - z_1\sqrt{n_2/n_1}}{1-z_1^{\,2}/(2n_1)} \right\}, \quad \text{where } Z \sim N(0,1), \\
&= \Phi\left\{ \frac{z_1\sqrt{n_2/n_1} - z_{crit}}{1-z_1^{\,2}/(2n_1)} \right\}.
\end{aligned}
$$

Reference List

(1) Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996 September 13;273(5281):1516-7.

(2) Runyon RP, Haber A, Pittenger DJ, Coleman KA. Fundamentals of Behavioral Statistics. ed 8. ed. New York: McGraw-Hill College; 1995.

(3) Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. Genet Med 2002 March;4(2):45-61.

(4) Vieland VJ. The replication requirement. Nat Genet 2001 November;29(3):244-5.

(5) Göring HH, Terwilliger JD, Blangero J. Large upward bias in estimation of locus-specific effects from genomewide scans. Am J Hum Genet 2001 December;69(6):1357-69.

(6) Hirschhorn JN, Altshuler D. Once and again-issues surrounding replication in genetic association studies. J Clin Endocrinol Metab 2002 October;87(10):4438-41.

(7) Dahlman I, Eaves IA, Kosoy R, Morrison VA, Heward J, Gough SC, Allahabadia A, Franklyn JA, Tuomilehto J, Tuomilehto-Wolf E, Cucca F, Guja C, Ionescu-Tirgoviste C, Stevens H, Carr P, Nutland S, McKinney P, Shield JP, Wang W, Cordell HJ, Walker N, Todd JA, Concannon P. Parameters for reliable results in genetic association studies in common disease. Nat Genet 2002 February;30(2):149-50.

(8) Colhoun HM, McKeigue PM, Davey SG. Problems of reporting genetic associations with complex outcomes. Lancet 2003 March 8;361(9360):865-72.

(9) Ott J. Association of genetic loci: Replication or not, that is the question. Neurology 2004 September 28;63(6):955-8.

(10) Hodge SE. Linkage analysis versus association analysis: distinguishing between two models that explain disease-marker associations. Am J Hum Genet 1993 August;53(2):367-84.

(11) Halliburton R. Introduction to Population Genetics. New Jersey: Prentice Hall; 2003.

(12) Schork NJ, Fallin D, Thiel B, Xu X, Broeckel U, Jacob HJ, Cohen D. The future of genetic case-control studies. In: Rao DC, Province MA, editors. Genetic Dissection of Complex Traits.CA: Academic Press; 2001. p. 191-212.

(13) Rosenberg NA, Nordborg M. A general population-genetic model for the production by population structure of spurious genotype-phenotype associations in discrete, admixed or spatially distributed populations. Genetics 2006 July;173(3):1665-78.

(14) Devlin B, Roeder K. Genomic control for association studies. Biometrics 1999 December;55(4):997-1004.

(15) Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 2001 November;60(3):155-66.

(16) Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000 June;155(2):945-59.

(17) Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. Am J Hum Genet 2000 July;67(1):170-81.

(18) Heiman GA, Hodge SE, Gorroochurn P, Zhang J, Greenberg DA. Effect of population stratification on case-control association studies. I. Elevation in false positive rates and comparison to confounding risk ratios (a simulation study). Hum Hered 2004;58(1):30-9.

(19) Gorroochurn P, Hodge SE, Heiman G, Greenberg DA. Effect of population stratification on case-control association studies. II. False-positive rates and their limiting behavior as number of subpopulations increases. Hum Hered 2004;58(1):40-8.

(20) Gorroochurn P, Heiman GA, Hodge SE, Greenberg DA. Centralizing the non-central chi-square: a new method to correct for population stratification in genetic case-control association studies. Genet Epidemiol 2006 February 24;30(4):277-89.

(21) Rosenthal R. Cumulating Evidence. In: Keren G, Lewis C, editors. A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues.Hillsdale,N.J.: Lawrence Erlbaum Associates; 1993. p. 519-59.

(22) Oakes M. Statistical inference: A commentary for the social and behavioral sciences. New York: Wiley; 1986.

(23) Gigerenzer G. The superego, the ego, and the id in statistical reasoning. In: Keren G, Lewis C, editors. A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues.Hillsdale,N.J.: Lawrence Erlbaum Associates; 1993. p. 311-39.

(24) Maxwell SE, Delaney HD. Designing Experiments and Analyzing Data: A Model Comparison Perspective. 2nd ed. Mahwah, NJ: Lawrence Erlbaum; 2004.

(25) Greenwald AG, Gonzalez R, Harris RJ, Guthrie D. Effect sizes and p values: what should be reported and what should be replicated? Psychophysiology 1996 March;33(2):175-83.

(26) Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. Psychol Methods 2000 June;5(2):241-301.

(27) Posavac EJ. Using p values to estimate the probability of a statistically significant replication. Understanding Statistics 2002;1(2):101-12.

(28) Daly LE, Bourke GJ. Interpretation and Uses of Medical Statistics. 5th ed. Oxford: Blackwell Science Ltd; 2000.

(29) Sohn D. Statistical Significance and Replicability: Why the Former Does not Presage the Latter. Theory Psychology 1998 June 1;8(3):291-311.

(30) Tversky A, Kahneman D. Belief in the law of small numbers. In: Keren G, Lewis C, editors. A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues.Hillsdale,N.J.: Lawrence Erlbaum Associates; 1993. p. 341-9.

(31) Heils A, Haug K, Kunz WS, Fernandez G, Horvath S, Rebstock J, Propping P, Elger CE. Interleukin-1beta gene polymorphism and susceptibility to temporal lobe epilepsy with hippocampal sclerosis. Ann Neurol 2000 December;48(6):948-50.

(32) Tan NC, Mulley JC, Berkovic SF. Genetic association studies in epilepsy: "the truth is out there". Epilepsia 2004 November;45(11):1429-42.

(33) Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. Nat Genet 2001 November;29(3):306-9.

(34) Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 1995;57(1):289-300.

(35) Wolf FM. Meta-Analysis. Thousand Oaks, CA.: Sage Publications; 1986.

(36) Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions. ed 3 ed. New York: Wiley; 2003.

(37) Murphy KR, Myors B. Statistical Power Analysis. Mahwah, N.J.: L. Erlbaum Associates; 2004.

(38) Cohen J. The earth is round ($p < .05$). American Psychologist 1994;49(12):997-1003.

(39) Royall RM. Statistical Evidence: A Likelihood Paradigm. London: Chapman and Hall; 1997.

(40) Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med 1999 June 15;130(12):995-1004.

(41) Tatsuoka M. Effect size. In: Keren G, Lewis C, editors. A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues.Hillsdale,N.J.: Lawrence Erlbaum Associates; 1993. p. 461-79.

(42) Rosenthal R. Parametric Measures of Effect Size. In: Cooper H, Hedrick PW, editors. The Handbook of Research Synthesis.New York: Russell Sage Foundation; 1994. p. 231-44.

Table 1 The meaning of a replication (adapted from Rosenthal (21, p. 541))

		second study	
		$p_2 \leq \alpha$*	$p_2 > \alpha$
first study	$p_1 \leq \alpha$	Successful replication	Failure to replicate
	$p_1 > \alpha^{\dagger}$	Failure to Replicate	Failure to establish effect

p_1= p-value of first study, p_2 = p-value of second study, α = test size for both studies.

*for a true replication, the effect in the second study must be in the same direction as that in the first. † a second study might be contemplated even when the first study was not significant if (i) p_1 was only slightly larger than α, (ii) the nominal α used was extremely small, or (iii) there was sufficient biological justification.

Table 2 Contingency table and chi-square test statistics for first and second studies.

			i[th] study (i = 1,2)
	Genotype with at least one marker allele	Genotype with no marker allele	row total
Disease	a_i	b_i	n_i
Non-disease	c_i	d_i	n_i

$$X_i^2 = \frac{2(a_i d_i - b_i c_i)^2}{n_i (a_i + c_i)(b_i + d_i)}$$

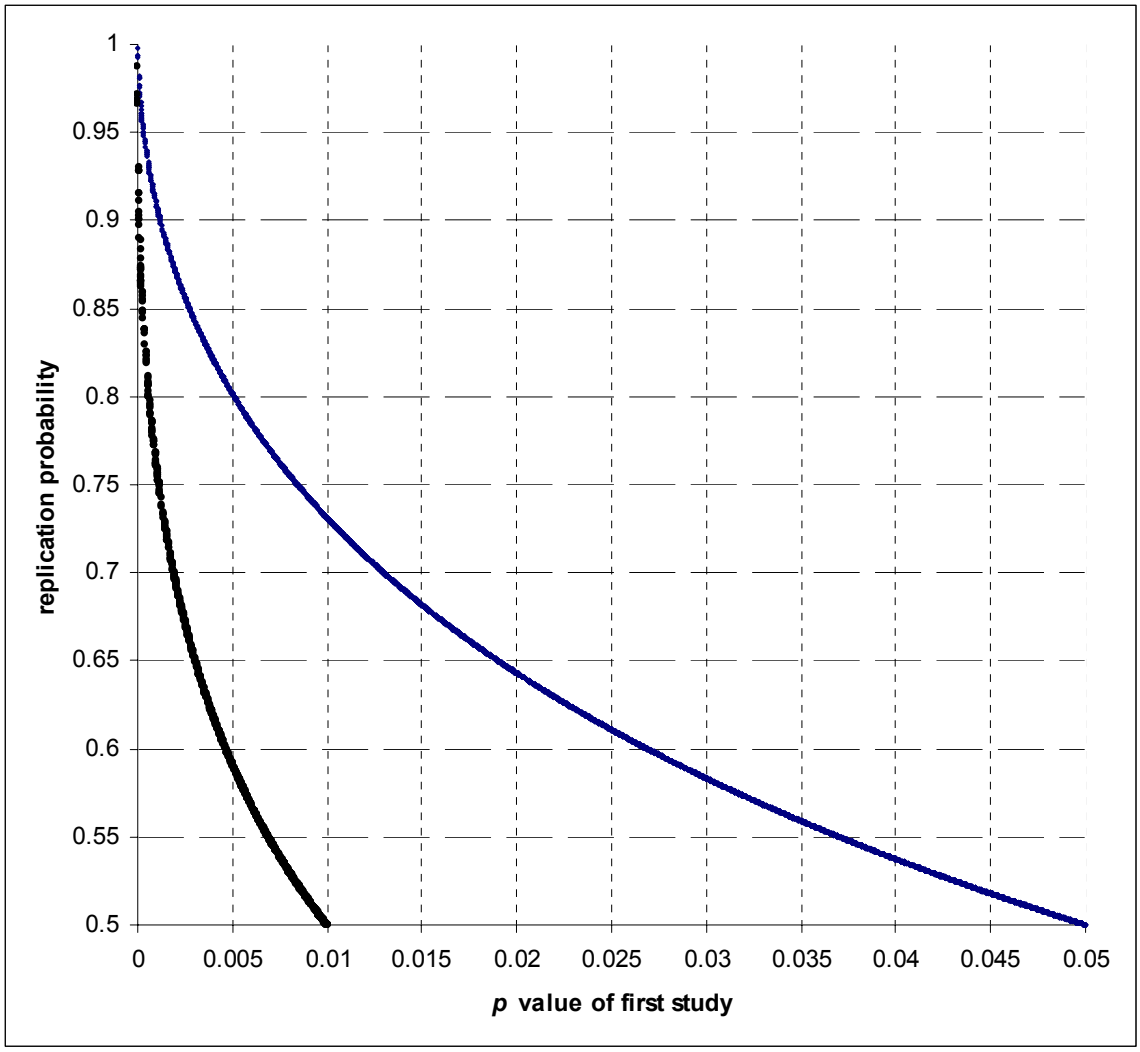

Fig. 1 Variation of replication probability as a function of the *p*-value of the chi-square test in the first study. The first and second studies are assumed to have the same sample size. For top graph, α = .05; for bottom graph, α = .01. For other values of α, use Eq. (1).

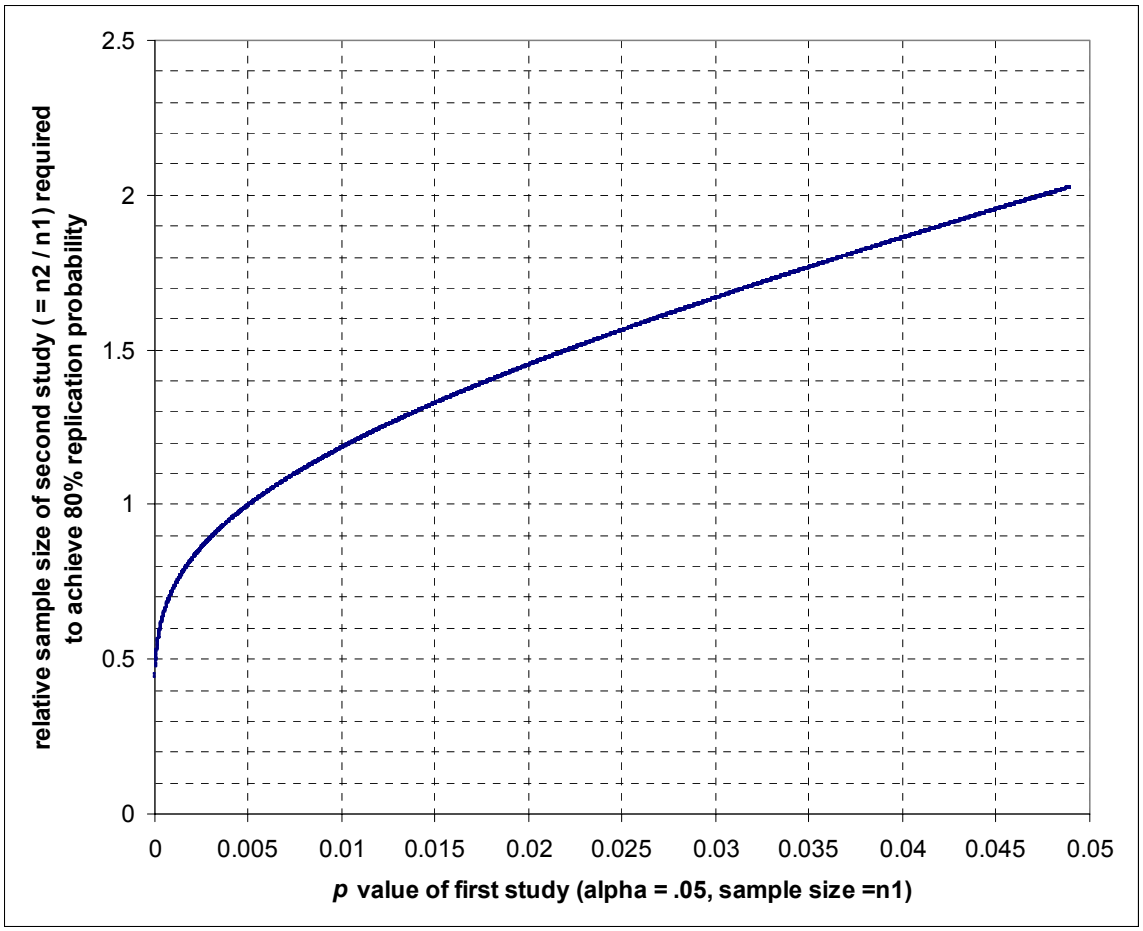

Fig. 2 Variation of the relative sample size (number of cases or controls) required for the second study as a function of the *p*-value of the chi-square test in the first study to achieve a replication probability of 80%.