Genetic Association Studies in Ischaemic Stroke

Replication Failure and Prospects

D. Martijn O. Pruissen^a L. Jaap Kappelle^a Frits R. Rosendaal^{c–e} Ale Algra^{a–c}

^aDepartment of Neurology, Rudolf Magnus Institute of Neuroscience, and ^bJulius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, Departments of ^cClinical Epidemiology,

^dThrombosis and Haemostasis and ^eEindhoven Laboratory for Experimental Vascular Medicine, Leiden University Medical Center, Leiden, The Netherlands

Key Words

Ischaemic stroke · Replication failure · Genetic association

Abstract

Although hundreds of genetic association studies of ischaemic stroke have been published, the failure to replicate associations has led to scepticism about their findings. Possible explanations for this failure are: (1) a false-positive association in the initial study; (2) a false-negative association in a replication study; (3) methodological differences (e.g. study populations or study designs). We review underlying causes for replication failure, such as small sample size, multiple testing and publication bias, and methods to deal with these problems. We also make suggestions about the design of genetic association studies in ischaemic stroke with regard to stroke subtype classification, candidate pathways, subgroups, intermediate phenotypes and potential clinical impact.

Copyright © 2009 S. Karger AG, Basel

Introduction

Ischaemic stroke is a complex (i.e. multifactorial) disease caused by a combination of genetic and environmental factors [1]. The heritability of stroke is in part due to

the heritability of hypertension and large vessel atherosclerosis [2]. Genetic research makes use of two main approaches to identify gene variants related to disease: linkage and association studies. Linkage relies on the fact that the recombination of gene sequences is more likely to occur between loci that are further apart on the genome than between loci closer together. Thus, the coinheritance of two distant loci that occurs more often than would be expected by chance indicates linkage of a disease to a locus. Linkage analysis in family samples makes it possible to identify distant genetic markers in families that would otherwise be missed in general population samples because the recombination of distant loci occurs only sporadically, and not within a few generations. The polygenic nature of common complex diseases complicates the linkage approach. The second approach is to study the association between allelic variation within specific genes and cerebral ischaemia, for instance, by comparing genotype frequencies in stroke patients and healthy controls. A gene variant that is more common in patients than in controls may cause stroke or be located close to the true causal variant. The use of association studies in complex diseases has become popular as a result of lower power in family studies and advances in high throughput genotyping.

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2009 S. Karger AG, Basel
1015–9770/09/0273–0290\$26.00/0

Accessible online at:
www.karger.com/ced

Ale Algra, MD
University Medical Center Utrecht, Julius Center for Health Sciences and Primary Care
Room STR 6.131, PO Box 85500
NL–3508 GA Utrecht (The Netherlands)
Tel. +31 88 755 9350, Fax +31 88 755 5485, E-Mail a.algra@umcutrecht.nl

Table 1. Causes for replication failure in genetic association studies

False-positive association	multiple testing small sample size population stratification publication bias
False-negative association	small sample size accounting for multiple testing genotyping error phenotyping error superfluous adjustment for confounding
Differences between study populations	different ethnicity phenotyping heterogeneity general or genetically isolated population
Different study designs and methods	family members or unrelated individuals case-control or cohort study genotyping heterogeneity

Replication Failure

The identification of PDE4D (phosphodiesterase 4D) as a novel susceptibility gene for ischaemic stroke raised expectations [3], but subsequent studies tempered this enthusiasm: 5 studies replicated this association, but 4 did not [4]. The impact of PDE4D on ischaemic stroke is currently unclear. A review of 166 reported genetic associations in various common diseases identified only 6 associations that were consistently replicated [5]. This failure to replicate findings has led to scepticism. In principle, there are 3 possible explanations for replication failure: (1) false-positive association in the initial study; (2) false-negative association in a replication study; (3) methodological differences, such as different study populations or study designs. Table 1 gives an overview of the causes of replication failure.

False-Positive Associations

Multiple testing, including subgroup analyses or additional analyses with alternative outcome measures or different modes of inheritance, is an important cause of false-positive associations. Whether and how one should account for multiple testing in genetic association studies is a matter of debate. The Bonferroni correction method proposes lowering the *p* value threshold for statistical significance to 0.05 divided by the number of tests per-

formed. However, the Bonferroni correction is considered to be overly conservative in genetic studies [6]. Promising methods to account for multiple testing are permutation and the false-discovery rate [7]. The false-discovery rate expresses false-positive test results as a proportion of positive test results, and is an equivalent of the false-positive rate in diagnostic studies.

False-positive associations also arise with small samples. A meta-analysis of 370 studies of 36 genetic associations found a smaller sample size in the initial study to be a major predictor of replication failure [8]. Subgroup analyses may also induce false-positive associations because the sample size is reduced.

Population stratification, also known as admixture, may also give rise to false-positive associations: if an ethnic subgroup in a study population has a higher disease incidence, any variant that is more common in that subgroup than in other parts of the population may be falsely associated with the disease. However, the limited empirical support for undetected population stratification suggests that population stratification is a minor cause of false-positive associations [9]. Genetic associations with ischaemic stroke were similar among study populations of Asian and European descent in a large meta-analysis, suggesting that ethnicity has little effect on genetic associations, at least in ischaemic stroke [10]. Methods to deal with population stratification include: matching by family, self-reported ancestry, genotyping ancestry-informative variants and statistical methods such as genomic control [7]. Genomic control exploits the fact that population stratification generates ‘overdispersion’ of statistics used to assess association. By testing multiple gene variants throughout the genome, only some of which are relevant to the disease of interest, the degree of overdispersion can be estimated and taken into account.

Another potential source of false-positive associations is publication bias. However, a meta-analysis of 301 studies of 25 genetic associations indicated that publication bias is limited in genetic association studies [11]. In an effort to reduce publication bias, several journals have proposed that such studies should adhere to standard criteria [12, 13]. Journals should also be willing to publish good quality negative studies and scientists should be encouraged to share their unpublished negative findings. The extent of publication bias can be investigated in meta-analyses by showing an asymmetrical distribution of study results in relation to sample sizes (visualized in a funnel plot).

False-Negative Associations

A small sample size is an important cause of false-negative associations, and analyses within small subgroups are also prone to miss true associations. Meta-analyses of genetic association studies in ischaemic stroke report effect sizes (relative risks) below 1.5 for associated variants [14], but most studies to date were not powered to detect such small effect sizes. A meta-analysis of 301 studies of 25 genetic associations found an excess of replications not explained by publication bias in 11 of the 25 associations. Underpowered studies with false-negative results probably contributed to inconsistent replication because pooled analysis of all studies yielded consistent replication for 8 out of 11 gene variants while 117 of the 154 single studies reported negative results [11]. This emphasizes the importance of performing power calculations.

The methods used to account for multiple testing also lead to false-negative associations, especially in studies testing a large number of variants. Each statistical method involves a tradeoff between detecting false-positive findings and preserving true associations. The false-discovery rate approach substantially increases the power to detect true associations at the cost of only a few false-positive associations [15].

Another source of false-negative associations is misclassification of the genotype or disease, which can be minimized by using standardized procedures, confirming or replicating the genotype and estimating the Hardy-Weinberg equilibrium [16].

Unnecessary adjustment for confounding can theoretically lead to both false-negative or false-positive findings because adjustment can increase or decrease an odds ratio. However, adjustment always leads to a wider confidence interval due to an increased mathematical standard error, which makes false-negative associations more likely. Environmental factors (e.g. antithrombotic treatment) are unlikely to be associated with genotype frequencies and to confound genetic associations [17].

Methodological Differences

Methodological differences, e.g. differences in study population or study design, are a third cause of replication failure. Ethnicity, phenotype and genotyping methods can differ between studies. Such differences can also affect the validity of meta-analyses and multicentre studies. However, as mentioned above, ethnicity probably has

a limited effect on genetic differences in disease occurrence [10].

Phenotype heterogeneity in study populations can arise if different inclusion and exclusion criteria are used for a specific disease; however, even if inclusion criteria are identical, aetiological heterogeneity can still exist [18]. Gene-environment interactions can also contribute to phenotype heterogeneity. For example, the use of oral contraceptives increases the risk of ischaemic stroke among carriers of a coagulation factor XIII gene variant [19]. Thus, differences in oral contraceptive use between studies can contribute to a difference in reported genetic risks in these studies.

Differences in study design and methods can also contribute to variance in findings between association studies. Family-based designs are less feasible in ischaemic stroke because the parents of stroke patients are often dead. A meta-analysis of 93 genetic associations from both family-based and unrelated case-control studies found the overall odds ratio of an association to be similar in the two types of study [20]. Both case-control and prospective cohort designs are suitable for association studies and provide equally valid risk estimates, provided that case and control sampling are performed properly. Cohort studies allow for absolute risk estimates and simultaneous investigation of different outcomes, but are more expensive and time consuming. Lastly, genotyping techniques can vary among laboratories and introduce differential genetic effects between studies [21].

Two Different Hypotheses on Genetic Variation in Common Disease

The 'common disease common variants' hypothesis proposes that most of the genetic risk for common diseases is due to common variants [22]. A meta-analysis and a simulation study provide data to support the presence of common susceptibility variants with small effects [11, 23]. Although common variants are unlikely to contribute to risk stratification for individual patients, their impact on common disease may still be important at a population level because of their wide distribution.

Alternatively, multiple rare variants may be largely responsible for the risk for common diseases [24, 25]. Rare variants will probably not be detected in genetic association studies, but will instead require candidate gene sequencing [25]. The impact of rare variants on ischaemic stroke is unknown and merits further study.

Genome-Wide Association and Candidate Gene Approaches

The scope of genetic association studies is shifting from the candidate gene approach to coverage of the entire genome, and several genome-wide association (GWA) studies in common diseases have been published recently. The scale of genome-wide coverage continues to grow, and even nucleotide sequencing of the entire genome can be expected in the future. Indeed, patient recruitment to acquire large sample sizes will probably lag behind technical advances. Copy number variation of DNA sequences is another promising type of genomic variation, and has been detected in 12% of the genome [26]. A first study of copy number variation in ischaemic stroke did not reveal strong effects, but the sample size was small [27].

GWA studies can lead to the identification of new susceptibility genes and pathways, such as the association between potassium transport genes and ischaemic stroke [28]. Another advantage of this approach is 'hypothesis-free' association testing, where there is no need to select candidate genes based on linkage, functionality, previous association or personal preference. GWA also enables correction for population stratification. However, these studies are limited by false-negative associations (due to accounting for multiple testing) and an inability to identify rare variants.

Advances in genome-wide variation coverage do not invalidate the candidate gene approach. Hypothesis testing continues to be a valid scientific method to target candidate genes and to optimally apply limited resources. Although population stratification and publication bias are less likely to affect GWA studies, both GWA studies and candidate gene studies are susceptible to other causes of replication failure, such as small sample size and phenotyping error. Moreover, accounting for multiple testing can cause more false-negative findings in GWA studies as compared with candidate gene studies.

Regardless of the genetic approach used, a solid study design, including adequate sample sizes and adequate phenotyping, remains essential for producing reliable data. Multicentre studies and meta-analyses of individual patient data may be needed to achieve sample sizes large enough to detect small effect sizes, gene-gene interactions and gene-environment interactions. Moreover, it is important to study all the variants of a particular gene before excluding its involvement in a specific disease. If rare variants are thought to be involved, gene sequencing is indicated. Independent replication and functionality

studies remain essential to confirm new associations identified in copy number variation, GWA or candidate gene studies.

Future Studies in Ischaemic Stroke

Researchers face several choices when designing genetic studies of patients with ischaemic stroke with regard to subtyping, candidate pathways, subgroups, intermediate phenotypes and potential clinical impact.

Several genetic variants have been specifically associated to large or small vessel disease, for example, the factor XIII Val34Leu variant and small vessel disease [29]. Since most studies performed to date were underpowered for subtype analyses [30], it is essential to recruit a large enough sample and to clearly define subtype classifications. In this respect, a classification based on clinical features and neuroimaging findings alone may yield more valid subtypes than classification systems that include vascular risk factors such as carotid stenosis [31].

Promising candidate pathways include inflammation [32] and hypertension [33]. Pathway-based analysis of GWA studies may provide more meaningful information than analysis based on the lowest p values [34]. Larger genetic effects may be detected in subgroups with high heritability. For example, subgroups of younger patients or patients with hypertension are interesting because a family history of stroke has been associated with younger age and previous hypertension [33].

Intermediate phenotypes such as carotid intima-media thickness or white matter lesions are often more homogeneous than ischaemic stroke itself, and quantitative traits allow for more powerful statistical analysis than dichotomous outcome measures [33].

Researchers should also consider the aim of their study. Given the small effects of most variants on ischaemic stroke, genetic testing is as yet not suitable for diagnostic testing, identifying those at risk or providing a prognosis [35]. In clinical practice, treatment of acute ischaemic stroke needs to be initiated long before the results of genetic testing become available. However, genetic testing may give insight into the aetiology of stroke, and may be promising for predicting the response to treatment (pharmacogenetics). A simulation study has shown genotype-specific antithrombotic treatment to be theoretically feasible [35].

Conclusions

Genetic association studies in ischaemic stroke have yielded conflicting results. Important causes of replication failure include small sample size, multiple testing and publication bias. Most causes of replication failure can be dealt with by paying attention to the study design. Genome wide association, copy number variation and candidate gene studies can all contribute to unravelling

the genetic susceptibility to ischaemic stroke. Confirmation of genetic associations in replication and functionality studies remains essential. Researchers can choose to select a specific stroke subtype, a specific candidate pathway or a specific patient subgroup. Aetiological studies and pharmacogenetics are more promising than studies aimed at identifying diagnostic or prognostic genetic factors.

References

- Flossmann E, Schulz UG, Rothwell PM: Systematic review of methods and results of studies of the genetic epidemiology of ischemic stroke. *Stroke* 2004;35:212–227.
- Flossmann E, Schulz UG, Rothwell PM: Potential confounding by intermediate phenotypes in studies of the genetics of ischaemic stroke. *Cerebrovasc Dis* 2005;19:1–10.
- Gretarsdottir S, Thorleifsson G, Reynisdottir ST, et al: The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nat Genet* 2003;35:131–138.
- Rosand J, Bayley N, Rost N, de Bakker PI: Many hypotheses but no replication for the association between PDE4D and stroke. *Nat Genet* 2006;38:1091–1092.
- Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K: A comprehensive review of genetic association studies. *Genet Med* 2002;4:45–61.
- Cordell HJ, Clayton DG: Genetic association studies. *Lancet* 2005;366:1121–1131.
- Balding DJ: A tutorial on statistical methods for population association studies. *Nat Rev Genet* 2006;7:781–791.
- Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG: Replication validity of genetic association studies. *Nat Genet* 2001;29:306–309.
- Cardon LR, Palmer LJ: Population stratification and spurious allelic association. *Lancet* 2003;361:598–604.
- Ariyaratnam R, Casas JP, Whittaker J, Smeeth L, Hingorani AD, Sharma P: Genetics of ischaemic stroke among persons of non-European descent: a meta-analysis of eight genes involving approximately 32,500 individuals. *PLoS Med* 2007;4:e131.
- Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN: Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 2003;33:177–182.
- Chanock SJ, Manolio T, Boehnke M, et al: Replicating genotype-phenotype associations. *Nature* 2007;447:655–660.
- Dichgans M, Markus HS: Genetic association studies in stroke: methodological issues and proposed standard criteria. *Stroke* 2005;36:2027–2031.
- Casas JP, Hingorani AD, Bautista LE, Sharma P: Meta-analysis of genetic studies in ischemic stroke: thirty-two genes involving approximately 18,000 cases and 58,000 controls. *Arch Neurol* 2004;61:1652–1661.
- Dupuis J, O'Donnell CJ: Interpreting results of large-scale genetic association studies: separating gold from fool's gold. *JAMA* 2007;297:529–531.
- Pompanon F, Bonin A, Bellemain E, Taberlet P: Genotyping errors: causes, consequences and solutions. *Nat Rev Genet* 2005;6:847–859.
- Colhoun HM, McKeigue PM, Davey SG: Problems of reporting genetic associations with complex outcomes. *Lancet* 2003;361:865–872.
- Healy DG: Case-control studies in the genomic era: a clinician's guide. *Lancet Neurol* 2006;5:701–707.
- Pruissen DM, Slooter AJ, Rosendaal FR, van der Graaf Y, Algra A: Coagulation factor XIII gene variation, oral contraceptives, and risk of ischemic stroke. *Blood* 2008;111:1282–1286.
- Evangelou E, Trikalinos TA, Salanti G, Ioannidis JP: Family-based versus unrelated case-control designs for genetic associations. *PLoS Genet* 2006;2:e123.
- Clayton DG, Walker NM, Smyth DJ, et al: Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* 2005;37:1243–1246.
- Lander ES: The new genomics: global views of biology. *Science* 1996;274:536–539.
- Peng B, Kimmel M: Simulations provide support for the common disease-common variant hypothesis. *Genetics* 2007;175:763–776.
- Pritchard JK: Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 2001;69:124–137.
- Reitsma PH, Rosendaal FR: Past and future of genetic research in thrombosis. *J Thromb Haemost* 2007;5(suppl 1):264–269.
- Redon R, Ishikawa S, Fitch KR, et al: Global variation in copy number in the human genome. *Nature* 2006;444:444–454.
- Matarin M, Simon-Sanchez J, Fung HC, et al: Structural genomic variation in ischemic stroke. *Neurogenetics* 2008;9:101–108.
- Matarin M, Brown WM, Scholz S, et al: A genome-wide genotyping study in patients with ischaemic stroke: initial analysis and data release. *Lancet Neurol* 2007;6:414–420.
- Slowik A, Dziedzic T, Pera J, Figlewicz DA, Szczudlik A: Coagulation factor XIII Val34Leu polymorphism in patients with small vessel disease or primary intracerebral hemorrhage. *Cerebrovasc Dis* 2005;19:165–170.
- Pruissen DM, Kappelle LJ, Rosendaal FR, Algra A: Prothrombotic gene variation in patients with large and small vessel disease. *Neuroepidemiology* 2008;31:89–92.
- Jackson C, Sudlow C: Are lacunar strokes really different? A systematic review of differences in risk factor profiles between lacunar and nonlacunar infarcts. *Stroke* 2005;36:891–901.
- Libby P: Inflammation in atherosclerosis. *Nature* 2002;420:868–874.
- Schulz UG, Flossmann E, Rothwell PM: Heritability of ischemic stroke in relation to age, vascular risk factors, and subtypes of incident stroke in population-based studies. *Stroke* 2004;35:819–824.
- Wang K, Li M, Bucan M: Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* 2007;81:1278–1283.
- van Rijn MJ, van Duijn CM, Slooter AJ: Impact of genetic testing on complex diseases. *Eur J Epidemiol* 2005;20:383–388.