Problems of reporting genetic associations with complex outcomes

Helen M Colhoun, Paul M McKeigue, George Davey Smith

Inability to replicate many results has led to increasing scepticism about the value of simple association study designs for detection of genetic variants contributing to common complex traits. Much attention has been drawn to the problems that might, in theory, bedevil this approach, including confounding from population structure, misclassification of outcome, and allelic heterogeneity. Other researchers have argued that absence of replication may indicate true heterogeneity in gene-disease associations. We suggest that the most important factors underlying inability to replicate these associations are publication bias, failure to attribute results to chance, and inadequate sample sizes, problems that are all rectifiable. Without changes to present practice, we risk wastage of scientific effort and rejection of a potentially useful research strategy.

The post-genome era has been responsible for a flood of new data submitted for publication. Journals are littered with studies reporting an association between genetic variation and disease-related outcomes. With several million single-nucleotide polymorphisms (SNPs) catalogued in the human genome, the number of possible genetic associations that can be tested is limited only by the rate at which laboratories can type these polymorphisms. For such data to be useful, procedures are needed that will distinguish between real and spurious associations and will collate the mass of data for any given polymorphism or outcome.

Researchers recognise that few published reports of significant associations can be replicated unequivocally. A typical example is the initial positive associations between the G460W polymorphism of alpha-adducin and hypertension, which have not been replicated by others. This uncertainty is a general occurrence. As Cardon and Bell state, it is “impossible to consider the veracity of a new report of complex disease association in the light of past knowledge, because so few ‘real’ associations are known”. This situation has led some commentators to question the value of genetic association studies, suggesting that they should be restricted to the study of polymorphisms that have been shown to have a direct effect on gene function. In the first part of this article, we discuss some of the possible explanations why findings from so many genetic association studies are inconsistent, and then we propose strategies for dealing with the problem.

Is the association study strategy futile?

Can any association study ever be expected to detect genetic determinants of a complex disease? Several researchers have argued that attempts to locate disease genes must account for the complexity of biological and social pathways to disease and the context dependency of the effect of many risk factors. In short, the assertion here is that genes do not generally act in a simple additive manner, but through complex networks involving gene-gene and gene-environment interactions, so that association approaches will always be hopelessly simplistic and reductionist. However, although many interaction effects will doubtless exist that can affect power to detect average genetic effects and ability to replicate associations, important additive effects of genes can be expected, simply because much of the rationale for searching for complex disease genes is based on the presence of parent-offspring resemblance for the trait which is not caused by shared environment. The degree to which parents and offspring resemble each other for genetic reasons is attributable to the additive rather than interactive effects of genes. Empirically, comparisons of monozygotic twin and parent-offspring concordance accord with presence of strong additive effects for many complex traits that have been considered. Thus, a so-called reductionist strategy of studying genes one at a time should yield useful results, even when epistatic or gene-environment interaction effects are not being modelled simultaneously.

Other researchers have pointed to difficulties inherent in reliance on linkage disequilibrium between marker alleles and disease alleles to detect associations. Unless a polymorphism that is known to affect gene function can be identified in advance, detection of association between genotype and outcome is usually reliant on the presence of association (linkage disequilibrium) between alleles at an undetermined disease locus and alleles at the nearby marker loci that is actually being genotyped. When many disease-causing alleles are present in a gene, so-called allelic heterogeneity, marker studies will have little power to detect associations with disease. However, although allelic heterogeneity has been reported
for many rare recessive diseases, population theory suggests that for genes accountable for most population risk for common diseases, the allelic spectrum will be small, though empirical evidence for this suggestion does not yet exist. In other words, for common diseases, only a few alleles in a given gene are likely to contribute to disease. As we discuss below, allelic heterogeneity should not cause inconsistency between studies of common diseases between most populations either.

**Why are reports of associations between genotype and outcome so frequently inconsistent?**

In principle, it is possible to measure associations between genotype and outcome more reliably than in classic epidemiological studies, because biases that are unavoidable in epidemiological studies can be largely eliminated when studying genetic associations. The exposure (genotype) can be measured after disease onset.

| Panel 1: Reasons for inability to replicate genetic associations |
|----------------------|-------------------------------------------------|-------------------------------------------------|
| **Category**         | **Possible reasons**                            | **Comment**                                    |
| True variation in the underlying association between genotype and outcome between populations studied | Disease-causing allele is in linkage disequilibrium (i.e., is associated) with a different allele at the marker locus (i.e., the locus under study) in different groups | Unlikely that allelic associations will vary much between continental populations, apart from when disease-causing variant is rare or has been subject to differential selection pressure |
|                      | Allelic heterogeneity exists between ethnic groups: different variants within the same gene contribute to disease risk | Unlikely that much allelic heterogeneity will be present within continental populations apart from when disease-causing variant is rare or has been subject to differential selection pressure |
|                      | The association is modified by other genetic or environmental factors that vary between the groups studied | Effect modification by genes unlikely to account for failure to replicate studies in similar populations. Modification by environmental factors more likely, not especially when absolute risk of disease varies |
| Biases vary between studies | Selection bias: population at risk consists of strata that vary by allele frequencies at the marker locus, or the population is admixed, and the probability of being selected as a case or control varies by stratum | Avoidable by careful sampling strategy |
|                      | Differential misclassification of exposure: for instance when cases and controls have been genotyped in separate batches with genotyping error rates that vary between batches | Easily avoided by appropriate laboratory procedures |
| Confounding by population substructure | Differential misclassification of outcome: possible if genotype is known when outcome is classified | Unlikely, because outcome is usually defined in advance of genotyping |
|                      | Population is divided into strata that vary by disease risk and by allele frequencies at the marker locus | Unlikely to be a serious problem in most studies: when confounding is a problem, it can be controlled in study design by restriction or use of family-based controls, or in the analysis by quantifying and controlling for substructure |
| Absence of power leading to false-negative results in studies that do not replicate earlier positive findings | Inadequate sample sizes, or differences in the efficiency of the sampling strategy (for instance sampling the entire distribution of a trait, instead of the tails of the distribution) | Need to establish whether a negative study had adequate power to detect an association of the size originally reported |
|                      | Non-differential misclassification of outcome | Even substantial misclassification of outcome can be compensated for by modest increases in sample size |
|                      | Failure to consider that the initial effect size reported is an inflation of the true effect size | Replication studies should be powered to detect effect sizes that are smaller than the initial effect size reported, especially when the initial study had low power |
| False positive results by chance in initial positive studies | Case mix heterogeneity in an apparently homogenous outcome between populations studied: for instance in a study of stroke, mix of haemorrhagic and thrombotic subtypes may vary between populations | Unlikely to be an explanation for failure to replicate studies in similar populations with similar case sampling strategies |
|                      | Multiple testing and publication bias: multiple loci are assessed in each study, many statistical tests are done, and multiple studies are undertaken but only positive results are reported | The most likely reason for failure to replicate |
without systematic error. Selection bias in sampling of cases or controls is less serious in genetic association studies than in studies of environmental exposures because it is unlikely to perturb genetic associations unless population stratification exists. Several authors have listed possible reasons for inconsistency in results of genetic association studies (panel 1).\textsuperscript{14,17,22} Several points are worth emphasising.

**True variation of underlying association between populations**

True variation in the presence or size of an association between an allele and disease could cause associations to be non-replicable.\textsuperscript{15,23} Such variation could arise if: (1) different disease-causing alleles predominate in different study populations, or variation in the degree of linkage disequilibrium between marker and disease alleles exists; or (2) allele frequencies are similar, but heterogeneity in the size of the effect of the disease gene between study settings is present. We consider these possibilities below.

**Allelic associations that vary between study settings**—For common alleles contributing to disease risk, allele frequencies will not vary much between similar populations. For instance, between subpopulations of European ancestry, allele frequencies do not vary widely (the standardised variance of allele frequencies is less than 0·01).\textsuperscript{24} Furthermore, for common diseases, population theory predicts that the half-life for the ancestral allelic spectrum will be very long, in the order of 1·3 million years. In other words, the same set of disease alleles in a gene is likely to persist in most human population groups since they have diverged much more recently than this time.\textsuperscript{25}

Whether allelic associations (the relation between marker allele and linked disease allele) vary between populations depends on the time back to the most recent common ancestor of the disease-causing variant. When disease-causing variants of a gene are rare (eg, the breast cancer-associated BRCA1 gene) or have been subject to strong selection pressure (eg, some haemoglobinopathies), the time back to the most recent common ancestor is likely to be short. Accordingly, the haplotype on which this common ancestor was present may vary between populations. However, for a common variant, the most recent shared ancestor of all existing copies is likely to be distant in time, antedating the split of human populations into continental groups. Thus, associations between common variants at nearby loci do not generally vary much between the main human subpopulations.\textsuperscript{23,26} For example, the thermolabile variant of the 5-10 methylenetetrahydrofolate reductase (MTHFR) gene is associated with a common haplotype in people from Israel, Japan, and Ghana.\textsuperscript{27}

**Effect modification by other genetic or environmental factors**—True heterogeneity of effect size could arise from effect modification by other genetic or environmental factors, if these factors vary between study settings. For the reasons discussed in the preceding paragraph, the frequencies of common alleles or haplotypes—including alleles at other so-called modifier loci—are unlikely to vary much between populations or ethnic groups originating from the same continent. Thus, inconsistency between studies in populations or ethnic groups originating from the same continent is unlikely to be attributable to gene-gene interaction.

For heterogeneity of size of genetic effect to result from the effect of variation in prevalence of an environmental factor, fairly extreme circumstances must hold: for instance, if the effect of genotype on disease risk is entirely restricted to people exposed to the environmental factor, and prevalence of this exposure varies twofold between populations, a twofold variation in average effect size of genotype on disease risk would exist. Nonetheless, when large variations in the absolute risk of disease between populations under study are present then variation in effect size because of gene-environment interaction is a potential explanation for inability to replicate associations.

However, in studies of populations of similar genetic background and with similar rates of disease, it is unlikely that heterogeneity of effect size can be ascribed to gene-environment interaction. At present, many genetic association studies do not include data on known environmental risk factors, and though this inclusion would allow examination of whether environmental factors contributed to inability to replicate results, modelling of the gene-environment interaction does not, in general, make it easier to detect genetic association.\textsuperscript{28}

With respect to empirical evidence for heterogeneity of effect sizes, in a meta-analysis of 36 genetic associations,\textsuperscript{14} showed some evidence for lack of fit of a simple so-called fixed-effects model, in which all studies are estimating the same effect size (odds ratio) without error. Whether this finding results from true heterogeneity of effect size between studies, or from publication bias, remains to be established.

**Misclassification of outcome**

One explanation for failure to replicate findings has been that, in some studies, misclassification of outcome has weakened the association with genotype.\textsuperscript{29,30} Apart from when studying a rare subtype of a common disease, the level of misclassification is unlikely to be enough to explain persistent failure to replicate associations with diseases for which diagnostic criteria have been standardised. For instance, in a case-control study of type 2 diabetes, suppose that 10% of patients classified by the investigator as cases have another disorder (such as latent autoimmune diabetes of adulthood), and 2% of those classified as controls have undiagnosed type 2 diabetes. This degree of misclassification would weaken the noted association only slightly, reducing a true odds ratio of, say, 1·7 to 1·6 for a control allele frequency of 20%. A modest increase in sample size (about 1·3-fold) would be sufficient to compensate for this weakening of association.

**Population stratification**

When the population under study consists of a mixture of two or more subpopulations that have different allele frequencies and disease risks (for environmental or genetic reasons unrelated to the allele under study), associations between genotype and outcome could be confounded by population stratification. This confounding could give rise to associations of disease with genotype at loci that are not linked to any locus that affects disease susceptibility. These associations would be regarded as false positive, by contrast with associations that result from confounding by allelic association between tightly linked loci. Even if no difference in disease frequency between population strata were present, spurious associations could, in theory, arise if the probability of a case or control being selected into the study were not independent of population stratum of origin (ie, selection bias).

The extent to which stratification actually poses problems for association studies has been extensively discussed.\textsuperscript{29,30,32} Researchers have calculated that only in
extreme situations (few ethnic groups, great differences in disease and genotype frequencies) will any substantial confounding occur, and these situations should usually be obvious to investigators. Few examples of studies exist in which false-positive associations arose from hidden population stratification and in which the initial investigators did not consider population stratification as the explanation. A possible example is a report that prostate cancer in African-American populations is associated with a polymorphism (in the CYP3A4 gene) at which allele frequencies differ strikingly between west Africans and Europeans. The most likely explanation for this association is that in this admixed population, both CYP3A4 genotype and risk of prostate cancer are independently related to the proportion of the genome that is of west African ancestry. This situation is, of course, extreme: risk of prostate cancer is about twofold higher in people of west African descent than in Europeans, great stratification (variation of individual admixture proportions) within the African-American population is present, and allele frequencies at the CYP3A4 marker locus show large differences between west Africans and Europeans.

When population stratification cannot be controlled by standard epidemiological methods (restriction or matching by demographic background), statistical techniques to model it directly, using a panel of marker loci, are under development. Control for stratification and admixture will also eliminate selection bias and will remove confounding attributable to differential distribution of environmental risk factors by ethnic origin.

Variation in power between studies

An obvious cause of failure to replicate findings is that subsequent studies are underpowered. Important to the power is the underlying size of effect (genotypic relative risk or odds ratio) to be detected, and this effect will typically be small in genetic association studies. For diseases that have only modest familial aggregation (sibling recurrence risk ratios of 3 or less—e.g., with type 2 diabetes), we can infer that only a few common polymorphisms of large effect (genotypic risk ratios of 2 or more) can exist, though many polymorphisms of smaller effect (genotypic risk ratios 1·5 or less) may be present.

When choice of candidate genes is guided by previous evidence from family linkage studies, to expect large effect sizes may be reasonable, since otherwise linkage would not have been detected. When polymorphisms have been chosen for study because they have been shown to have effects on potentially important intermediate phenotypes, the effect that the researcher is trying to detect is generally much smaller. For example, individuals homozygous for the thermolabile variant of MTHFR have homocysteine levels that are 2·6 μmol/L higher on average than those homozygous for the wild-type MTHFR. A meta-analysis of observational studies showed that a difference in homocysteine of this magnitude would be expected to be associated with a relative risk of coronary heart disease of just 1·13 (95% CI 1·08–1·19). Thus a large sample size would be needed to detect an association between the thermolabile variant of MTHFR and coronary heart disease, and small negative studies would themselves be inconclusive. Indeed a meta-analysis from 40 studies involving a total of 11 162 cases and 12 758 controls was needed to show such an effect. In the latter, homozygosity for the thermolabile variant of MTHFR was associated with an odds ratio of 1·16 (95% CI 1·05–1·28) for coronary heart disease, consistent with the magnitude of effect expected from the observational data.

Similarly, in a study of angiotensin-converting enzyme (ACE) insertion/deletion (I/D) polymorphisms and coronary heart disease, initial findings of high relative risks in some subgroups were implausible in view of the small differences in intermediate phenotype associated with this variant; the effect size 1·10 (1·00–1·21) in the large case-control study is reasonable, but a sample size of nearly 5000 cases and 6000 controls was needed to detect the effect, and it is still far from statistically robust. Effects that are detected by linked marker polymorphisms are even weaker, because they depend on confounding by alleles at the disease susceptibility locus.

In situations in which the initial positive finding indicates a true association, a possible reason for failure to replicate is that subsequent studies may be inadequately powered. Since, even without publication bias, investigators are more likely to attempt to replicate positive than negative findings, estimates of effect size in subsequent studies will regress to the true effect size, which will be less extreme. Subsequent studies must therefore be powered to detect smaller effect sizes than estimated by the initial study.

Failure to exclude chance as an explanation in some studies, and publication bias

We suggest that failure to exclude chance is the most likely explanation for difficulty in replication of reports of genetic associations with complex diseases. For most diseases of interest, hundreds of known genes are possible candidates, and in most of these genes, dozens of polymorphisms are known or can be easily identified by screening of the gene. Around the world, thousands of such polymorphisms are tested for disease associations every week. Even if none of these genotypes is associated with outcome, we can expect many associations that are significant at 5% or less to arise frequently by chance alone. When the initial test of association with genotypes or alleles does not reach significance, further exploration of subgroup-specific or haplotype-specific effects is possible, increasing the chance that at least one significant association will be detected.

Exacerbation of the problem of spurious chance findings is publication bias. Negative results in association studies, especially smaller ones, may not generally be submitted for publication, let alone accepted. This renders any systematic review of published results misleading. Empirical evidence exists for the role of publication bias in genetic association studies, at least for some associations. For example, considerable interest has been shown in the association between the ACE I/D polymorphism and risk of myocardial infarction. In the figure, odds ratios from 19 studies included in a meta-analysis are plotted against their SE in a funnel plot. This graph shows that the effect is large in small case-control studies but only modest in large studies, which accords with publication bias. In the largest study done to date (about 5000 cases and 6000 controls), the estimated risk ratio was just 1·10 (CI 1·00–1·21). The finding that initial striking findings in genetic association studies fail to be matched by those of later studies seems to be a general occurrence.

Possible approaches to exclusion of results seen by chance

Adoption of more stringent criteria for exclusion of chance as an explanation for noted associations between genotype and outcome will reduce this problem. But how stringent should the criteria be?

For personal use. Only reproduce with permission from The Lancet Publishing Group.
Bayesian approach

Many researchers find correction for multiple testing counterintuitive and view it as an unnecessary hurdle imposed by statisticians. In this next section, we elaborate a rationale for more stringent significance levels based on Bayesian arguments, which we think researchers will find more intuitive.

We regard the interpretation of results from an association study as a decision problem. Researchers studying a genotype–outcome association must choose the appropriate course of action when results are available: to abandon further study of this association; to continue collecting data; or to declare that evidence of association has been detected (laying the grounds for further studies of the molecular basis of this effect and potential interventions based on this understanding). The rational decision will depend on the probabilities they attach to each possible result of their decision (missing a true association, investigating a false association, detecting a true association, and so on) and on the values (utilities or regrets) they place on these results. To the extent that these values are likely to differ between, for example, journal editors, readers, and researchers, the criteria for declaring significance might also vary.49,50 Nonetheless, we might reasonably suppose there is consensus that criteria should be set so that the probability that an association is false positive, given that criteria for declaring significance are met, is for example no greater than about 5%. Or, in other words, the ratio of true-positive to false-positive associations that an author or journal publishes should be about 20:1.

Although often mistaken as such, this ratio is not the same as a 5% significance level, which is a probability of 5% of declaring that an association exists if the null hypothesis were true. The probability of a given significant association being true or false is dependent both on the significance level and the overall proportion of hypotheses being tested for which the null hypothesis is true.51 The table illustrates the latter point. First, suppose that in 90% of all studies undertaken the null hypothesis is true. Second, suppose that the average power of association studies is about 80% for detection of whatever magnitude of odds ratio is of interest. Accordingly, in 1000 studies, 100 true associations will exist and we will detect 80 of these as being significant at the 5% level. Of the remaining 900 non-associations, we will declare 5% or 45 as being significant associations. Overall, of the 125 associations declared to exist, 36% (45/125)—not 5%—will be false.

These figures are of course arbitrary, but they serve to illustrate the point that the proportion of significant associations reported at any given p value depends on the p value, the overall proportion of hypotheses being tested that are true, and the statistical power (and thus sample size) of the study. The argument is somewhat simplified, since it assumes that what is being tested in an association study is two discrete alternatives of association versus no association. However, in reality, the threshold for the odds ratio or risk ratio that one interprets as constituting association is arbitrary and set by the researcher at the time of the power calculation.

If there were agreement on the prior probability of a true association, then we could decide rationally on a significance threshold that would ensure the ratio of true-positive to false-positive associations is about 20:1. In practice, researchers would be unlikely to agree on these

### Table: Relation between significance level and predictive power of a positive test of association under assumptions about power and proportion of all hypotheses being tested that are true

<table>
<thead>
<tr>
<th>Result of experiment</th>
<th>Polymorphism really is associated with disease</th>
<th>Polymorphism is not associated with disease</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Association declared to exist</td>
<td>80</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>Association not declared</td>
<td>20</td>
<td>855</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>900</td>
<td>1000</td>
</tr>
</tbody>
</table>

For personal use. Only reproduce with permission from The Lancet Publishing Group.
prior probabilities. However, what we can say with reasonable certainty is that under present circumstances, in view of the number of associations that are never replicated, the ratio of true-positive to false-positive associations is probably not much better than 1:20. Therefore, to achieve posterior odds of 20:1 (95-2% of all reported associations being valid) would mean a 400-fold more stringent p value under present circumstances—if assuming that power was maintained at the same level by whatever sample size increase was necessary—and that the prior probability of association remained the same.

If we consider that the ratio of true-positive to false-positive associations at present is even lower than this ratio, say 1:50, we would have to use p values about 1000 times more stringent than at present. Since a p value of $5 \times 10^{-3}$ is the typical threshold presently used for genetic associations, we should probably be using significance thresholds in the range of $5 \times 10^{-1}$ to $5 \times 10^{-3}$. Thus, we suggest that, in general, in candidate-gene studies of complex traits, the conventional threshold for declaration of statistical evidence of association—a p value of $5 \times 10^{-3}$—should be reduced to about $5 \times 10^{-5}$.

The p value of $5 \times 10^{-5}$ could be interpreted in a frequentist reasoning as an adjustment for about 1000 tests, or in a Bayesian framework as ensuring a ratio of true positive to false-positive associations of 20:1. Declaration of association should not be made in simplistic terms of present or absent, but rather authors should comment on the strength of evidence for association of their data, and should consider other factors, such as evidence for functionality of the polymorphism. As the table shows, this stringency may not be needed in situations when the prior probability of association is much higher than average. For example, when studying direct associations of genetic polymorphisms with some aspect of the gene product (such as polymorphisms of MTHFR genes and homocysteine concentrations), the prior probability of association is higher than that seen when doing association studies with a wide range of candidates with no or only weak established functional links with a disease under investigation. Similarly, if in the future the overall proportion of tests of association that are true were substantially increased, this stringent p value threshold could be relaxed. The ability to preselect SNPs that tag common haplotypes in a gene might also increase the prior probability of association when a candidate gene is being studied.

On the other hand, the capability for genome-wide association studies will allow all genes to be tested whether or not they have about 90% power to detect odds ratios as low as 1.25 at control allele frequencies of 30%. Of course, for very small effects, the required sample size will be very large. For example, with a control allele frequency of 30%, to detect an odds ratio of 1.15 at p$=5 \times 10^{-4}$ would need about 14 000 cases and 14 000 controls.

Thus, serious study of genetic associations with complex outcomes will need larger case-control collections than are typically available to researchers at present—typically thousands of cases and controls, rather than hundreds. With multicentre collaborations to recruit cases and controls, assembly of such large case-control collections is feasible. Although these sample sizes seem daunting, consider that in the more complex area of clinical trials, for which prospective assessment of participants is needed, sample sizes of 20 000 or so have been achieved. That researchers quantitatively assess the trade-off between spending scarce resources on ensuring little or no misclassification of disease or doing larger simpler studies is important. For many common phenotypes, probable levels of misclassification can be tolerated, and the more efficient strategy will be to use a less specific but more easily applicable set of entry criteria. These collections have only to be laid down once: after they have been irreversibly made anonymous, DNA samples and clinical data can be shared freely. Once these sample banks have been assembled and DNA extracted, the cost of genotyping large numbers of samples at a locus is modest: present rates in high throughput laboratories are lower than $0.50$/genotype (US$0.50/genotype). If all such collections are registered, and users of each collection submit their genotype data to one database of all genotypes scored on the samples, to keep publication bias to a minimum should be possible.

**Replication studies**

The recommendation that large case-control banks are needed to test for associations is set against the present reality that most researchers have only small collections available to them. In an attempt to deal with the limitations imposed by these small sample sizes, replication of associations before declaration of evidence as convincing is gaining increasing acceptance. However, some clarity about the role of replication studies is needed. These studies do have advantages, in that biases in any initial study are unlikely to be the same in subsequent studies. However, researchers have a tendency to think that one weak association followed by another—sometimes in a different subgroup—constitutes strong evidence, when clearly this is not the case. Only if precisely the same hypothesis is being tested can replication really be said to have been done.

Theoretically, a replication study should need a less stringent p value because a more specific hypothesis is being tested. However, as noted above, replication studies are more likely to give effect sizes smaller and closer to the true effect size than initial published reports of an association. A

---

**Reporting subgroup analyses**

Even the simplest genetic association study can be analysed in several different ways—for instance, by testing the effects of alleles, genotypes, or haplotypes. Additional analyses by subgroup—eg, age, sex, categories of phenotype—can be undertaken. Subgroup analyses are typically done when no robust main effect is seen, to guarantee publication of some apparently positive finding. Since the number of possible subgroup analyses that can be undertaken is large, significant results obtained in such analyses should be treated with even more scepticism than tests for the average genetic effect across all subgroups combined.

---

**Implications for sample size**

The increase in sample size needed to maintain statistical power to detect a given association at more stringent p values is challenging but feasible, as the required sample power to detect a given association at more stringent stringency may not be needed in situations when the prior probability of association may be much higher than average. For example, when studying direct associations of genetic polymorphisms with some aspect of the gene product (such as polymorphisms of MTHFR genes and homocysteine concentrations), the prior probability of association is higher than that seen when doing association studies with a wide range of candidates with no or only weak established functional links with a disease under investigation. Similarly, if in the future the overall proportion of tests of association that are true were substantially increased, this stringent p value threshold could be relaxed. The ability to preselect SNPs that tag common haplotypes in a gene might also increase the prior probability of association when a candidate gene is being studied.

On the other hand, the capability for genome-wide association studies will allow all genes to be tested whether or not they have about 90% power to detect odds ratios as low as 1.25 at control allele frequencies of 30%. Of course, for very small effects, the required sample size will be very large. For example, with a control allele frequency of 30%, to detect an odds ratio of 1.15 at p$=5 \times 10^{-4}$ would need about 14 000 cases and 14 000 controls.

Thus, serious study of genetic associations with complex outcomes will need larger case-control collections than are typically available to researchers at present—typically thousands of cases and controls, rather than hundreds. With multicentre collaborations to recruit cases and controls, assembly of such large case-control collections is feasible. Although these sample sizes seem daunting, consider that in the more complex area of clinical trials, for which prospective assessment of participants is needed, sample sizes of 20 000 or so have been achieved. That researchers quantitatively assess the trade-off between spending scarce resources on ensuring little or no misclassification of disease or doing larger simpler studies is important. For many common phenotypes, probable levels of misclassification can be tolerated, and the more efficient strategy will be to use a less specific but more easily applicable set of entry criteria. These collections have only to be laid down once: after they have been irreversibly made anonymous, DNA samples and clinical data can be shared freely. Once these sample banks have been assembled and DNA extracted, the cost of genotyping large numbers of samples at a locus is modest: present rates in high throughput laboratories are lower than $0.50$/genotype (US$0.50/genotype). If all such collections are registered, and users of each collection submit their genotype data to one database of all genotypes scored on the samples, to keep publication bias to a minimum should be possible.

---

**Replication studies**

The recommendation that large case-control banks are needed to test for associations is set against the present reality that most researchers have only small collections available to them. In an attempt to deal with the limitations imposed by these small sample sizes, replication of associations before declaration of evidence as convincing is gaining increasing acceptance. However, some clarity about the role of replication studies is needed. These studies do have advantages, in that biases in any initial study are unlikely to be the same in subsequent studies. However, researchers have a tendency to think that one weak association followed by another—sometimes in a different subgroup—constitutes strong evidence, when clearly this is not the case. Only if precisely the same hypothesis is being tested can replication really be said to have been done.

Theoretically, a replication study should need a less stringent p value because a more specific hypothesis is being tested. However, as noted above, replication studies are more likely to give effect sizes smaller and closer to the true effect size than initial published reports of an association. A

---

**Implications for sample size**

The increase in sample size needed to maintain statistical power to detect a given association at more stringent p values is challenging but feasible, as the required sample size scales with only the logarithm of the p value. A rough working rule is that if the threshold p value is lowered by $n$ powers of 10, the sample size needed to maintain statistical power of 90% increases by around $n$-fold. Thus, a roughly three-fold increase in sample size is needed if the threshold p value is lowered from $5 \times 10^{-2}$ to $5 \times 10^{-4}$. To give an idea of what this increase would mean, at this level of significance, studies of 5000 cases and 5000 controls would
practical aspect of this finding is that although less stringent p-value thresholds may be appropriate for replication studies, smaller sample sizes than the initial study might not be appropriate, because effect sizes will usually be smaller. Ultimately, the best approach is not to have many underpowered studies—of which only a biased sample are published—but fewer large studies; thus the establishment of large case-control collections is essential.

Some researchers might argue that having many small studies is preferable to a few large ones on the grounds that they can provide information about heterogeneity of effect size. In practice however, large case-control studies will usually be based on multicentre collections, so they can also provide information about possible heterogeneity of effect size between different study settings with the advantage of having been done with standardised methods.

**Keeping publication bias to a minimum**

For clinical trials, publication bias is kept to a minimum by prospective registers and systematic reviews. For genetic association studies, a different approach is needed, because most negative results will never even reach conference proceedings, protocol papers will not usually have been published to suggest that the study is ongoing, and an effective mechanism for establishment of prospective registers of proposed analyses is not feasible. To propose that journals should publish all negative studies that are done is unfeasible, thus other methods will have to be adopted.

The most promising way to avoid this bias is to provide alternative ways of publishing. Specifically, as others have advocated, internet-based reporting must become more widespread, but to succeed, scientists must be confident that this type of publication will be well regarded by their peers, and this respect will take time to establish. Scientists with negative findings should be encouraged to submit them in a brief format for web-based publishing. Ultimately, this publication would allow the proportion of null hypotheses that are false (the prior probability of association, in Bayesian terms) to be estimated for genotype-disease associations in general. Journal editors might consider restriction of paper publication by using significance criteria that aim to reduce the proportion of reports that are false positives to an acceptable amount, and when referees are satisfied that the issues summarised in checklist format in panel 2 have been considered. Other study results could be disseminated through web-based publishing.

**Summary of data in systematic reviews**

Even if publication bias can be reduced, and stringent criteria are adopted for declaration of significance, the problem will remain of how to collate this profusion of data to extract useful information. For any given genotype-disease association, considerable effort is needed to assemble all available studies. Some databases containing reviews of genetic variants and reported associations with various outcomes already exist; for instance OMIM (Online Mendelian Inheritance in Man) and HuGENet (Human Genome Epidemiology Network). As data accumulate, to maintain up-to-date reviews of each genotype-outcome association is becoming a sisyphean task. A more feasible approach may be to establish an automated system for meta-analyses of each genotype-outcome association that has been studied. For an outcome such as diabetic nephropathy, for example, one could construct a web-based database in which submitted data are classified by polymorphism and outcome, and a meta-analytic plot is continually updated. Researchers would be encouraged to submit negative and positive findings in a standard format, and users would be able to assess a continually-updated summary estimate of the effect associated with the allele or genotype. Data with attribution and a simple description of the study population and methods could then be included in the updated meta-analyses proposed above. For this database to work in practice, the development of mechanisms to reward researchers who contribute to these databases will be necessary.

In summary, consensus between scientists and journal editors is needed about better ways to interpret and manage data emerging from genetic association studies so we can reduce the number of spurious findings that are declared significant and allow all available data to be collated without bias. We have proposed some guidelines, although we recognise that these are somewhat arbitrary, and welcome debate on the issues.

**Conflict of interest statement**

None declared.

**Acknowledgments**

We are grateful to Paul Burton, Andrew Hattersley, and Jonathon Steene who provided helpful comments on an earlier draft of this article.
References


16 Strohman RA. Maneuvering in the complex path from genotype to phenotype. Science 2002; 296: 701–03.


41 我要删除这一条。