Validation in Genetic Association Studies

Inke R. König
Submitted: 22nd September 2010; Received (in revised form): 2nd November 2010

Abstract
Validation of genetic associations is understood to be a cornerstone for the scientific credibility of the results. To approach this topic, the general concept of genetic association studies is introduced briefly, followed by how the term ‘validation’ is used in the context of genetic association studies. As a central issue, reasons for the importance of validation and for failure of validation will be described.

Genetic association studies have increasingly gained in importance for unraveling the genetics of complex diseases. Despite their value, problems to confirm their results are rather common, so that the need for validation is emphasized by the scientific community. In the following, I will approach this topic by briefly introducing the wider perspective of genetic association studies as a study type. Since there seems to be some confusion about the use of terms, the different possible meanings of validation in the context of genetic association studies are disentangled. Then, reasons for the importance of validation are given. In the final section, I will consider why replication and validation may fail and close with brief recommendations.

WHAT ARE GENETIC ASSOCIATION STUDIES?
Very generally, in genetic association studies, we investigate specific associations between a phenotypic variable on the one hand and a genetic variable on the other. The phenotypes here can be of any kind: we may use a binary variable, for example, for an affection status such as myocardial infarction yes or no, a quantitative trait such as blood pressure, or even survival endpoints, although this is rare.

As genetic variables, we make use of genetic markers. These are defined as being special genetic loci with a known location in the genome and with at least one difference between at least two individuals that is detectable in the laboratory. The genetic marker most often used in current genetic association studies is the single nucleotide polymorphism (SNP) [1]. Basically, this is a variation at only a single base in the DNA that comes usually in two variants (alleles) in the human population with a frequency of the rare allele of at least 1%, therefore a polymorphism. And it is associated with an affection status if one of its alleles occurs more often than expected by chance in affected individuals than in non-affected ones. More detailed descriptions are found, e.g. in references [2, 3].

Genetic association can be investigated using different study designs. The most important distinction for the current purpose is whether family data are used or data from unrelated individuals. For family data, the classical design involves the recruitment of trios, that is, of affected children with both of their parents. There are a number of advantages to this design. Most importantly, one may simultaneously analyze genetic association and genetic linkage, which is the co-segregation of the marker and the disease within families. This attenuates a number of problems that come with association studies, such as possible confounding by ethnicity (population stratification). Additionally, this may allow for the identification of genetic effects that deviate from the classical modes of inheritance, such as imprinting, and of epigenetic phenomena [4]. As a result, the requirements for validation are different than for studies based on unrelated individuals. In the following, I therefore focus only on association studies using data from unrelated individuals.

Here, two very different scenarios need to be distinguished. In the first, study designs are experimental
designs in which the probands are exposed to an independent variable in a planned way. For instance, in a controlled clinical trial of Phase III, probands may receive different interventions according to a randomization scheme. Since the genetic factor cannot be manipulated experimentally, it may enter in these studies as an additional variable. For example, in a pharmacogenetic clinical trial, it may be investigated whether the efficacy of an intervention depends on a specific genotype [5]. Hence, these studies require that an association between the genotype and specific phenotypes that are related to metabolism has already been shown.

In the other, more common, scenario, observational studies are carried out in which an association between an exposure and the outcome is merely observed. The possible study designs resemble those from classical epidemiological studies. Thus, for binary phenotypes, case–control or cohort studies are common, whereas cross-sectional studies less frequently applied. Briefly, in the case–control design, a sample of affected and unaffected individuals is recruited. These are then analyzed with respect to their genotype. In contrast, in the cohort design, a sample of probands is recruited without the knowledge of their disease status. The probands are investigated with regard to their genotypes, and it will then prospectively be observed who develops the disease. Both of these designs have their advantages and disadvantages, which are discussed in detail in the literature [6, 7]. From a statistical point of view, a major difference lies in how genetic effects may be estimated: if probands are recruited prospectively and independent of the disease status as in a cohort study, relative risks can be utilized. In case–control studies, however, risks generally cannot be estimated. In contrast, a comparison of the odds of disease is possible between two groups, so that the odds ratio is the estimate of choice.

Another distinction between different association studies regards the kind of genetic regions that are investigated at a time. For a long time, candidate–gene association studies were performed. Here, only a few genetic markers are analyzed in candidates, which are defined beforehand based on either prior knowledge about the location of a predisposing gene [8]. A number of differences are important to acknowledge in the context of validation. On the technical side, GWA studies make use of microarray chip technology, and the most commonly used platforms are provided by Affymetrix and Illumina. These can easily process up to a million of SNPs in thousands of probands per week, but the quality control of the resulting data is usually more complex than in low- or middle-throughput technologies [8, 9]. Secondly, prior knowledge feeding candidate studies might also enable the researcher to select SNPs that may represent the causal variant for the disease itself. In contrast, SNPs in a GWA study are selected for a good coverage of the genome, so that a SNP might be associated with a disease because it lies within the vicinity of the causal variant so that there is an association between the two variants, what is termed linkage disequilibrium [10]. A final difference is that in GWA studies, the effect sizes are usually smaller and required sample sizes larger. As a result, the samples are often less well characterized, and the controls might not be well matched to the cases but drawn from population-based cohorts.

WHAT IS MEAN BY VALIDATION IN THE CONTEXT OF GENETIC ASSOCIATION?

Within the wider context of genetic association studies, the term ‘validation’ has a number of different uses. For instance, before any genetic association is investigated in itself, quality control of the data plays an important role. Here, one step is the validation of the marker genotypes, i.e. the genetic information. Different genotyping methods in the laboratory are used with different degrees of reliability, and the validation of genotypes is often understood as the concordance between different typing methods (e.g. [11]). In contrast, the terms ‘repeatability’ and ‘reproducibility’ refer to the concordance of genotypes when the same methods are applied in the same and a different laboratory, respectively [12].

If the genotype information is used as part of a genetic test, the usage of terms is slightly different. Criteria applied to recommend a genetic test for clinical practice are derived from the ACCE model, which is an acronym for the criteria analytical validity, clinical validity, clinical utility and ethical, legal and social implications [13]. Specifically, analytical validity denotes the ability of the test to precisely
determine the genetic variants technically, which is related to the previous definition. Clinical validity, in turn, refers to the quality of a genetic test as a predictor for the disease in question. This is determined by the strength of the association and also on how well the genotype is able to differentiate between affected and unaffected individuals. This is evaluated using classical measures of diagnostic accuracy such as sensitivity, specificity, predictive values or receiver operating characteristic curves [14].

Yet another use of the term validation here is in the context of validating statistical models. The result of the association analysis might be a complex statistical model that, for instance, explains the risk for myocardial infarction taking into account a number of SNPs and classical risk factors in a multivariate combination. To avoid an optimistic assessment of this model, this subsequently has to be internally and externally validated before it can be put into practice [15]. In fact, this can be viewed as a more general case of the validation of a genetic association itself, which is described in the following.

The focus here is on the specific validation of a genetic association. For this, it is helpful to consider the wider perspective of how a genetic variant that is responsible for a disease might be identified [16]. This begins with the knowledge or hypothesis that there is a genetic background to the disease. After this, possibly family and linkage studies are performed to define a mode of inheritance and a crude localization. In the third step, an association analysis identifies a more precise locus or even a single variant, which involves replication, validation and thorough investigation of the region and affected phenotypes. In subsequent steps, the conferred risk may be estimated, and functional studies provide evidence about the pathophysiological mechanism. Even though functional results might be viewed as the eventual validation of the association results, I will concentrate here on the third step of association, replication and validation.

It should first be noted that these different terms—validation, replication and confirmation—are unfortunately used more or less interchangeably to indicate that a result is reproduced in another study group. To avoid confusion, we have recently suggested distinguishing the two following scenarios [17]:

(1) Original and confirmation sample are drawn from the same population, and systematic differences are reduced to a minimum. This design aims at a replication of a genetic association as defined below and is equivalent to reproducibility or internal validation in the context of statistical models [15].

(2) The confirmation sample stems from a population which is different than that from which the original sample was drawn. Differences between these populations may concern the ethnic background, the phenotype definition, the recruitment or sampling strategy and the time point of investigation. With this design, validation of the genetic association is attempted (termed temporal or external validation of statistical models).

More precisely, to validate a result means to obtain similar findings under modified influencing factors such as ethnicity, phenotype, time, etc. The consequence is that a validated genetic association shows a greater generalizability than a replicated association. As an example, the region 9p21 on chromosome was found to be associated with coronary artery disease in Caucasian populations [18, 19], which was subsequently replicated in Caucasian populations multiple times (for instance [20]). Recently, this association was also found in other populations such as Indians [21], which would be viewed as a validation.

From this distinction of terms, it becomes clear that the step of replication and validation comprises a process in itself. First, an association is established and confirmed through replication [22]. For this, the recommended concept is what has been termed exact replication [23, 24]. Generally speaking, this involves the examination of the same genetic variant in independent but similar data sets using the same analysis method. More specifically, the NCI-NHGRI Working Group on Replication in Association Studies set up precise criteria for establishing a positive replication that are shown in Table 1 [25].

In contrast to the exact replication, local replication refers to the analysis of the original marker plus other markers in the same region that were not part of the original experiment. This fine-mapping plays an important role in GWA studies when it may be assumed that the associated marker is not the causal variant and possibly not even the variant that best tags it [26]. Strictly speaking, this is already one step in the validation process. Finally, the validation may comprise the generalization and more thorough
investigation of a genetic association regarding its homogeneity across different populations and subgroups as well as in varying phenotypes [22].

**WHY IS VALIDATION IMPORTANT?**

The first reason for why association results need to be replicated and validated is found in the study design itself. Here, we have to keep in mind that genetic association studies are a special case of classical epidemiological association studies. Just like in these, the establishment of a causal relationship from an association is impossible directly, since experimental designs including, for instance, randomization cannot be used. Thus, to infer causality of a genetic factor (the exposition) to a phenotype (the disease), we need to consider classical criteria such as those proposed by Hill [27] (Table 2).

From a statistical point of view, the requirement of replication and validation is the most crucial criterion. In a similar way, Ioannidis et al. [22] defined three sets of criteria for credibility for the specific case of assessing genetic association, one of them being replication and validation.

The second reason why validation is important is more specific for genetic associations. It has been shown repeatedly that in the first study of an association, the effect is overestimated, and that there is only a modest correlation between effects in first and in subsequent studies on the same association [25, 28, 29]. Reasons for this discrepancy are that, first, there might certainly be true differences between the populations under study. Although this might be clarified in subgroup analyses, the power is usually too low for these. Second, there is a phenomenon known as ‘winner’s curse’ [30] or ‘Jackpot effect’ [31] originating in the fact that the associations with the strongest effects are inflated [28]. Related to regression to the mean [32], this occurs primarily because with a small sample, a weak effect becomes significant only if the effect is overestimated. This phenomenon is aggravated by a selective reporting of the analyses, possibly biased interpretation of results and publication and other forms of bias [28, 29, 33]. The result is that the first report of an

---

**Table 1: Criteria for positive replication**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Explanation</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replication sample size should be sufficient.</td>
<td>Replication data set(s) should be independent from initial data set(s).</td>
<td></td>
</tr>
<tr>
<td>Replication phenotype should be similar.</td>
<td>Replication effect should be similar in magnitude and in same direction.</td>
<td></td>
</tr>
<tr>
<td>Replication population should be similar.</td>
<td>Replication marker should be the same or in very high linkage disequilibrium, selected with strong rationale.</td>
<td></td>
</tr>
<tr>
<td>Genetic model in replication should be the same.</td>
<td>$P$-value of combined analysis should be smaller than initial $P$-value.</td>
<td></td>
</tr>
<tr>
<td>Report of replication should be in same level of detail as for initial report.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adapted from reference [25].

---

**Table 2: Conditions to infer causality from an association**

<table>
<thead>
<tr>
<th>No.</th>
<th>Condition</th>
<th>Explanation</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Strength</td>
<td>Stronger associations are more likely to be causal.</td>
<td>Given the plethora of disease-predisposing factors in complex diseases, this is not necessarily true for genetic associations.</td>
</tr>
<tr>
<td>2</td>
<td>Consistency</td>
<td>Association is reproducible in independent studies in different settings using different methods.</td>
<td>Hill [27] suggested that the differences between the studies should not be too large. Thus, this criterion falls somewhere between the 'exact replication' and 'validation'.</td>
</tr>
<tr>
<td>3</td>
<td>Specificity</td>
<td>Exposition is, in the ideal case, necessary and sufficient.</td>
<td>This is rarely realistic in classical as well as genetic epidemiological studies, and, according to Hill [27], must not be over-emphasised.</td>
</tr>
<tr>
<td>4</td>
<td>Temporality</td>
<td>Exposition precedes the disease.</td>
<td>This can generally be viewed as given for genetic association [22].</td>
</tr>
<tr>
<td>5</td>
<td>Biological gradient</td>
<td>Dose—response relationship: with a stronger exposition the likelihood of the disease increases.</td>
<td>This can be questioned for genetic associations, because, a priori, an additive genetic model is not more plausible than another one.</td>
</tr>
<tr>
<td>6</td>
<td>Plausibility</td>
<td>Association has a biological explanation.</td>
<td>Conditions 6, 7 and 8 should be viewed together: in some lucky cases, a biological explanation or analogical situation might be available; however, this really depends upon the biological knowledge of the day.</td>
</tr>
<tr>
<td>7</td>
<td>Analogy</td>
<td>Result coincides with relationships between similar expositions and diseases.</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Coherence</td>
<td>Result does not conflict with existing knowledge.</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Experiment</td>
<td>Controlled change in exposition alters disease.</td>
<td>This is impossible to underline genetic associations.</td>
</tr>
</tbody>
</table>

Adapted from reference [27].
association is more likely to be a false positive and the effect overestimated, which again emphasizes the need for a replication or validation of the findings.

In a GWA study that typically investigates complex diseases, the effect is often even more pronounced than in candidate gene studies, because the true effects are usually weaker. It should, however, be noted that the first study in a genome-wide setting rarely has the aim of exactly estimating genetic effects but only of identifying them [30], and a more correct specification of effects can then be followed. On the other hand, it has almost become common practice to combine the results from first and subsequent studies, so that pooled effect estimates are again biased. Recently, suggestions have been made how to correct for this [34, 35]. Still, it needs to be pointed out that if data are combined across stages—in a multi-stage design, a sequential design, or in meta-analyses—a positive result can always be driven by the large effect from the first study only, even if it is unbiased. Thus, this does not meet the requirement of replication or validation, which implies that the same evidence is shown in independent studies.

WHAT IF REPLICATION OR VALIDATION FAILS?
Although a replication lessens the likelihood of false positives, it does not totally eliminate it, so that it does not necessarily mean that the association is true. However, we worry more if a genetic association cannot be replicated. Certainly, this might first hint at a false positive result in the first report. This might be explainable by statistical chance; for instance, with a local significance level of 5%, a power of 90% and 5 out of 500 000 SNPs being truly associated, one out of two SNPs reaching significance represents a false positive [36]. Also, a number of biases such as population stratification can be the reason for a false positive finding. On the other hand, non-replication may be due to chance and to a false negative finding in the replication study. In addition to looking into the possibly insufficient power and flaws or biases of the replication study, it is also recommended to investigate a possible heterogeneity between the study samples [37]. Phrased differently, one might have attempted replication through a sample that was different enough to explain genetic differences.

If a genetic association cannot be validated, there is even more room for speculation [24]. Here, there might be genuine differences between the populations under study regarding, for instance, ethnic background, phenotype definition or exposure to other risk factors. Therefore, non-validation might even provide interesting clues about the underlying genetic architecture [38].

In conclusion, I would like to make the following pleas: first, the described terms should be used consistently and precisely. In this line, it is helpful for the interpretation of association results when it has clearly been stated whether replication or validation was attempted. Secondly, the design of a study needs to match the study purpose. For example, for a replication study, the power calculation needs to take into account the likely upward bias in the first study. Finally and most generally, adhering to the recommendations [25, 39] will help investigators to avoid bias and readers to evaluate reports immensely.

Key point
- Replication and validation play a prominent role in the process of identifying genetic variants responsible for complex diseases. To correctly perform and interpret the analyses, it is necessary to use the related terms consistently and precisely. Adhering to general recommendations will help both investigators to avoid bias and readers to evaluate reports.

References


