Genetic factors appear to be important in multifactorial stroke pathogenesis,1,2 but the underlying molecular basis remains uncertain. Over the last decade, increasing numbers of studies investigating associations between DNA sequence variants and stroke have been published, and this trend is likely to accelerate.3 Increasingly rapid and cheaper genotyping technologies are becoming available, allowing large numbers of polymorphisms to be determined. As a consequence, the dimensions of studies and complexity of the data will increase.

There have been many reports of “significant” associations, but the majority have not been replicated. This has raised concerns about association studies and complex genetics in general. Because of such concerns, a number of journals have proposed guidelines for reporting genetic association studies.4,5 Stroke does not have such formal guidelines. Nevertheless, there are a number of methodological aspects that are important in planning and interpreting such studies in stroke. This article addresses these issues, provides recommendations on how to optimize study design, and suggests a set of standard criteria for genetic association studies in stroke.

Defining the Target Phenotype
The biological mechanisms underlying specific stroke subtypes differ substantially, and specific mutations or polymorphisms may contribute selectively to the risk of a particular stroke subtype.6,7 The classification systems for categorizing stroke subtypes vary between studies.8 The most widely used systems are the Trial of Org 10172 in Acute Stroke Treatment (TOAST) and Oxfordshire Community Stroke Project (OSCP) system. The OSCP system uses clinical examination rather than investigations to classify into clinical syndromes; it is not a pathophysiological classification and is therefore less suited to investigating risk factors for specific subtypes. Pathophysiological classification systems such as TOAST are more appropriate, although their accuracy is dependent on the quality of the clinical information and the experience of the raters. Therefore, the proportion of patients receiving a computed tomography (CT) scan, MRI, imaging of the cerebral vasculature, echocardiography, and other investigations is important. Using predefined phenotyping protocols and central rating by specially trained raters9 improves the quality of stroke categorization and helps to identify biologically meaningful and reproducible stroke subtypes.

Certain genes, such as those influencing neuronal responses to injury, may predispose to all forms of stroke. Therefore, looking for genetic associations with stroke in general remains a justifiable approach and has successfully been used in the past to identify gene loci10 and genes implicated in stroke risk.11 Nevertheless, accurate phenotyping and performing separate analyses according to stroke subtypes are essential. Focusing on particular stroke subtypes...
Association studies compare the frequency of specific DNA sequence variants (alleles) in groups of individuals in a case-control design. An allele is said to be associated with the disease if its frequency differs between cases and controls more than would be predicted by chance. For complex traits, association studies are more efficient than linkage studies because they have greater statistical power to detect several genes of small effect. However, they also pose important methodological challenges.

An important consideration is the selection of appropriate controls. This must be from a genetically similar population to avoid the introduction of bias attributable to population stratification. The prevalence of some potential stroke risk polymorphism varies markedly between ethnic groups, making matching cases and controls for ethnicity important. Defining ethnicity may be difficult in populations with marked admixture, and some groups recommend recording ethnicity and geographic origin of parents as well as cases. However, even when efforts are made to match for ethnicity, unrecognized population structure can cause false associations and failure to detect genuine associations. Statistical methods have been proposed to account for population structure so that association studies can proceed even when population stratification is present. One way to avoid population stratification is to use family-based controls. The transmission disequilibrium test (TDT) was developed to match cases and controls on their immediate genetic background: the parents. In late-onset conditions such as stroke, parents are often not alive, making TDT unrealistic. The sibling TDT, which requires sibs to be recruited, is potentially feasible in stroke but is still a major undertaking.

Another challenge lies in the selection of target genes and sequence variants. Up to now, association studies have been limited to the investigation of candidate genes selected because of a priori hypotheses about their etiological role in a disease. Thus, these studies depend on the ability to predict functional candidate genes and polymorphisms. There are strategies to optimize and prioritize their selection. Candidate genes may be selected from regions that have been identified through genomewide scans. Alternatively, genes may be prioritized because they are known to encode for specific proteins related to the disease process. Criteria for the selection of specific polymorphisms include the type of polymorphism and its frequency in the population. As a rule, alleles with frequencies of ≥5% are necessary. Lower allele frequencies pose a challenge to sample size.

Recently, investigators have begun performing large-scale association studies involving multiple candidate genes and polymorphisms at a time. This can introduce the problem of spurious associations attributable to multiple hypothesis testing. In contrast to candidate gene studies, genomewide association studies use markers that are evenly spaced throughout the genome, without regard to their function or location in a particular gene. Therefore, no previous hypothesis is required. This approach has been shown to be feasible but poses significant challenges in terms of sample sizes, genotyping resources, and data analysis.
Data Analysis and Statistical Issues

So far, most association studies have been limited to single polymorphisms (single-nucleotide polymorphisms, repeat polymorphisms, insertions, or deletions). More recently, studies have looked at the joint effects of single markers on a haplotype level. Haplotypes are a set of alleles that are physically close to each other on the DNA strand and are thus inherited as a unit, (ie, a set of alleles that are in strong linkage disequilibrium [LD]). Association between a polymorphism or haplotype and a trait, such as stroke, does not necessarily imply causality. Instead, the polymorphism/haplotype under investigation may be in LD with the causative sequence variant (which may be unknown). In this case, the associated polymorphism or haplotype represents a marker for the causative sequence variant, and more detailed studies are required to identify the causative sequence alteration.

Increasing the number of markers or performing subgroup analyses on stroke subtypes inevitably leads to the problem of multiple testing and the possibility of false-positive associations if correction for multiple testing is not carried out. The loss of statistical power using traditional methods such as the Bonferroni procedure has led many authors to neglect their interest in replication or phenotyping errors), and inappropriate statistical methods.

True variability in association between different populations may exist because of differences in: (1) the frequency of disease-causing alleles, (2) the pattern of association between disease-causing alleles and single markers or haplotypes under investigation, or (3) interacting genetic or environmental factors. Thus, nonreplication does not necessarily imply lack of causality but might point to the need for additional studies. In view of these difficulties, it has been proposed recently that replication studies should undertake more extensive genotyping, comprising all genetic variation within a candidate gene. However, it is still unclear how extensively one needs to analyze a gene to exclude association in another population.

In cases of positive replication, the first study to report an association often finds a stronger effect than seen in subsequent studies, particularly if the initial results were obtained through genomewide scans. Therefore, replication studies should be powered to detect effect sizes that are smaller than the initial effect size reported. Meta-analyses may then assist in estimating the population-wide effects of genetic risk factors, although careful account needs to be taken of the effect of publication bias.

Suggestions for Standard Criteria

This discussion has documented the methodological considerations required to increase the likelihood of a valid and replicable study. Major points are listed in Table 2. Few studies will yet meet all criteria; in particular, current sample sizes are smaller than those optimal. We propose standard criteria for implementation in stroke studies to ensure quality and enable replication studies and meta-analyses:

- The a priori hypotheses of the study should be stated upfront.
- Power calculations should be provided to demonstrate that the study is sufficiently powered to test all hypotheses of the study.
- For replication studies, the sample should be large enough to detect or exclude an effect of somewhat smaller size than in the initial study.
- Any overlap with previously published studies (same polymorphisms or polymorphisms in LD with polymorphisms from a previous study in the same sample of patients or controls) should be indicated.

### Table 1. Sample Sizes (cases) as a Function of Minor Allele Frequency and Genotypic Relative Risk (multiplicative model)

<table>
<thead>
<tr>
<th>Minor Allele Frequency</th>
<th>0.01</th>
<th>0.05</th>
<th>0.10</th>
<th>0.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relative Risk</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td>&gt;20000</td>
<td>4587</td>
<td>2444</td>
<td>1401</td>
</tr>
<tr>
<td>1.5</td>
<td>3981</td>
<td>845</td>
<td>456</td>
<td>268</td>
</tr>
<tr>
<td>2.0</td>
<td>1199</td>
<td>259</td>
<td>142</td>
<td>87</td>
</tr>
<tr>
<td>3.0</td>
<td>403</td>
<td>89</td>
<td>51</td>
<td>33</td>
</tr>
</tbody>
</table>

Sample sizes are calculated for an equal No. of controls, a power of 0.8 and a type 1 error \( P \) value of 0.05. Calculations were done using the Genetic Power Calculator.
TABLE 2. Methodological Aspects That Enhance the Significance of an Association Finding

<table>
<thead>
<tr>
<th>Aspect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large sample size</td>
</tr>
<tr>
<td>Rigorous phenotypic assessment in patients and in controls</td>
</tr>
<tr>
<td>Low genotyping error rate</td>
</tr>
<tr>
<td>Genomic controls or other techniques to account for population stratification</td>
</tr>
<tr>
<td>Low P value (corrected for multiple testing)</td>
</tr>
<tr>
<td>Odds ratio or attributable risk is high</td>
</tr>
<tr>
<td>Replication in ≥1 independent populations</td>
</tr>
<tr>
<td>Gene makes biological sense</td>
</tr>
<tr>
<td>Allele affects the gene in a meaningful way*</td>
</tr>
<tr>
<td>Presence of a gene–dose response relationship*</td>
</tr>
<tr>
<td>Functional data demonstrating a biological effect of the at-risk allele</td>
</tr>
</tbody>
</table>

*It may be difficult to determine whether an allele affects a gene in a meaningful way. If the functional relevance of the polymorphism is unknown, additional experiments examining its biological effect should be considered.

- A description of how cases and controls were sampled should be provided. Controls should be ethnically matched and derived from the same source population as cases. They should allow for the possibility of using other methods to account for population stratification besides matching for ethnicity.
- The phenotyping protocol should be specified. In particular, the proportion of patients having CT/MRI, imaging of large extracranial and intracranial arteries (eg, by ultrasound or magnetic resonance angiography), and echocardiography should be detailed.
- Phenotyping of patients should be done according to a standard pathophysiological classification such as the TOAST system and ideally by a central rater particularly for multicenter studies.
- When controlling for conventional risk factors, the definitions and methods used to determine presence or absence of a risk factor should be specified.
- Gene–environment interactions should be sought, but studies need to be powered to allow these analyses.
- Investigators involved in genotyping should be blinded with regard to phenotypes.8
- The genotyping error should be stated and low.
- It should be indicated whether marker genotypes were in Hardy–Weinberg equilibrium (HWE) in cases and controls. Strong deviation from HWE (especially in controls) points to the possibility for genotyping errors.40
- In an initial association study, data on major stroke subtypes should be presented whenever possible.
- Genotype frequencies as well as allele frequencies should be presented for all groups analyzed.
- The relative risks (odds ratios and 95% CIs) or attributable risks should be indicated.
- Uncorrected P values should be reported to facilitate meta-analyses, but adjustment for multiple testing should still be performed.
- Authors should check for a possible gene–dose effect. If present, the results are usually more convincing. If there is no such effect, it should be discussed whether the mechanism for a dominant or recessive effect is explicable from a biological viewpoint.
- Authors should provide a complete listing of all phenotypes, single markers, and haplotypes analyzed.
- By checking at these criteria during the planning phase of a study, more consistent quality will be obtained.

Conclusion

Association studies remain a powerful tool to identify genetic risk factors for stroke. However, the results from initial reports need to be treated with caution, and replication is essential. Properly conducted negative studies are of interest to the field, especially when addressing previously published claims. To minimize publication bias, journals should be willing to publish high-quality replication studies regardless of whether the results are positive or negative. Efforts should not be wasted on poorly designed studies. Insufficient sample size remains a key issue in most association studies, and collaborative multicenter efforts are needed to solve this problem.

Acknowledgments

We thank Bertram Müller-Myyhosk for helpful discussions.

References

Martin Dichgans and Hugh S. Markus

*Stroke*. 2005;36:2027-2031; originally published online July 28, 2005; doi: 10.1161/01.STR.0000177498.21594.9e

*Stroke* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2005 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/36/9/2027

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in *Stroke* can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to *Stroke* is online at:
http://stroke.ahajournals.org//subscriptions/