Published in final edited form as:
*Stroke*. 2010 April ; 41(4): 579–580. doi:10.1161/STROKEAHA.109.576769.

# Stroke genome-wide association studies: the large numbers imperative

**James F. Meschia, MD**
Department of Neurology, Mayo Clinic, 4500 San Pablo Road, Jacksonville, Florida 32224

James F. Meschia: meschia.james@mayo.edu

With a swiftly moving, highly technical field, it helps to have a reliable guide. In this issue of *Stroke*, Lanktree and colleagues provide a timely overview of genomic analysis applied to stroke.[1] Readers will find a clear and concise synopsis of single-nucleotide polymorphisms (SNPs), copy number variations, listings of the strengths and limitations of genome wide association studies (GWAS) as well as a distillation of the findings from GWAS performed in 6 cohorts (5 of which were ischemic stroke cohorts). The review also covers a topic not usually found in clinical reviews, namely techniques to visually display quantitative information. Excellence in statistical graphics should, among other things, avoid distorting what the data have to say, present many numbers in a small space, and make large data sets coherent.[2] The Manhattan and Q-Q plots are excellent examples of statistical graphics that have become invaluable for interpreting GWAS results.

Along with visualization comes interpretation of data in the context of GWAS. Clinical investigators are well aware of the problem of multiple testing from such settings as interim and subgroup analyses in clinical trials, which can lead to wildly spurious conclusions, such as concluding that aspirin only helped individuals of certain astrological signs in the second International Study of Infarct Survival (ISIS-2).[3] GWAS simply escalates the problem of multiple testing by orders of magnitude. Because SNP-based GWAS test hundreds of thousands of SNPs per subject, a significant association requires a very low p-value. The Wellcome Trust Case Control Consortium (WTCCC) used $P < 5 \times 10^{-7}$ as the cut-off for genome-wide significance.[4] Others have chosen to pre-specify genome-wide significance with greater stringency at $P < 5 \times 10^{-8}$, corresponding to the 5% significance level adjusting for the number of independent tests estimated in HapMap for individuals of European ancestry.[5]

How has stroke performed in the significance "high-jump" competition? Not particularly well. Of the loci on chromosomes 4, 11, 12, 14, 16 and 22 with associations with ischemic stroke reported to have reached genome-wide significance, none has been replicated across studies. 1 This may be the result of biases, including the so-called winner's curse.[6] Differential effects under different exposures (e.g., tobacco smoking), association with correlated phenotypes (e.g. atrial fibrillation or diabetes mellitus), differences in ascertainment schemes, genotype misclassification, or marker polymorphism in variable linkage disequilibrium with the causative variant across populations may alternatively explain the heterogeneity.[7]

There is what might be called a large numbers imperative when it comes to GWAS of a complex disorder like ischemic stroke. It is unlikely that any single study, even a multi-center study, will ever achieve the sample size necessary to yield a credible result. An uncommon level of world-wide collaboration must emerge. To put things in perspective, the Venice criteria state that to earn an 'A' rating in terms of amount of evidence in a genetic association study requires a sample size exceeding 1,000 combined cases and controls (assuming a 1:1 ratio) in the least common genetic group of interest.[8] The less common the risk allele is; the greater the sample size requirement.

The stroke community is striving to meet the large numbers imperative. The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE), which consists of five community-based cohorts, recently reported the results of its ischemic stroke GWAS.[9] The WTCCC is currently conducting a three-stage GWAS of ischemic stroke under the leadership of Dr. Hugh Markus. Genome-wide genotyping has been completed for approximately four thousand cases in stage one. The Ischemic Stroke Genetics Consortium (ISGC), a loose federation of investigators, first met in Boston, MA on April 28, 2007. The ISGC now involves 73 investigators across 16 countries.10 The ISGC initially focused on large-scale candidate gene replication studies.11 It is currently turning its attention to organizing a genome-wide association study.

To meet the large numbers imperative, meta-analysis is almost unavoidable. There are reasons to believe that an appropriately powered meta-analysis in ischemic stroke is feasible. First, SNP imputation techniques have evolved such that results from studies using diverse gene chip platforms can be pooled without losing statistical power.[12] This is important because it can save on the considerable expense of re-genotyping. Second, under certain conditions, meta-analysis of summary results can be as efficient statistically as joint analysis of individual participant data (also known as mega-analysis).[13] This is important because many investigators might be unwilling or unable to share data at the individual participant level due to privacy concerns.

The methodological advantages of cohort studies are well known.14 However, the large numbers imperative is not likely to be satisfied by compiling incident cases alone. CHARGE consortium included fewer than 1,200 incident cases of ischemic stroke.[9] Stroke centers can more efficiently generate far larger numbers of cases of ischemic stroke, but concerns have been raised regarding the validity of case-control ischemic stroke GWAS. One concern about such studies relates to the possibility of Neyman (prevalence-incidence) bias.[15] Such bias could be problematic if the genetic determinants of ischemic stroke are also those that increase risk of death from stroke and if large percentages of patients die before they have the opportunity to donate DNA. To date no genetic variant has been shown to be an unequivocal determinant of both ischemic stroke and risk of death from stroke. Further, the proportion of cases that die before being screened for enrollment in a genetic association study is likely small if recruitment occurs in the setting of an inpatient stroke service, particularly if surrogate consent is permitted.[16]

Finally, sample size and phenotypic heterogeneity tend to be inversely related. Restrict the phenotype enough, and stroke begins to look more like an orphan disease than a common disease. If studies include phenotypically diverse stroke, it would behoove investigators to characterize strokes among the cases in great detail. Rather than forcing cases into a limited set of mutually exclusive categories, it might be more productive to capture within phenotypic data sets the results of studies that were done to evaluate cardiac and cerebrovascular status along with the results of brain imaging. Semiautomated approaches like the web-based Causative Classification of Stroke system can help to systematically structure diverse clinical data sets across multiple studies.[17] However, restructuring clinical data sets with source documentation is laborious, time consuming, and expensive. Other approaches to phenomics might be more efficient.

Funding agencies may also need to consider the value of new case recruitment under a uniform protocol that includes deep phenotyping.

As investigators strive to make definitive, consistently reproducible discoveries of the genetic determinants of ischemic stroke, one is likely to see GWAS of increasing sample size and increasing use of meta-analytic techniques.[18]

## References


1. Lanktree MB, Dichgans M, Hegele RA. Advances in genomic analysis of stroke: what have we learned and where are we headed? Stroke 2010;41:825–832. [PubMed: 20167918]
2. Tufte, E. The visual display of quantitative information. Cheshire: Graphics Press; 1983.
3. Sleight P. Debate: Subgroup analyses in clinical trials: Fun to look at - but don't believe them! Curr Control Trials Cardiovasc Med 2000;1:25–27. [PubMed: 11714402]
4. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–678. [PubMed: 17554300]
5. Benjamin EJ, Rice KM, Arking DE, Pfeufer A, van Noord C, Smith AV, Schnabel RB, Bis JC, Boerwinkle E, Sinner MF, Dehghan A, Lubitz SA, D'Agostino RB Sr, Lumley T, Ehret GB, Heeringa J, Aspelund T, Newton-Cheh C, Larson MG, Marciante KD, Soliman EZ, Rivadeneira F, Wang TJ, Eiriksdottir G, Levy D, Psaty BM, Li M, Chamberlain AM, Hofman A, Vasan RS, Harris TB, Rotter JI, Kao WH, Agarwal SK, Stricker BH, Wang K, Launer LJ, Smith NL, Chakravarti A, Uitterlinden AG, Wolf PA, Sotoodehnia N, Kottgen A, van Duijn CM, Meitinger T, Mueller M, Perz S, Steinbeck G, Wichmann HE, Lunetta KL, Heckbert SR, Gudnason V, Alonso A, Kaab S, Ellinor PT, Witteman JC. Variants in zfhx3 are associated with atrial fibrillation in individuals of european ancestry. Nat Genet 2009;41:879–881. [PubMed: 19597492]
6. Nakaoka H, Inoue I. Meta-analysis of genetic association studies: Methodologies, between-study heterogeneity and winner's curse. J Hum Genet 2009;54:615–623. [PubMed: 19851339]
7. Khoury MJ, Bertram L, Boffetta P, Butterworth AS, Chanock SJ, Dolan SM, Fortier I, Garcia-Closas M, Gwinn M, Higgins JP, Janssens AC, Ostell J, Owen RP, Pagon RA, Rebbeck TR, Rothman N, Bernstein JL, Burton PR, Campbell H, Chockalingam A, Furberg H, Little J, O'Brien TR, Seminara D, Vineis P, Winn DM, Yu W, Ioannidis JP. Genome-wide association studies, field synopses, and the development of the knowledge base on genetic variation and human diseases. Am J Epidemiol 2009;170:269–279. [PubMed: 19498075]
8. Ioannidis JP, Boffetta P, Little J, O'Brien TR, Uitterlinden AG, Vineis P, Balding DJ, Chokkalingam A, Dolan SM, Flanders WD, Higgins JP, McCarthy MI, McDermott DH, Page GP, Rebbeck TR, Seminara D, Khoury MJ. Assessment of cumulative evidence on genetic associations: Interim guidelines. Int J Epidemiol 2008;37:120–132. [PubMed: 17898028]
9. Ikram MA, Seshadri S, Bis JC, Fornage M, DeStefano AL, Aulchenko YS, Debette S, Lumley T, Folsom AR, van den Herik EG, Bos MJ, Beiser A, Cushman M, Launer LJ, Shahar E, Struchalin M, Du Y, Glazer NL, Rosamond WD, Rivadeneira F, Kelly-Hayes M, Lopez OL, Coresh J, Hofman A, DeCarli C, Heckbert SR, Koudstaal PJ, Yang Q, Smith NL, Kase CS, Rice K, Haritunians T, Roks G, de Kort PL, Taylor KD, de Lau LM, Oostra BA, Uitterlinden AG, Rotter JI, Boerwinkle E, Psaty BM, Mosley TH, van Duijn CM, Breteler MM, Longstreth WT Jr, Wolf PA. Genomewide association studies of stroke. N Engl J Med 2009;360:1718–1728. [PubMed: 19369658]
10. Stroke genetics. Retrieved on January 1, 2010 from www.strokegenetics.org
11. Gschwendtner A, Bevan S, Cole JW, Plourde A, Matarin M, Ross-Adams H, Meitinger T, Wichmann E, Mitchell BD, Furie K, Slowik A, Rich SS, Syme PD, MacLeod MJ, Meschia JF, Rosand J, Kittner SJ, Markus HS, Muller-Myhsok B, Dichgans M. Sequence variants on chromosome 9p21.3 confer risk for atherosclerotic stroke. Ann Neurol 2009;65:531–539. [PubMed: 19475673]
12. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 2007;39:906–913. [PubMed: 17572673]
13. Lin DY, Zeng D. Meta-analysis of genome-wide association studies: No efficiency gain in using individual participant data. Genet Epidemiol 34:60–66. [PubMed: 19847795]
14. Grimes DA, Schulz KF. Cohort studies: Marching towards outcomes. Lancet 2002;359:341–345. [PubMed: 11830217]
15. Grimes DA, Schulz KF. Bias and causal associations in observational research. Lancet 2002;359:248–252. [PubMed: 11812579]
16. Chen DT, Case LD, Brott TG, Brown RD Jr, Silliman SL, Meschia JF, Worrall BB. Impact of restricting enrollment in stroke genetics research to adults able to provide informed consent. Stroke 2008;39:831–837. [PubMed: 18258838]

17. Ay H, Benner T, Arsava EM, Furie KL, Singhal AB, Jensen MB, Ayata C, Towfighi A, Smith EE, Chong JY, Koroshetz WJ, Sorensen AG. A computerized algorithm for etiologic classification of ischemic stroke: The causative classification of stroke system. Stroke 2007;38:2979–2984. [PubMed: 17901381]

18. Zeggini E, Ioannidis JP. Meta-analysis in genome-wide association studies. Pharmacogenomics 2009;10:191–201. [PubMed: 19207020]