Predictive Genomics of Cardioembolic Stroke

Rachel Badovinac Ramoni, DMD, ScD; Blanca E. Himes, PhD; Michele M. Sale, PhD; Karen L. Furie, MD, MPH; Marco F. Ramoni, PhD

Abstract—Cardioembolic stroke is a complex disease resulting from the interaction of numerous factors. Using data from Genes Affecting Stroke Risk and Outcome Study (GASROS), we show that a multivariate predictive model built using Bayesian networks is able to achieve a predictive accuracy of 86% on the fitted values as computed by the area under the receiver operating characteristic curve relative to that of the individual single nucleotide polymorphism with the highest prognostic performance (area under the receiver operating characteristic curve = 60%). (Stroke. 2009;40[suppl 1]:S67-S70.)

Key Words: Bayesian networks ■ genetics ■ ischemic stroke ■ prediction ■ risk factors

Ischemic strokes account for 87% of all strokes in the United States,1 and in whites, the cardioembolic subtype predominates among all strokes of identifiable origin.2,3 Considerable progress has been made in identifying individual single nucleotide polymorphisms (SNPs) associated with ischemic stroke.4 However, a multivariate approach is required to develop a predictive model of a complex phenotype like ischemic stroke. The ability to predict stroke would allow clinicians to identify and counsel individuals who have a high probability of developing the phenotype. In addition, an accurate predictive model would identify potential novel drug targets. We present the first predictive genomic model of cardioembolic stroke in a white population using data from the Genes Affecting Stroke Risk and Outcome Study (GASROS).

Untangling the origins of a complex disease like cardioembolic stroke requires the use of multivariate models.5 The typical solution is logistic regression, which has been used to examine up to 3 genotypes and 4 clinical risk factors concurrently in relation to ischemic stroke.6 This approach has its limits, however, because when the susceptibility to disease is caused by the interaction among several genes, the number of parameters required to fit a logistic regression model increases exponentially, rapidly demanding unfeasible sample sizes. Overcoming this limitation requires novel statistical modeling that goes beyond traditional methods. The Bayesian network is a proven analytic approach with properties that nominate it as ideally suited to the task of discovering predictive models of complex diseases. Using this approach, a researcher can efficiently search thousands of variables for the most probable model of dependency with a given phenotype in an automated fashion7 and can use this model for prediction.8,9 Importantly, with a given sample size, Bayesian networks are able to concurrently account for a larger number of variables than logistic regression because of their ability to represent models conditional on the phenotype and therefore scale better as the number of simultaneous factors increases.10 The ability of Bayesian networks to identify a predictive model for a complex phenotype has been demonstrated in the context of stroke in individuals with sickle cell anemia, who were each genotyped at 108 SNPs in 80 candidate genes. From these data, a Bayesian network consisting of 25 SNPs and 4 clinical factors was generated; the model was found to have a predictive accuracy of 98.2% in an independent population.11

Identifying a predictive model of ischemic stroke in a general population presents challenges beyond that of stroke in sickle cell anemia. In particular, a general population presents a more heterogeneous portfolio of constituent phenotypes and a more heterogeneous racial spectrum. Because of this, we limited our analysis to the predominant ischemic stroke subtype and the majority race in the GASROS population: cardioembolic stroke (36.3% of cases) and white subjects (94.9% of cases and control subjects).
All GASROS subjects were older than 18 years of age at the time of enrollment, and informed consent from the subject (cases and control subjects) or proxy (cases only) was obtained. Exclusion criteria for both cases and control subjects were active malignancy (other than skin tumors), end-stage renal failure, or end-stage hepatic dysfunction. Potential cases were ascertained from 2002 to 2005 by reviewing emergency department and outpatient neurology clinic logs to identify consecutive adult patients treated at Massachusetts General Hospital, Boston, Mass, with symptoms of suspected stroke. CT or MRI was used to exclude nonischemic etiologies. Unusual causes of stroke (eg, central nervous system vasculitis, migraine, vasospasm, infection, known genetic disorder, or procedural complication) were also excluded. A single stroke specialist (K.L.F.) reviewed the diagnostic studies to determine the diagnosis of stroke and TOAST classification. Control subjects were healthy age- and sex-matched individuals selected from the practices of a large, multiphysician primary care internal medicine practice at Massachusetts General Hospital who were screened for absence of stroke symptoms using the Questionnaire for Verifying Stroke-Free Status. Potential control subjects with a known history of carotid stenosis, transient monocular blindness, transient ischemic attack, ischemic or hemorrhagic stroke, systemic embolization, other thrombotic episodes, or surgery within the last 3 months were excluded.

To develop a predictive model of cardioembolic stroke, we extracted 569 subjects from the GASROS population (146 cases of cardioembolic stroke and 423 control subjects). Blood was drawn from all subjects, DNA was extracted, and genotyping was performed using a custom candidate gene 1536-plex Illumina (Illumina Inc, San Diego, Calif) SNP panel at the Wake Forest University School of Medicine Center for Human Genomics (Winston-Salem, NC). A total of 1313 SNPs were subjected to the Bayesian network analysis using Bayesware Discoverer (www.bayesware.com); the resulting Bayesian network is shown in Figure 1. The network describes the joint association for 37 SNPs (blue) and race (pink) with the risk of cardioembolic stroke (red).
mainder of the SNPs on the panel were excluded due to failure of the Hardy-Weinberg equilibrium test, to low prevalence of the minor allele (less than 1%), or to having more than 10% of missing data.

Figure 1 shows the network generated by the analysis, in which 37 SNPs in 20 genes and intronic regions modulate, together with race, the risk of cardioembolic stroke (case). The network in the figure contains only the SNPs directly modulating the risk of stroke and their mutual dependencies.

To validate the predictive accuracy of the model, we computed for each subject in the sample the probability of stroke given the subject’s genotype and compared the probability with the observed value. The probability of stroke given the genotype of an individual subject was calculated using the clique algorithm. The predictive accuracy of the model was evaluated by calculating the receiver operator characteristic (ROC) curves. Convex hulls were estimated for each ROC curve using the Qhull algorithm implemented in MATLAB (MathWorks, Natick, Mass). Figure 2 displays the ROC curve for such a classifier, showing a predictive accuracy, computed as area under the ROC curve, of 86%. The robustness of the model to sampling variability was evaluated using a procedure known as crossvalidation, which consists of resampling the cases and control subjects used to construct the model in turn, re-estimating the model parameters and predicting the occurrence of stroke in each subject. The predictive accuracy estimated by this method was 80%, quite close to the 86% of the predictions on the fitted values. By contrast, the Table shows that the predictive accuracy on the fitted values achieved by using one SNP at a time is close to random, represented the 50% area under the ROC, and is never greater than 60%, a performance largely inferior to the 86% accuracy achieved using the entire network.

These results suggest that accurate SNP-based predictive models of cardioembolic stroke are within reach, but their development will require the deployment of methods able to capture the interactions of the genetic variations underpinning the disease mechanisms. Assuming that these results will hold when validated on independent populations, they represent a lower bound of the predictive performance that will be achieved through the analysis of genome-wide association studies, because we will be able to include in the models genetic risk factors that are today unsuspected. From a clinical point of view, these predictive models will prove especially useful for cardioembolic stroke, because it...
has been found to be the least heritable of all types of stroke.\textsuperscript{14,15}

**Sources of Funding**
This research was supported, in part, by the NHGRI/National Institutes of Health (R01 HG003354), National Institute of Dental and Craniofacial Research (K08 DE016956), National Institute of Neurological Diseases and Stroke (K23 NS042720), the Mallinckrodt GCRC at Massachusetts General Hospital, grant number M01 RR-01066, National Center for Research Resources, National Institutes of Health, and the Deane Institute of Integrative Research in Stroke and Atrial Fibrillation. DNA panels from the National Institute of Neurological Diseases and Stroke Human Genetics Resource Center DNA and Cell Line Repository (http://ccr.coriell.org/ninds) were used in this study.

**Disclosures**
None.

**References**
Predictive Genomics of Cardioembolic Stroke
Rachel Badovinac Ramoni, Blanca E. Himes, Michele M. Sale, Karen L. Furie and Marco F. Ramoni

Stroke. 2009;40:S67-S70; originally published online December 8, 2008;
doi: 10.1161/STROKEAHA.108.533273

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/40/3_suppl_1/S67

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/