American Journal of Epidemiology
Vol. 165, No. 3
DOI: 10.1093/aje/kwk006
Advance Access publication November 2, 2006

Practice of Epidemiology

Logic Regression for Analysis of the Association between Genetic Variation in the Renin-Angiotensin System and Myocardial Infarction or Stroke

Charles Kooperberg[1,2], Joshua C. Bis[3,4], Kristin D. Marciante[3,4], Susan R. Heckbert[3,4], Thomas Lumley[2,3], and Bruce M. Psaty[3,4,5]

[1] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA.
[2] Department of Biostatistics, School of Public Health and Community Medicine, University of Washington, Seattle, WA.
[3] Cardiovascular Health Research Unit, University of Washington, Seattle, WA.
[4] Department of Epidemiology, School of Public Health and Community Medicine, University of Washington, Seattle, WA.
[5] Department of Health Services, School of Public Health and Community Medicine, University of Washington, Seattle, WA.

Received for publication October 17, 2005; accepted for publication June 13, 2006.

Recent developments in genetic sequencing technology now make it possible to genotype large numbers of single nucleotide polymorphisms (SNPs) in large samples. Many association studies using SNP data are now being carried out. Typically, these observational studies establish whether certain haplotypes or individual SNPs are associated with a health outcome. Few methods exist for finding interaction effects among multiple SNPs or between SNPs and environmental factors. In this paper, the authors describe logic regression, an exploratory method with which to identify interactions for further research. They illustrate this method using data from a US case-control study of myocardial infarction and stroke (1995–1999) carried out among 1,614 persons in Washington State who were genotyped for 32 SNPs on five genes in the renin-angiotensin system.

epidemiologic methods; epistasis, genetic; models, statistical; polymorphism, single nucleotide; regression analysis

Abbreviations: ACE, angiotensin-converting enzyme; AGT, angiotensinogen; AGTR, angiotensin II receptor; RAS, renin-angiotensin system; REN, renin; SNP, single nucleotide polymorphism.

Over the last few years, the number of studies characterizing associations between single nucleotide polymorphisms (SNPs) and disease outcomes has increased dramatically. With the recent developments in genetic sequencing technology, the number and size of these studies will increase further. While single SNP association analyses are straightforward to carry out, they do not make efficient use of the genomic structure.

For many SNP association studies, interest will not be limited to identifying individual SNPs or haplotypes associated with a disease outcome but, equally important, will also involve the identification of interactions between SNPs within a gene (as in a haplotype effect), between genes, or between genes and environmental factors such as drugs, smoking, and alcohol consumption. Few methods exist for finding interaction effects of multiple SNPs or between SNPs and environmental factors (1, 2). Reasons for this shortage of methods may be the large number of potential interactions, which makes it practically impossible to examine all interaction models, and the requirement of multiple-comparisons correction for all models that are examined.

In judging methods for identifying interactions, it is useful to differentiate between three types of interactions that can be of interest.

Correspondence to Charles Kooperberg, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M3-A410, Seattle, WA 98109-1024 (e-mail: clk@fhcrc.org).

Within-gene interactions. Identification of interactions between SNPs within the same gene, or within the same region of a chromosome, is an alternative to using reconstructed haplotypes for association studies. An advantage of using haplotypes is that there are relatively few common ones within a block, and it is thus an easy way to model interactions. However, haplotypes depend on which tagSNPs, which "tag" the desired haplotype block, are genotyped, and there is no guarantee that another study will identify the same haplotypes. If several haplotypes all have increased or decreased risk, it may not be straightforward to recognize which SNPs are associated with the disease. Moreover, it is conceivable that no individual haplotype is significantly associated with the outcome while a simple combination of fewer SNPs is significant, since haplotypes may include SNPs that are not relevant to the disease risk. A within-gene interaction of the type "a variant at SNP 1 and a variant at SNP 3" could potentially identify groups of haplotypes that together produce increased risk.

Between-gene interactions. The term "epistasis" was coined to describe the effect of masking of the phenotypic effects of one gene by a second gene. More generally, epistasis is considered an interaction between different genes. Epistasis is thought to play a significant role in complex diseases (3, 4). The identification of epistasis can be problematic, since if the effect of one locus is masked by the effect of another locus, the power to detect the first locus is probably reduced.

Gene-environment interactions. The computational complexity of gene-environment interactions, with respect to the difficulty of identifying such interactions, falls somewhere between the complexity of "within-gene" interactions and "between-gene" interactions. For a study with 25,000 SNPs, there may be approximately $25{,}000^2/2$ possible two-SNP interactions, but there are only $25{,}000q$ SNP-environment interactions, where q is the (usually small) number of environmental factors under consideration.

Logic regression (5) is an exploratory regression method that is designed for analysis with binary predictors when interest is in possible interactions between predictors. As such, it is well suited for SNP association studies. In this paper, we will briefly describe the logic regression method and apply it to data on 349 myocardial infarction patients, 202 stroke patients, and 1,063 controls for whom 32 SNPs on five genes in the renin-angiotensin system (RAS) were genotyped.

MATERIALS AND METHODS

RAS data

The RAS plays a central role in maintenance of vascular tone and in salt and water homeostasis. Renin (REN) cleaves angiotensinogen (AGT) to produce angiotensin I, which is converted by angiotensin-converting enzyme (ACE) to angiotensin II, a potent constrictor. Evidence suggests that the *235T* allele of the *AGT* gene is associated with increased risk of hypertension and elevated AGT levels; other related genotypes, such as angiotensin II receptor types 1 and 2 (*AGTR1* and *AGTR2*), have been associated with hypertension and cardiovascular complications.

We carried out this study to investigate whether the effect of ACE inhibitors on risk of incident nonfatal myocardial infarction or stroke differs by *ACE* or *AGT* genotype. Our analysis was conducted within a population-based case-control study of members of the Group Health Cooperative of Puget Sound (western Washington State) aged 30–79 years with pharmacologically treated hypertension. The study design has been described elsewhere (6). Cases were persons who survived an incident myocardial infarction or stroke during the period 1995–1999, and controls were eligible if they did not have a history of myocardial infarction or stroke. Controls were randomly sampled from the Group Health Cooperative enrollment files and were frequency-matched to the myocardial infarction cases on age decade, sex, and calendar year of identification.

Current use of antihypertensive medications was determined using computerized pharmacy records. Individual drugs were grouped into major classes: diuretics, β-blockers, ACE inhibitors, calcium-channel blockers, and vasodilators. Diuretics included both loop and thiazide diuretics.

The SNPs for the RAS data were identified in the Seattle-SNPs Variation Discovery Resource (http://pga.gs.washington.edu) by genomic resequencing. Patterns of linkage disequilibrium were used to select a subset of SNPs that tagged major common patterns of variation. Using this method, three tagSNPs for *ACE*, eight tagSNPs for *AGT*, 12 tagSNPs for *AGTR1*, three tagSNPs for *AGTR2*, and six tagSNPs for *REN* were genotyped. *AGTR2* is located on the X chromosome; thus, analyses were conducted separately for men and women. More details can be found in the paper by Marciante et al. (7).

The current analysis using logic regression tries to identify combinations of SNPs and drug classes that are associated with increased or decreased disease risk. These are slightly different interactions than the ones for which the study was designed.

Logic regression

Assuming a link function h of interest, a traditional interaction model is

$$h\{E(Y|X)\} = \beta_0 + \beta_1 X_1 + \beta X_2 + \beta_3 X_1 X_2,$$

where Y is the disease phenotype and X_1 and X_2 are indicators of genotypes at two different loci. The parameter β_3 models the interaction. Locus X_i could be a categorical variable with three levels, so the global test for interaction has 4 degrees of freedom. More restrictive tests based on subsets of the predictors can be used to improve power. In particular, interaction models that are interpretable without main effects are potentially more powerful than the traditional interaction model, since they use fewer degrees of freedom.

The logic regression model is

$$h\{E(Y|X)\} = \beta_0 + \sum_{i=1}^{m} \beta_i L_i + \sum_{i=1}^{p} \beta_{i+p} Z_{i+p}.$$

Here, $h(\cdot)$ is a link function relating the response and the covariates, such as the identity function for continuous outcomes or the logit function for binary outcomes. Each of the logic trees L_i is a Boolean combination of binary predictors $X_j, j = 1, \ldots, J$, such as ((X_7 and X_{13}^c) or X_5), where 1 equals "true," 0 equals "false," and the superscript c refers to the complement. The Z_i are additional confounders.

Logic regression is an adaptive algorithm which for a given model selects those L_i that minimize the residual sum of squares or the deviance. Typically the number of logic trees m selected is small (between 1 and 3), and the L_i can be interpreted as "risk factors." We say that the logic tree in the equation above has three leaves. The method is described in detail by Ruczinski et al. (5), and software is available from the R Foundation for Statistical Computing (http://cran.r-project.org) as an R/CRAN package (8).

Optimization of the logic regression model is carried out using a (stochastic) simulated annealing algorithm employing an irreducible Markov chain. At any stage of this algorithm, a logic tree gets modified by replacing predictors (like "X_7") or operators (like "and") or by changing the form of the tree (such as adding or deleting another "or X_i"). Modifications are proposed at random; if the proposed model is an improvement over the current model, then it is accepted (it replaces the current model), while if the proposed model is worse than the current model, it is accepted with a probability that depends on the stage of the algorithm and how much worse the proposed model is.

Logic regression for SNP data

Each SNP is coded into two binary covariates: $X_i^d = 1$ if a person has at least one variant allele and $X_i^r = 1$ if the person has two variant alleles; both are 0 otherwise. One can see that X_i^d and X_i^r code the dominant and recessive effects of SNP i, respectively. Logic regression has been applied successfully to the simulated SNP data of the 12th Genetic Analysis Workshop and to a study of heart disease (9, 10).

Model selection

For adaptive regression methods like logic regression, model selection is needed, since more complicated models typically fit data better, even if there is no signal. Model selection can make use of a simple penalty on the model complexity (size), such as Akaike's Information Criterion (11), or it can involve the use of a separate test set, cross-validation, or permutation tests. Within the logic regression software, three model selection tools are available.

1. Permutation tests can be used to globally assess whether any combination of SNPs is associated with the response: The outcome (e.g., case-control status) is permuted at random, and the quality of the fit (e.g., deviance) on the real data is compared with the quality of the fit on the permuted data.
2. Conditional permutation tests can be used to assess whether combinations of SNPs that are more complicated (because they involve either more complicated expressions or more expressions) than a particular model are more strongly associated with response than the current model. For these tests, the permutations are carried out such that we are guaranteed that even on the permuted data, a particular model can be fitted. Any improvement of the quality of the fit beyond that model is noise and can be compared with the quality of the fit on the real data.
3. Cross-validation can be used to assess which complexity of the model has the best predictive performance. For cross-validation, the data are divided into 10 equal parts. Then, 10 times, one partition is left as a test set, and for each possible model complexity the best model is selected using nine out of 10 parts, after which the predicted deviance on the remaining test set part is computed. For each level of complexity, the 10 predicted deviances are added, and the complexity with the smallest overall predicted deviance is selected.

Typically, the conditional permutation approach will suggest larger models than cross-validation, since the conditional permutation approach assesses an association between the predictors and the response, while cross-validation requires that such an association be strong enough to reduce the prediction error.

Missing data

In our RAS data, only 3 percent of the genotypes were missing. Forty-four percent of the cases and controls had at least one missing genotype, and each SNP had at least one missing genotype. Because any combination of SNPs could be involved in an interaction, a complete case analysis would require elimination of 44 percent of the cases and controls. Since we had several SNPs in each of the genes, a haplotype reconstruction was possible. Haplotypes were inferred using a Bayesian population genetic model that uses coalescent theory (PHASE (12)). A reconstructed haplotype implied a value for a missing genotype. When the imputed genotypes were ambiguous, we used the estimated haplotype probabilities as case weights. In some situations, the haplotype reconstruction was ambiguous, but each of these haplotypes implied the same genotype, so the missing genotypes could be imputed unambiguously.

Logic regression for haplotype data

Logic regression selects Boolean combinations of SNPs, thereby implicitly grouping haplotypes. In some situations, the expressions generated by logic regression may be hard to interpret, since even some of the most elementary expressions may involve unrelated SNPs. An alternative is then to require the most elementary expressions, such as "X_1 or X_2" or "X_3 and X_4," to involve SNPs within the same gene. Here we ran the publicly available version of logic regression software on haplotypes. For each haplotype, we created two binary predictors based on whether the subject had at least one copy of the haplotype or at least two copies of the haplotype, as for the SNP data.

TABLE 1. Results of a logic regression null-model permutation test for the association between single nucleotide polymorphisms and myocardial infarction or stroke*

	Null model†	Permutation test			
		SNPs‡		SNPs and drugs	
		Best score§	% smaller¶	Best score	% smaller
SNPs within gene					
Myocardial infarction					
ACE‡	1,576.295	1,562.428	82.8	1,564.504	56.1
AGT‡	1,575.217	1,557.370	3.9	1,554.695	8.0
AGTR1‡	1,569.675	1,552.151	23.3	1,550.955	29.3
AGTR2—women	666.411	658.503	20.2	648.606	6.2
AGTR2—men	908.060	906.670	68.0	896.425	40.8
REN‡	1,575.359	1,561.800	12.1	1,559.775	65.7
Across all genes except *AGTR2*	1,579.185	1,533.631	35.1	1,533.655	45.7
Stroke					
ACE	1,109.807	1,100.121	11.7	1,097.308	44.7
AGT	1,108.108	1,100.248	83.1	1,095.293	75.9
AGTR1	1,106.974	1,088.881	16.5	1,078.989	1.3
AGTR2—women	596.789	591.272	47.3	588.157	88.0
AGTR2—men	505.318	503.077	50.9	491.993	20.1
REN	1,109.358	1,095.940	13.4	1,091.296	15.4
Across all genes except *AGTR2*	1,111.037	1,053.773	1.9	1,057.859	6.9
		Haplotypes		Haplotypes and drugs	
		Best score	% smaller	Best score	% smaller
Haplotypes within gene					
Myocardial infarction					
ACE	1,567.419	1,564.271	75.6	1,555.808	55.8
AGT	1,572.797	1,556.432	4.3	1,553.558	8.7
AGTR1	1,560.084	1,552.967	70.4	1,544.052	7.9
AGTR2—women	665.999	656.600	15.9	648.520	10.6
AGTR2—men	908.060	906.670	68.0	896.425	46.7
REN	1,572.493	1,562.664	35.6	1,559.775	65.7
Across all genes except *AGTR2*	1,579.185	1,531.476	4.3	1,531.476	7.6
Stroke					
ACE	1,102.619	1,093.208	6.7	1,089.554	37.5
AGT	1,106.720	1,099.525	79.8	1,091.894	37.1
AGTR1	1,098.847	1,083.309	1.7	1,079.989	1.3
AGTR2—women	596.460	588.570	28.6	586.280	80.6
AGTR2—men	505.318	503.077	45.7	492.437	29.0
REN	1,107.076	1,092.825	7.4	1,085.934	2.4
Across all genes except *AGTR2*	1,111.037	1,065.171	5.0	1,062.886	7.8

* For both myocardial infarction and stroke, the table shows results from analyses of all SNPs on a particular gene, all haplotypes for a particular gene, and either of those in combination with the various drug classes.

† Deviance for the model with no logic trees (slight differences in the scores between different genes are due to small probabilities of particular imputed SNPs that were dropped).

‡ SNP, single nucleotide polymorphism; ACE, angiotensin-converting enzyme; AGT, angiotensinogen; AGTR, angiotensin II receptor; REN, renin.

§ Deviance of the fitted model with, at most, one logic tree with four leaves on the actual data.

¶ Percentage of the permutations that had a better test score: small percentages are suggestive of an association.

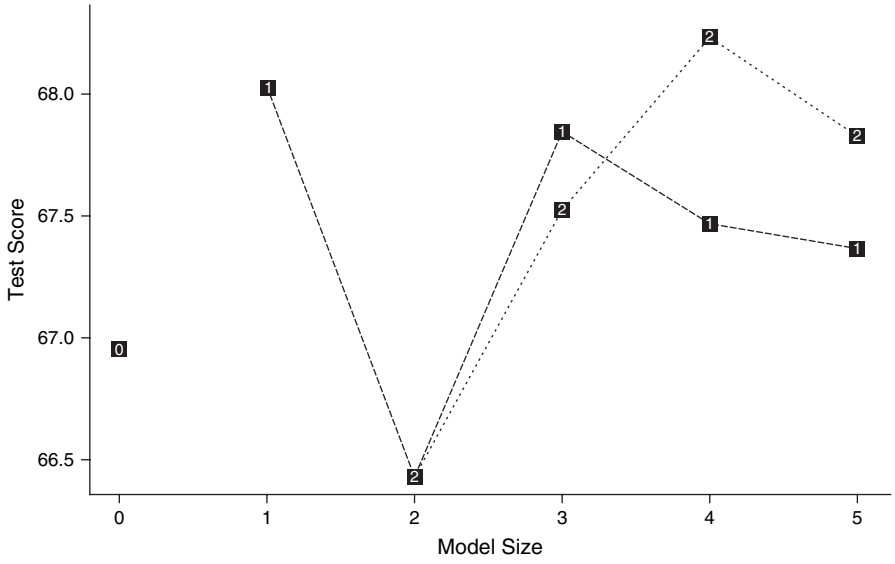

FIGURE 1. Cross-validation results for associations of antihypertensive drugs and angiotensin II receptor type 2 (*AGTR2*) single nucleotide polymorphisms with myocardial infarction in women. The plot shows the cross-validation test set deviance ("Test Score") for models with a specific number of logic trees (numbers in squares) and total number of leaves ("Model Size"). Models with the smallest test set deviance have the best predictive performance.

Processing of the data

Each of the 32 SNPs was replaced by two binary covariates, using the coding described above in the section "Logic regression for SNP data." Missing genotypes were inferred using the PHASE haplotype reconstruction. For analyses of single genes, we used case weights when genotypes were ambiguous, ignoring probabilities that were smaller than 0.05; the total probability covered by these cases was smaller than 0.5 percent of the data for every analysis. For the analysis using all genes, we imputed the most likely genotype, since using case weights for each SNP in each gene would lead to a much larger data set when there were multiple genes. An alternative is to use multiple imputations for missing genotypes. Because, in the current data, the percentage of missing genotypes was small, there was no practical difference in performance between these approaches, as we confirmed for selected analyses; when the percentage of missing genotypes is larger, a multiple-imputations approach leads to the least biased results (13). For the analysis using haplotypes, we used those reconstructed by PHASE and otherwise coded them in the same way as we code SNPs. We used 29 haplotypes that were estimated to occur at a frequency of more than 5 percent; thus, the number of haplotypes was very similar to the number of SNPs in this study. For the drug-genotype interactions, each of the drug classes was coded as a binary variable. In our current analysis, we did not control for other covariates, although they could be added in the logic regression model in a straightforward fashion.

The simulated annealing chains had 1,000,000 model evaluations for the actual model fitting runs and 500,000 model evaluations for the permutation tests and cross-validation runs. Other options used the defaults, unless indicated otherwise.

RESULTS

Below we describe our analysis of the RAS data using logic regression. In a haplotype analysis, Marciante et al. (7) did find an association between *AGT* haplotypes and myocardial infarction in these data.

Global permutation tests

Table 1 shows the results of global permutation tests. The tests were carried out with a maximum model size of one logic tree with four leaves. A small model size allows the algorithm to do an almost exhaustive search of all models, yielding permutation p values with little noise.

Most genes and gene-drug interactions appeared not to be associated with the outcomes. For most permutation tests, well over 5 percent of the permutations yielded scores (deviances) that were smaller than the score of the best model. The *AGT* SNPs, and to a lesser extent the *AGT* haplotypes, showed some association with myocardial infarction. The *AGTR2* SNPs in combination with drugs suggested a possible association with myocardial infarction in women. The *REN* haplotypes, and to a lesser extent the *REN* SNPs, showed some association with stroke. The *AGTR1* SNPs and haplotypes together with the drugs showed a possible association with stroke. Interestingly, all SNPs or haplotypes combined showed a stronger association with stroke than any of the associations with individual genes.

In the remaining analysis, we concentrated on the possible associations of drugs and *AGTR2* SNPs with myocardial

TABLE 2. Results from a conditional permutation test for the interaction of angiotensin II receptor type 2 (*AGTR2*) single nucleotide polymorphisms with drugs for myocardial infarction in women

No. of logic trees	Total no. of leaves in tree(s)	Initial score*	Best score†	% smaller‡
0	0	666.411	612.119	2.0
1	1	661.935	612.119	1.6
1	2	652.551	612.119	15.6
1	3	650.369	612.119	20.8
1	4	645.897	612.119	39.6
1	5	642.280	612.119	58.0
2	2	652.551	612.119	13.6
2	3	649.921	612.119	15.6
2	4	647.185	612.119	27.6
2	5	642.152	612.119	48.0

* Deviance of the model conditional on which the permutation was carried out (and beyond which only noise was fitted).

† Deviance of the fitted model with, at most, two logic trees with eight leaves on the actual data.

‡ Percentage of the permutations that had a better score than the best score. A small percentage in this column suggests that a model *larger than* the one conditional on which the permutation was carried out may fit the data better than that one.

TABLE 3. Results from a conditional permutation test for the association of angiotensinogen (*AGT*) single nucleotide polymorphisms with myocardial infarction

No. of logic trees	Total no. of leaves in tree(s)	Initial score*	Best score†	% smaller‡
0	0	1,575.217	1,537.941	11.6
1	1	1,570.368	1,537.941	22.8
1	2	1,565.938	1,537.941	43.6
1	3	1,558.675	1,537.941	78.8
1	4	1,557.273	1,537.941	86.0
1	5	1,555.848	1,537.941	90.8
2	2	1,565.938	1,537.941	41.6
2	3	1,558.675	1,537.941	78.4
2	4	1,556.529	1,537.941	86.0
2	5	1,553.279	1,537.941	90.0

* Deviance of the model conditional on which the permutation was carried out (and beyond which only noise was fitted).

† Deviance of the fitted model with, at most, two logic trees with eight leaves on the actual data.

‡ Percentage of the permutations that had a better score than the best score. A small percentage in this column suggests that a model *larger than* the one conditional on which the permutation was carried out may fit the data better than that one.

infarction in women, of *AGT* SNPs with myocardial infarction, of drugs and *AGTR1* haplotypes with stroke, and of a combination of all SNPs with stroke.

Associations of hypertensive drugs and *AGTR2* SNPs with myocardial infarction in women

In figure 1, the cross-validation results of the analysis of the *AGTR2* SNPs and drugs for myocardial infarction in women suggest that a model with one or two logic trees with two leaves in total has the best predictive performance among the models examined. This agrees with the results of the conditional permutation tests shown in table 2, since that table indicates that the best model would be larger than a model with one logic tree with one leaf but would not be larger than a model with one logic tree with two leaves or a model with two logic trees with two leaves. For table 2 and all further conditional permutation tests, the largest model size used was two logic trees with eight leaves each. If the maximum model size was the same as for table 1, the entry with zero trees and zero leaves in table 2 would be the same as the entry for drugs and *AGTR2* SNPs with myocardial infarction in women in table 1. For conditional permutation tests, where we try to determine the size of a model, we typically use a larger model than that used for the unconditional permutation tests. The best model with one logic tree with two leaves is

$$\begin{aligned}&\text{logit}(P(\text{myocardial infarction}|\textit{AGTR2}\ \text{SNPs, drug classes}))\\&\quad=-0.900-0.720\times((\geq 1\ \text{A allele at SNP rs17231429})\\&\qquad\text{and}\ (\text{no calcium channel blockers})).\end{aligned}$$

When analyzed as an explicit model, the t value for the logic tree is -3.70 and the odds ratio is 0.487. The permutation and cross-validation approaches implicitly correct for the number of models examined for this gene, but a p value for a selected logic regression model and a confidence interval for an odds ratio are not corrected and would thus be of limited value. Because we examined several genes and outcomes, the differences in figure 1 are modest, and since the percentages in table 2 are not much under 5 percent, we consider the results only suggestive of an association.

AGT SNP associations with myocardial infarction

Table 3 shows the results of the permutation tests for associations of *AGT* SNPs with myocardial infarction. The effect of the increased maximum model size compared with table 1 is that there does not seem to be much association between *AGT* SNPs and myocardial infarction. Typically, the magnitude of the association that is identified is smaller when the maximum model size is larger, since even in data where there is no signal, models that are large enough will show some association. The cross-validation analysis (figure 2) suggests that a model with three SNPs has some predictive power. This model, a model with one tree and three leaves, is

$$\begin{aligned}&\text{logit}(P(\text{myocardial infarction}|\textit{AGT}\ \text{SNPs}))\\&\quad=-0.953-0.564\times[(\geq 1\ \text{T allele for SNP rs2478523})\\&\qquad\text{and}\ ((2\ \text{G alleles for SNP rs2493132})\ \text{or}\\&\qquad(\geq 1\ \text{T allele for SNP rs7079}))].\end{aligned}$$

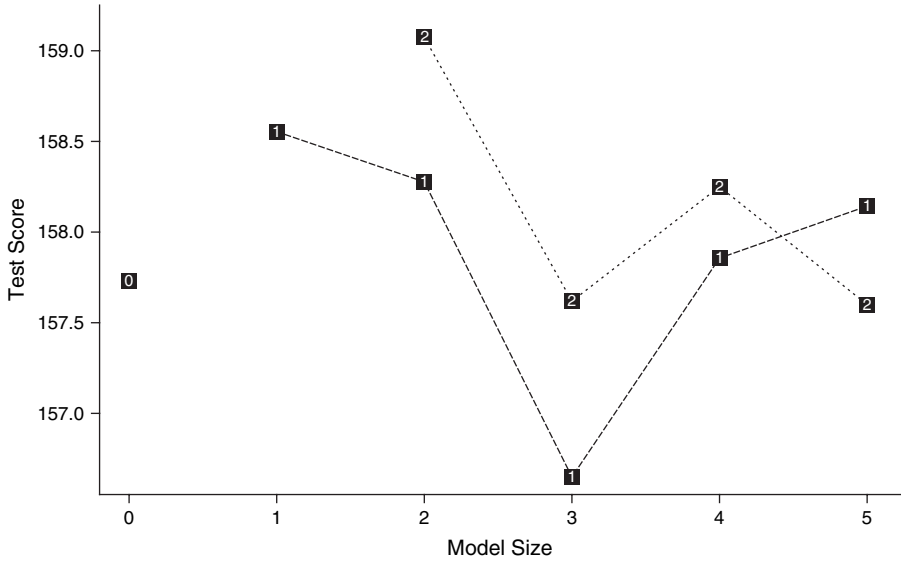

FIGURE 2. Cross-validation results for associations of angiotensinogen (*AGT*) single nucleotide polymorphisms with myocardial infarction. The plot shows the cross-validation test set deviance ("Test Score") for models with a specific number of logic trees (numbers in squares) and total number of leaves ("Model Size"). Models with the smallest test set deviance have the best predictive performance.

The three SNPs in this model are consecutive along the genome in the RAS data. This model has a t value of -3.96 and an odds ratio of 0.568. There is a limited correspondence between this model and the haplotype analysis of Marciante et al. (7). In that analysis, compared with the most common haplotype, two haplotypes are associated with increased risk (one statistically significant). For the analysis presented here, any participant with at least one copy of the most frequently occurring of these two haplotypes who was at increased risk would be in the high-risk group defined by logic regression, but this group also included some of the participants who were not at increased risk in the haplotype analysis.

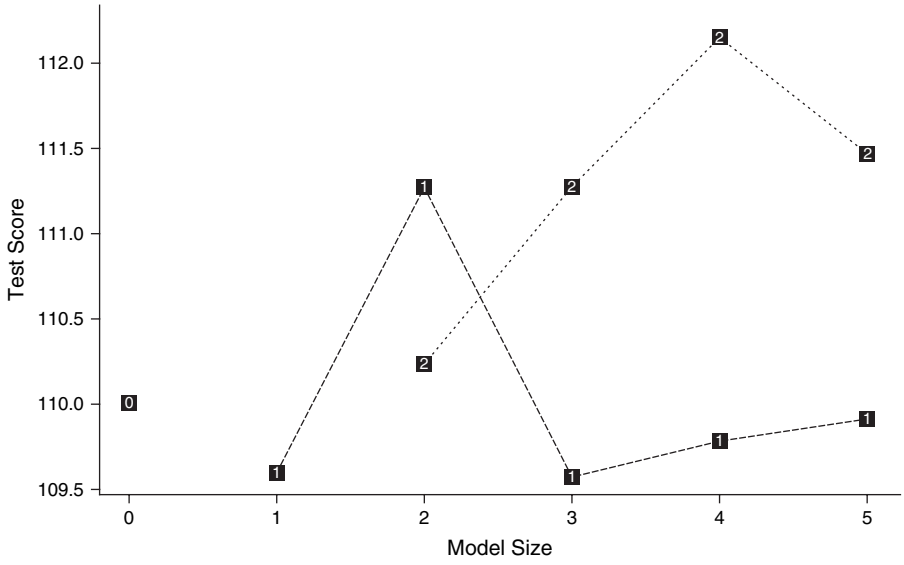

FIGURE 3. Cross-validation results for the interaction of angiotensin II receptor type 1 (*AGTR1*) haplotypes with drugs for stroke. The plot shows the cross-validation test set deviance ("Test Score") for models with a specific number of logic trees (numbers in squares) and total number of leaves ("Model Size"). Models with the smallest test set deviance have the best predictive performance.

TABLE 4. Results from a conditional permutation test for the interaction of angiotensin II receptor type 1 (*AGTR1*) haplotypes with drugs for stroke

No. of logic trees	Total no. of leaves in tree(s)	Initial score*	Best score†	% smaller‡
0	0	1,098.847	1,047.702	1.2
1	1	1,090.288	1,047.702	4.4
1	2	1,089.364	1,047.702	9.2
1	3	1,081.336	1,047.702	32.8
1	4	1,078.988	1,047.702	37.6
1	5	1,075.133	1,047.702	53.6
2	2	1,087.104	1,047.702	6.8
2	3	1,081.090	1,047.702	30.8
2	4	1,077.815	1,047.702	27.6
2	5	1,073.955	1,047.702	38.4

* Deviance of the model conditional on which the permutation was carried out (and beyond which only noise was fitted).

† Deviance of the fitted model with, at most, two logic trees with eight leaves on the actual data.

‡ Percentage of the permutations that had a better score than the best score. A small percentage in this column suggests that a model *larger than* the one conditional on which the permutation was carried out may fit the data better than that one.

TABLE 5. Results from a conditional permutation test for the association of all single nucleotide polymorphisms with stroke

No. of logic trees	Total no. of leaves in tree(s)	Initial score*	Best score†	% smaller‡
0	0	1,111.037	1,024.191	24.8
1	1	1,101.706	1,024.191	49.6
1	2	1,089.323	1,024.191	75.6
1	3	1,081.203	1,024.191	94.0
1	4	1,076.327	1,024.191	94.4
1	5	1,070.502	1,024.191	98.8
2	2	1,089.323	1,024.191	75.6
2	3	1,081.204	1,024.191	89.2
2	4	1,073.049	1,024.191	95.6
2	5	1,066.852	1,024.191	98.8

* Deviance of the model conditional on which the permutation was carried out (and beyond which only noise was fitted).

† Deviance of the fitted model with, at most, two logic trees with eight leaves on the actual data.

‡ Percentage of the permutations that had a better score than the best score. A small percentage in this column suggests that a model *larger than* the one conditional on which the permutation was carried out may fit the data better than that one.

Drug-haplotype interactions with stroke for the *AGTR1* gene

The cross-validation results of the analysis of the *AGTR1* haplotypes and drugs for stroke (figure 3) suggested that a model with a single logic tree with one or three leaves had the best predictive performance among the models examined. This agrees with the results of the conditional permutation tests shown in table 4, since this table indicates that the best model would be larger than one logic tree with one leaf and maybe even larger than a model with two leaves, but would not be larger than a model with three leaves.

The model with three leaves for these data is

$$\begin{aligned}&\text{logit}(P(\text{stroke}|\textit{AGTR1}\text{ haplotypes, drug classes}))\\&\quad=-1.734+1.360\times[(2\text{ copies of }\textit{AGTR1}\text{ haplotype G})\\&\qquad\text{or }((2\text{ copies of }\textit{AGTR1}\text{ haplotype D})\text{ and}\\&\qquad(\text{no }\beta\text{-blockers})]).\end{aligned}$$

(The haplotypes are labeled as in the paper by Marciante et al. (7).) Haplotypes D and G were two of the three high-risk haplotypes in the haplotype analysis (7). The *t* statistic for this model was 4.45, and the corresponding odds ratio was 3.90. The logic tree for this model identified a group of only 48 people.

Effect of all SNPs on stroke

Table 5 summarizes the results of the conditional permutation tests for the relation between all SNPs and stroke. These results suggest that the best model may be of size zero. This again seems to contradict the global permutation tests somewhat. Figure 4, showing the cross-validation results of the analysis of the relation between all SNPs and stroke, suggests that perhaps models with two or three SNPs in two logic trees are slightly better than other model sizes; the figure suggests that these models may have some predictive power. The best model with three SNPs is of some interest, as it combines effects of SNPs in three different genes:

$$\begin{aligned}&\text{logit}(P(\text{stroke}|\text{all RAS SNPs}))=-1.537-2.032\\&\quad\times[((2\text{ T alleles for }\textit{ACE}\text{ SNP rs17230372})\text{ or}\\&\quad(\geq 1\text{ T allele for }\textit{AGTR1}\text{ SNP rs17237596})\text{ and}\\&\quad(\geq 1\text{ T allele for }\textit{REN}\text{ SNP rs11571078})].\end{aligned}$$

The marginal *t* statistic for this model was 3.96, which, given the large number of models that were being examined, is not convincing of an association. The corresponding odd ratio was 0.131. Of the 146 participants for which the logic tree was true, only four had a stroke, while of the other 1,119, 198 had such an event. *ACE* SNP rs17230372 is also known as the ACE insertion-deletion variant. This polymorphism, the most widely studied polymorphism of the RAS variants, has been linked with cardiovascular disease in several studies (e.g., see the paper by Agerholm-Larsen et al. (14)).

DISCUSSION

It is widely acknowledged that the analysis of epistasis is challenging because of the large number of statistical models to be evaluated and the limited sample sizes if all

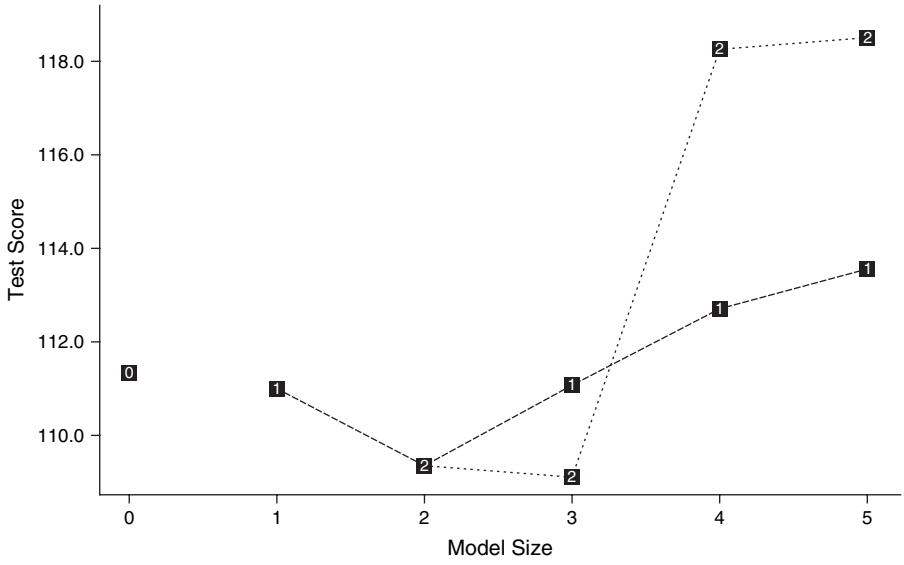

FIGURE 4. Cross-validation results for the association of all single nucleotide polymorphisms with risk of stroke. The plot shows the cross-validation test set deviance ("Test Score") for models with a specific number of logic trees (numbers in squares) and total number of leaves ("Model Size"). Models with the smallest test set deviance have the best predictive performance.

combinations of (tag)SNPs are considered. Although there are situations where there may be sufficient power to identify large gene-gene interactions (15), generally speaking a search for all combinations in large association studies will have limited power because of the number of tests being carried out.

In the next few years, the size of genome association studies is going to increase dramatically, both in the number of samples and (especially) in the number of SNPs genotyped. Researchers will want to find potential gene-gene or gene-environment interactions, and there is a pressing need for methods that can potentially identify such interactions. Besides logic regression, very few such methods currently exist (e.g., see the papers by Ritchie et al. (1) and Foulkes et al. (2)). Methods that employ adaptive selection of models and use a limited number of degrees of freedom (parameters) to model interactions are more likely to successfully identify interactions than methods that are less prudent with study resources.

In this paper, we have illustrated the use of logic regression, an adaptive regression method that uses a Boolean model structure well suited for SNP data and adaptive model selection, on data from a cardiovascular disease case-control study with 32 SNPs. Since logic regression is a well-defined procedure, model selection and multiple-comparisons corrections for the significance level are implicit and do not require further resampling or bootstrapping. Unless a test data set is available to verify selected logic regression models immediately, the adaptive model selection makes logic regression particularly appropriate as an exploratory method for identifying interactions for further research. Logic regression is intended for binary predictors, such as SNPs. If one wished to enter continuous predictors in the interactions, they would have to be dichotomized.

The types of interactions identified by logic regression are not "traditional" interactions, where one predictor modifies the effect of another predictor, but rather combinations of predictors that are associated with increased or decreased disease risk. With a single logic tree, such logic regression identifies a single group of persons at increased (decreased) risk; when the underlying risk profile is more complicated, additional logic trees may be needed.

As was the case for many of the SNP association studies that have been carried out, the association between the SNPs and the RAS data was not very strong. Marciante et al. (7) reached the same conclusion using a haplotype analysis. Nevertheless, we feel that these data illustrate how logic regression can be used to identify local interactions, drug-gene interactions, or gene-gene interactions in an automated fashion. In addition to case-control studies, logic regression can be used for cohort studies, survival analysis, and any other study design for which the model can be formulated as a generalized linear model.

Recently, we developed Monte Carlo logic regression (16). In this variation of logic regression, rather than identify interactions that have a significant association with a clinical outcome, the investigators identify larger numbers of potential interactions using a Markov chain Monte Carlo algorithm. An illustration of this approach on a cardiovascular disease data set with 779 participants and 89 SNPs can be found in the paper by Kooperberg and Ruczinski (16).

ACKNOWLEDGMENTS

This research was supported in part by grants CA53996 and CA74841 from the National Cancer Institute; by grants

HL43201, HL60739, HL68639, HL68986, and HL74745 from the National Heart, Lung, and Blood Institute; and by grants 9970178N and 0270054N from the American Heart Association.

Conflict of interest: none declared.

REFERENCES

1. Ritchie MD, Hahn LW, Roodi N, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 2001;69:138–47.
2. Foulkes AS, DeGrutola V, Hertogs K. Combining genotype groups and recursive partitioning: an application to HIV-1 genetics data. Appl Stat 2004;53:311–23.
3. Fijneman RJ, de Vries SS, Jansen RC, et al. Complex interactions of new quantitative trait loci, Sluc1, Sluc2, Sluc3, and Sluc4, that influence the susceptibility to lung cancer in the mouse. Nat Genet 1996;14:465–7.
4. Frankel WM, Schork NJ. Who's afraid of epistasis? Nat Genet 1996;16:371–3.
5. Ruczinski I, Kooperberg C, LeBlanc M. Logic regression. J Comput Graph Stat 2003;12:475–511.
6. Psaty BM, Smith NL, Heckbert SR, et al. Diuretic therapy, the α-adducin gene variant, and the risk of myocardial infarction or stroke in persons with treated hypertension. JAMA 2002; 287:1680–9.
7. Marciante KD, Bis JC, Rieder M, et al. Renin-angiotensin system gene haplotypes and the risk of myocardial infarction and stroke in pharmacologically treated hypertensive patients. (Abstract). Circulation 2005;111(suppl):e323.
8. R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2005.
9. Kooperberg C, Ruczinski I, LeBlanc ML, et al. Sequence analysis using logic regression. Genet Epidemiol 2001; 21(suppl 1):S626–31.
10. Ruczinski I, Kooperberg C, LeBlanc M. Exploring interactions in high-dimensional genomic data—an overview of logic regression, with applications. J Mult Anal 2004;90:178–95.
11. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, eds. Second International Symposium on Information Theory. Budapest, Hungary: Akademiai Kiado, 1973:267–81.
12. Stephens M, Smith M, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Gen 2001;68:978–89.
13. Dai JY, Ruczinski I, LeBlanc M, et al. Comparison of haplotype-based and tree-based SNP imputation in association studies. Genet Epidemiol (in press).
14. Agerholm-Larsen B, Nordestgaard BG, Tybjaerg-Hansen A. *ACE* gene polymorphism in cardiovascular disease: meta-analysis of small and large studies in whites. Arterioscler Thromb Vasc Biol 2000;20:484–92.
15. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Gen 2005;37:413–17.
16. Kooperberg C, Ruczinski I. Identifying interacting SNPs using Monte Carlo logic regression. Genet Epidemiol 2005;28: 157–70.