Tree-structured supervised learning and the genetics of hypertension

Jing Huang, Alfred Lin, Balasubramanian Narasimhan, Thomas Quertermous, C. Agnes Hsiung, Low-Tone Ho, John S. Grove, Michael Olivier, Koustubh Ranade, Neil J. Risch, and Richard A. Olshen

This paper is about an algorithm, FlexTree, for general supervised learning. It extends the binary tree-structured approach (Classification and Regression Trees, CART) although it differs greatly in its selection and combination of predictors. It is particularly applicable to assessing interactions: gene by gene and gene by environment as they bear on complex disease. One model for predisposition to complex disease involves many genes. Of them, most are pure noise; each of the values that is not the prevalent genotype for the minority of genes that contribute to the signal carries a “score.” Scores add. Individuals with scores above an unknown threshold are predisposed to the disease. For the additive score problem and simulated data, FlexTree has cross-validated risk better than many cutting-edge technologies to which it was compared when small fractions of candidate genes carry the signal. For the model where only a precise list of aberrant genotypes is predisposing, there is not a systematic pattern of absolute superiority; however, overall, FlexTree seems better than the other technologies. We tried the algorithm on data from 563 Chinese women, 206 hypotensive, 357 hypertensive, with information on ethnicity, menopausal status, insulin-resistant status, and 21 loci. FlexTree and Logic Regression appear better than the others in terms of Bayes risk. However, the differences are not significant in the usual statistical sense.

As interest in complex disease increases, there is increasing need for methodologies that address issues such as gene–gene and gene–environment interactions in a robust fashion. We report on a binary tree-structured classification tool, FlexTree, that addresses such needs. Our principal application has been to predict a complex human disease (hypertension) from single-nucleotide polymorphisms (SNPs) and other variables. FlexTree extends the approach of Classification and Regression Trees (CART; ref. 1); it retains CART’s simple binary tree structure. However, it differs in its ability to handle combinations of predictors. Furthermore, the construction of the tree respects family structures in the data. FlexTree involves coding categorical predictors to indicator variables, suitably scoring outcomes, backward selection of predictors by importance estimated from the bootstrap, and then borrowing from CART (1). We convert a problem of classification to one of regression without losing sight of classification. The approach we take resembles that of Zhang (2), although his focus is on different criteria for splitting, and he seems not to be concerned as we are about shaving the list of candidate features at each node. Each vector of predictor values can be located to a terminal node of the tree. Finding groups with high risk of disease may lend understanding to etiology and genetic mechanism. In most applications of classification trees, each split is on one feature. This seems inappropriate for polygenic disease, when no single gene is decisive, and the “main effect” may be a gene by environment interaction. With FlexTree as applied to data concerning certain Asian women (which will be described), we suggest complicated associations of hypertension and predictors. They include insulin resistance, menopausal status, and SNPs in a protein tyrosine phosphatase gene and the mineralocorticoid receptor gene. A particular strength of the approach taken here, as opposed to the approaches that involve schemes for voting among predictors or random selection of features, is its ability to make simple, clear choices of relevant features at each node. In particular, we find predictive, simple linear combinations of features, and we get at the issue of each genotype’s being predictive or a risk factor for an untoward outcome.

Algorithm

FlexTree borrows strength from both classification trees and regression; it creates a simple rooted binary tree with each split defined by a linear combination of selected variables. The linear combination is achieved by regression with optimal scoring. The variables are selected by a backward shaving procedure. Using a selected variable subset to define each split increases interpretability, improves robustness of the prediction, and prevents overfitting. FlexTree deals with additive and interactive effects simultaneously. Sampling units can be families or individuals, depending on the application.

Data Transformation.

In the situation of finding influential genes for a polygenic disease, the outcome is dichotomous disease status, affected versus disease-free, and the predictors are mutations at different loci. They are all qualitative. The response variable Y with L0 unordered classes and N observations is changed to an N by L0 dummy matrix with elements 0 or 1: Y ⇒ YN×L0. Each observation is transformed to an L0-dimensional indicator vector:

\[ \hat{y}_n \Rightarrow y_n \quad 1 \leq n \leq N \]

\[ y_n = (y_{n1}, y_{n2}, \ldots, y_{nl_0}) \quad y_{nl} = \begin{cases} 1, & \text{if } \hat{y}_n = c_l \\ 0, & \text{otherwise} \end{cases} \]

A predictor with K levels is represented by K − 1 instead of K columns in the design matrix to avoid singularity. One level of each variable is chosen as the baseline with the corresponding column removed. With J nominal categorical predictors Xi, 1 ≤ j ≤ J, each having N observations and Lj categories, the design matrix X is X = (1, X1, X2, \ldots, XJ) = XN×M, where M = 1 + Σ(Lj)

Abbreviations: SNP, single-nucleotide polymorphism; CART, Classification and Regression Trees; SAPPHIRE, Stanford Asian Pacific Program for Hypertension and Insulin Resistance.

© 2004 by The National Academy of Sciences of the USA
Regression and the Splitting Criteria. The regression framework is built on the transformed data. The binary splitting rule is defined as an inequality involving a linear combination of a subset of the predictors: \( X_{\text{select}} \beta \geq C \). The problem of classification is thus transformed to one of regression. Optimal scoring can be viewed as an asymmetric version of canonical correlation analysis and linear discriminant analysis (4). The optimal scores \( \Theta = \Theta_{m \times 1} = (\theta_1, \theta_2, \ldots, \theta_m)^T \) are determined as those that minimize the penalized sum of squares

\[
N^{-1}(|Y\Theta - X\beta|^2 + \beta^T \Omega \beta)
\]

under the constraint \(|Y\Theta|^2/N = 1\). \( \Omega \) is a penalty term introduced to avoid singularity, which occurs when there are many predictors involved or in the later stages of the partitioning, where the sample within a node is deliberately homogeneous. \( \Omega = \lambda I \), where we have taken \( \lambda = 0.001 \) in our computations, although obviously there could be other choices. Our motivation to use optimal scoring as a component of the algorithm is supported by findings of Hastie et al. (ref. 5, section 12.5.1). For any given score \( \Theta \), the penalized least-squares estimate of \( \beta \) is

\[
\hat{\beta} = (X^T X + \Omega)^{-1} X^T Y \Theta.
\]

Substitute Eq. 4 into formula 3, and the formula simplifies to

\[
1 - \frac{1}{N} [\Theta^T Y^T X (X^T X + \Omega)^{-1} X^T Y \Theta].
\]

\( \Theta \) that minimizes formula 5 is the eigenvector corresponding to the largest eigenvalue of the matrix.

\[
Y^T X (X^T X + \Omega)^{-1} X^T Y
\]

\( \Theta \) is standardized to satisfy \(|Y\Theta|^2/N = 1\). After the optimal scores are obtained, linear regression of the quantified outcome \( Z \) on \( X \) is applied:

\[
Z = Y\Theta = X\beta + \text{error}.
\]

is minimized, which entails estimated regression coefficients and outcome

\[
\hat{\beta} = (X^T X + \Omega)^{-1} X^T Z = (X^T X + \Omega)^{-1} X^T Y \Theta;
\]

\[
\hat{Z} = X(X^T X + \Omega)^{-1} X^T Z = X(X^T X + \Omega)^{-1} X^T Y \Theta.
\]

When \( X \) is full rank, simple linear regression is applied without a penalty term. A binary split is defined as \( Z = X\hat{\beta} \geq C \). \( C \) is chosen to maximize the impurity reduction

\[
\text{Optimal split} = \arg \max_n R_t - R_l - R_r,
\]

where \( R_t \) indicates the weighted generalized Gini index for the current node \( t \); \( R_l \) is one for the left daughter node of a given partition; and \( R_r \) is the one for the right daughter node (1). One critical question concerns how to choose the right subset of predictors on which to perform the regression. Another is how to measure the relative importance of each predictor involved.

“Backward Shaving” and Node-Specific Variable Ranking. Backward shaving begins with the full set of predictors. Selected proportions of them are successively shaved off until only a single predictor remains. This procedure produces nested families of predictors, from the full set to a single one. The shaving is based on the node-specific variable ranking defined by bootstrapped \( p \) values. Such \( p \) values are derived from a \( \chi^2 \) statistic as it applies to testing the null hypothesis that “related” regression coefficients are all zero:

\[
H_0: \beta_{(X)} = (0, 0, \ldots, 0) \quad \text{versus} \quad H_a: \beta_{(X)} \neq (0, 0, \ldots, 0).
\]

Here \( \beta_{(X)} \) indicates the subset of regression coefficients for predictor \( X \). We assume (without obvious loss) that the outcome is calculated as the appropriate entry in a vector of “optimal” scores, which are the coordinates of the first canonical variable that relates outcome and predictors, and \( \hat{\beta}_{(X)} \), the least-squares estimate of \( \beta_{(X)} \), has approximately a multinormal distribution

\[
\hat{\beta}_{(X)} \sim N(\mu, \Sigma),
\]

which implies that the solid ellipsoid of \( Z \) values satisfying \((Z - \mu)^T \Sigma^{-1} (Z - \mu) \leq \chi^2(\alpha)\) has probability about 1 - \( \alpha \). We use the bootstrap to estimate \( \mu \) and \( \Sigma \) and to evaluate variable importance. Observations are often correlated, and the predictors can be categorical. We first bootstrap families (or individuals, as the case may be) to obtain \( B \) independent bootstrap samples \((X_1, Y_1), (X_2, Y_2), \ldots, (X_B, Y_B)\). For each sample \((X_1, Y_1)\), we estimate the optimal score \( \Theta_{(1)} \) and the regression coefficients \( \beta_{(1)} \) as described in the previous section. We then compute the sample mean and sample covariance of \( \hat{\beta}_{(X)} \), from the \( B \) bootstrapped estimates \( \hat{\beta}_{(X)}^{(b)}_1, b = 1 \) to \( B \):

\[
\hat{\mu} = \text{Mean}(\hat{\beta}_{(X)}), \quad \hat{\beta}_{(X)}^{(b)}_1 = \frac{1}{B} \sum_{b=1}^B \hat{\beta}_{(X)}^{(b)};
\]

\[
\Sigma = \text{Cov}(\hat{\beta}_{(X)}) = \frac{1}{B - 1} \sum_{b=1}^B (\hat{\beta}_{(X)}^{(b)} - \hat{\beta}_{(X)})(\hat{\beta}_{(X)}^{(b)} - \hat{\beta}_{(X)}).
\]

Given \( \hat{\mu} \) and \( \Sigma \), the test statistic is

\[
X^2 = (\hat{\mu} - \mu)^T \Sigma^{-1} (\hat{\mu} - \mu) = \hat{\mu}^T \Sigma^{-1} \hat{\mu},
\]

which has approximately a \( \chi^2 \) distribution under the multinormal approximation. The \( p \) value derived from this test statistic corresponds to the \( 100(1 - p)\% \) ellipsoid-confidence-contour, where the boundary goes through the origin. In what follows, this is referred to as the \( T^2 \) \( \times \) \( p \) value. These \( p \) values give not only a measure of variable importance but also an order for shaving. We repeatedly shave off a certain number of “least important” variables until only the single most important variable is left, and create a nested sequence of variable subsets:

\[
S_{\text{full set}} = S^{(m)} \supset S^{(m-1)} \supset \cdots \supset S^{(2)} \supset S^{(1)} = \text{a single variable}.
\]

The subset \( S_1 = S^{(n)} \) that includes the least number of variables and yet achieves approximately the largest impurity reduction within a prespecified error margin is chosen. The error margin is set to be 5% of the largest impurity reduction.

\[
i_1 = \min_i \left[ \|G(S^{(i)}) \| \geq \left[ \max_{1 \leq j \leq m} G(S^{(j)}) - \text{error margin} \right] \right]
\]

Variables that are not included in \( S_1 \) are shaved off. In the case where \( S_1 = S^{(m)} \), the least significant variable is shaved off to keep the process going. Then \( S_1 \) is treated as the full set, and the
shaving procedure is repeated with new optimal scores to get an even smaller set. The procedure is continued until only one variable is left. This again gives us a sequence of nested variable subsets, full set = $S_0 \supset S_1 \supset S_2 \supset \ldots \supset S_{k-1} \supset S_k = \text{a single variable.}^{10}$ The optimal subset is determined by the mean of the cross-validated estimates; $k$-fold cross-validation ($k = 5$ or 10) is applied to each set in the nested sequence to estimate the impurity reduction associated with the subset.

$$\Delta R^v(S_t) = R_t - R^v_t(S_t) - R^v_t(S),$$

where $R_t$ is the impurity measure of the current node $t$ according to the cross-classified partition rule using variable subset $S_t$; $R^v_t(S_t)$ is the impurity measure of the left daughter node; $R^v_t(S)$ is the impurity measure of the right daughter node. Because single cross-validation is relatively unstable, it is repeated $M = 40$ times, each time with a different $k$-ary partition of the data. The means of the $M$ cross-validated estimates quantify the performance of each subset:

$$\overline{\Delta R^v(S_t)} = \frac{1}{M} \sum_{m=1}^{M} R^v_m(S_t).$$

The subset with the largest mean (within 1 SE) is chosen as the final variable subset to define the binary split. After the final set is chosen, a permutation test on the quantity given by Eq. 18 is performed to test whether there is any (linear) association between the outcome and the selected predictors. If the test statistic is larger than a preassigned threshold, we conclude there is still significant association and continue splitting the node. Otherwise, we stop. Beyond this subsidiary stopping rule, procedures for growing and pruning trees are exactly as in CART.

**Missing Value Imputation.** Genetic data are prone to problems with missing values. To avoid losing valuable information, we try to retain observations with some missing values in the analysis. Because genetic predictors are often at least moderately correlated, we use these correlations at least implicitly to establish a predictive model for the one with missing values, using all of the other predictors. In particular, we chose the standard classification tree method to serve as such a predictive model. It is flexible, being applicable to both categorical and continuous variables without distributional assumptions. Also, it deals with missing values effectively through surrogate splits and can succeed in situations where many predictors in the training set have missing values (5). Strictly speaking, our imputation scheme is valid only if data are “missing at random.”

**Testing Association, Genotype by Disease Status.** Given a SNP with three genotypes and disease status (yes or no) for each subject in a list of subjects, one can form the usual $2 \times 3$ table and ask whether the two variables are associated. The scheme by which SAPPHIRE subjects were gathered entails that in all but families with a single qualifying sib, observations within families are not independent (SAPPHIRE is described in a subsequent section.) We developed a test for association that respects family structures. This approach combines the bootstrap and permutation testing. Begin with the former. Suppose there are $F$ families and $N$ subjects. Pick at random $F$ people from among the $N$ with replacement. For each person chosen, include all sibs from his or her family in the bootstrap sample. This procedure has the effect of sampling families with probabilities proportional to numbers of eligible sibs. For each chosen family, permute the disease status when possible, always keeping total marginal numbers by genotype and phenotype constant. Now, using all “bootstrap families,” those that allowed nontrivial permutation and those that did not, compute the $\chi^2$ statistic for the $2 \times 3$ table. Repeat the process of bootstrapping $B$ times ($B = 1,000$ in our case), followed by permuting (when possible) and computation of the $\chi^2$ statistic each time. With $B$ bootstrap samples, there are now $B + 1$ values of the $\chi^2$ statistic, the “1” being the value for the original data. Order these $B + 1$ values from largest to smallest. If the “true” $\chi^2$ is say the 7th on the list, then the $p$ value for testing the null hypothesis of “no association” between genotype and phenotype is $7/(B + 1)$. Because this $p$ value depends on SNPs only marginally, it seems intuitively clear that attained significance levels should be higher than for the $T^2$-based values for regression coefficients associated with FlexTree. They are indeed much higher, as is shown.

**Simulations**

We report on simulations of the accuracy of FlexTree in associating a “disease process” with numbers of mutations. There are two standard models, each involving 30 loci of a diploid organism. For each model, the “probability of disease” is a nondecreasing function of the numbers of genes mutated away from prevalent type at six key sites. For each model this number is an integer between 0 and 12. Genotype at any of the remaining 24 loci is unrelated to the probability of disease. For Model A (for “additive”), the effects of mutations are additive in the following sense. If $M$ denotes the random numbers of mutations at the six key sites, and $D = [\text{Disease}]$, then with indicator function notation,

$$P(D|M) = 0.1 I_{M=4} + 0.4 I_{M=6} + 0.4 I_{M=7} + 0.1 I_{M=9}.$$  

[20]

The other model is one for which there is an utterly epistatic impact of numbers of mutations on the conditional probability of disease (7). Thus, for Model E (for “exact”),

$$P(D|M) = 0.1 + 0.8 I_{M=12}.$$  

[21]

Denote by $m_1, m_2, \ldots, m_6$ the numbers of mutations away from most prevalent type at (respective) sites 1, 2, 3, 4, 5, 6. Then under Model A, $m_1, \ldots, m_6$ are independently and identically distributed as

$$P(m_1 = j) = \left(\frac{2}{j}\right) 0.5^2 \quad j = 0, 1, 2 \quad i = 1, 2, \ldots, 6.$$  

[22]

So $M \sim \text{Binomial}(12, 0.5)$. Under Model E, $m_1, \ldots, m_6$ are also independently and identically distributed, but the distribution is different:

$$P(m_1 = j) = \left(\frac{2}{j}\right) 0.9^j 0.1^{6-j} \quad j = 0, 1, 2 \quad i = 1, 2, \ldots, 6.$$  

[23]

So $M \sim \text{Binomial}(12, 0.9)$. The unconditional probability of disease under A is $P(D) = 0.5$ because $P(D|M) = P(D|12 - M)$ and $M = 12 - M$. Under E, $P(D) = 0.9^{0.9} 0.1^{6} + (1 - 0.9^{0.9}) 0.1 = 0.326$. For both models, at the 24 “irrelevant” loci, the genotype is prevalent type or not with equal probabilities, and these loci are independent and identically distributed. It is far easier to have mutations with Model E than with Model A. With Model E, only an “exact set” of mutations entails enhanced risk of disease, whereas with Model A, as long as the mutations accumulate to pass a certain threshold, they
will entail high risk. For both models, the Bayes rule can be described by a binary tree with but a single “linear combination split,” which is exactly the case in FlexTree. For each model, we simulate a learning sample of 200 observations, each with information on 30 loci. In both cases, FlexTree produces a tree with one split with the exact six key loci being used correctly to define the linear inequality (i.e., the splitting rule) (see Table 1).

In what follows, we look at a wide range of possibilities, particularly at cases of 4, 8, 10, 12, 14, and 16 disease genes. In Table 2, the numbers in the first row in each cell are the estimated mean of the risk from 200 simulations, the first one for Model A and the second one for Model E. The numbers in the parentheses are the estimated standard deviations, accordingly.

Comparisons of Bayes risk for different estimators cannot be done by linear inequality (i.e., the splitting rule) (see Table 1).

Table 1. FlexTree performance on Models A and E with six effective genes

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model A</th>
<th>Model E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T^2$</td>
<td>$p$ value</td>
</tr>
<tr>
<td>1</td>
<td>0.237</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.283</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.249</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.234</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.250</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.254</td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity: 77/89 = 0.865
Specificity: 95/111 = 0.856

Table 2. Comparison for Models A and E on Bayes risk

<table>
<thead>
<tr>
<th>Model</th>
<th>FlexTree</th>
<th>CART</th>
<th>QUEST</th>
<th>Logic Regression</th>
<th>Random Forest</th>
<th>Bayes risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 genes</td>
<td>0.207, 0.242</td>
<td>0.324, 0.490</td>
<td>0.259, 0.427</td>
<td>0.368, 0.543</td>
<td>0.270, 0.216</td>
<td>0.180, 0.148</td>
</tr>
<tr>
<td></td>
<td>(0.030, 0.044)</td>
<td>(0.031, 0.040)</td>
<td>(0.032, 0.044)</td>
<td>(0.033, 0.061)</td>
<td>(0.031, 0.014)</td>
<td></td>
</tr>
<tr>
<td>6 genes</td>
<td>0.197, 0.177</td>
<td>0.335, 0.173</td>
<td>0.245, 0.254</td>
<td>0.327, 0.258</td>
<td>0.276, 0.223</td>
<td>0.186, 0.147</td>
</tr>
<tr>
<td></td>
<td>(0.027, 0.039)</td>
<td>(0.033, 0.038)</td>
<td>(0.030, 0.039)</td>
<td>(0.032, 0.045)</td>
<td>(0.030, 0.012)</td>
<td></td>
</tr>
<tr>
<td>8 genes</td>
<td>0.193, 0.219</td>
<td>0.369, 0.256</td>
<td>0.282, 0.409</td>
<td>0.357, 0.177</td>
<td>0.298, 0.219</td>
<td>0.147, 0.149</td>
</tr>
<tr>
<td></td>
<td>(0.029, 0.038)</td>
<td>(0.034, 0.038)</td>
<td>(0.032, 0.041)</td>
<td>(0.031, 0.039)</td>
<td>(0.033, 0.014)</td>
<td></td>
</tr>
<tr>
<td>10 genes</td>
<td>0.186, 0.206</td>
<td>0.362, 0.241</td>
<td>0.242, 0.219</td>
<td>0.384, 0.374</td>
<td>0.285, 0.227</td>
<td>0.152, 0.146</td>
</tr>
<tr>
<td></td>
<td>(0.028, 0.039)</td>
<td>(0.034, 0.038)</td>
<td>(0.033, 0.043)</td>
<td>(0.033, 0.058)</td>
<td>(0.030, 0.013)</td>
<td></td>
</tr>
<tr>
<td>12 genes</td>
<td>0.232, 0.225</td>
<td>0.405, 0.285</td>
<td>0.240, 0.384</td>
<td>0.357, 0.308</td>
<td>0.276, 0.183</td>
<td>0.175, 0.146</td>
</tr>
<tr>
<td></td>
<td>(0.030, 0.039)</td>
<td>(0.037, 0.043)</td>
<td>(0.030, 0.045)</td>
<td>(0.034, 0.049)</td>
<td>(0.030, 0.011)</td>
<td></td>
</tr>
<tr>
<td>14 genes</td>
<td>0.234, 0.221</td>
<td>0.368, 0.289</td>
<td>0.230, 0.261</td>
<td>0.369, 0.299</td>
<td>0.274, 0.310</td>
<td>0.132, 0.144</td>
</tr>
<tr>
<td></td>
<td>(0.029, 0.036)</td>
<td>(0.036, 0.038)</td>
<td>(0.030, 0.041)</td>
<td>(0.037, 0.050)</td>
<td>(0.032, 0.011)</td>
<td></td>
</tr>
<tr>
<td>16 genes</td>
<td>0.290, 0.180</td>
<td>0.417, 0.297</td>
<td>0.267, 0.246</td>
<td>0.408, 0.279</td>
<td>0.302, 0.178</td>
<td>0.156, 0.143</td>
</tr>
<tr>
<td></td>
<td>(0.032, 0.038)</td>
<td>(0.035, 0.041)</td>
<td>(0.029, 0.045)</td>
<td>(0.033, 0.046)</td>
<td>(0.034, 0.012)</td>
<td></td>
</tr>
</tbody>
</table>

The numbers in the first row in each cell are the estimated mean of the risk from 200 simulations, the first one for Model A and the second one for Model E. The numbers in the parentheses are the estimated standard deviations, accordingly.
FlexTree to SAPPHIRe data and try to identify important genes that influence hypertension. SAPPHIRe data consist of “affected” sib-pairs from Taiwan, Hawaii, and Stanford. Both concordant and discordant sib-pairs are included, although always the proband was hypertensive (13, 14). History of hypertension was obtained by interviewing family members. Blood lipid profiles were collected from medical records. Genetic information was obtained by using fluorogenic probes (Taqman from Applied Biosystems and Invader from Third Wave Technologies). This data set resembles many others in genetics and epidemiology because selection bias is an issue. Sample families are guaranteed to have at least one hypertensive sibling, which is patently not the case in general. There are data on 563 Chinese women (206 hypertensive, 357 hypertensive) with some information on SNPs on 21 distinct loci, menopausal status, insulin resistant status, and ethnicity. All SNP data are in Hardy-Weinberg equilibrium. Three SNPs are X-linked. We examine only women because one fundamental hypothesis of the SAPPHIRe project is that insulin resistance is a predisposing condition for hypertension, that is, an intermediate phenotype. See ref. 15. In work not yet published we have found that in SAPPHIRe the relationship between insulin resistance and hypertension is stronger in women than in men, with the trend driven by premenopausal women. Insulin resistance was quantified by a k-means clustering approach, as in ref. 16, applied to lipid profiles and metabolic data after age and body mass index were regressed out (up to and including quadratic terms). Two clusters were found, one clearly insulin resistant and the other clearly not.

The data first were processed for missing values by a method of imputation as cited above. Then FlexTree was applied, with equal prior probabilities and misclassification cost for hypotensives twice that for hypertensives. The final tree has three splits (Fig. 1). Among all factors, the most important is menopausal status. It is the sole condition for hypertension, that is, an intermediate phenotype. The number to the left of the slash is the number of hypotensives; the number to the right is the number of hypertensives. Here we assume equal prior probabilities and misclassification cost for hypotensives twice that for hypertensives.

![Fig. 1. Application of FlexTree to Chinese women in SAPPHIRe. The oval and rectangles, respectively, indicate internal and terminal nodes. The label assigned to each node is determined so as to minimize the misclassification cost. The number to the left of the slash is the number of hypotensives; the number to the right is the number of hypertensives.](image)

Table 3. Postmenopausal Chinese women, split two (SNP for nonprevalent type)

<table>
<thead>
<tr>
<th>Variable</th>
<th>$T^2$</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyp11B2.5INV (1/2 P, 2/2 R)</td>
<td>0.0144</td>
<td>0.369</td>
</tr>
<tr>
<td>Cyp11B2INV (R)</td>
<td>0.0242</td>
<td>0.431</td>
</tr>
<tr>
<td>Insulin Resistance (+, R)</td>
<td>0.0256</td>
<td>0.174</td>
</tr>
<tr>
<td>MRL12V (R)</td>
<td>0.0753</td>
<td>0.601</td>
</tr>
<tr>
<td>Cyp11B2X1INV (P)</td>
<td>0.117</td>
<td>0.466</td>
</tr>
<tr>
<td>AVPR2G12E (R)</td>
<td>0.162</td>
<td>0.408</td>
</tr>
<tr>
<td>AGT2R1A1166C (R)</td>
<td>0.220</td>
<td>0.579</td>
</tr>
<tr>
<td>HUT2SNP7 (R)</td>
<td>0.248</td>
<td>0.469</td>
</tr>
<tr>
<td>CD36x2aINV (R)</td>
<td>0.301</td>
<td>0.476</td>
</tr>
<tr>
<td>PTP114INV (P)</td>
<td>0.320</td>
<td>0.713</td>
</tr>
<tr>
<td>AGT2R2C1333T (P)</td>
<td>0.622</td>
<td>0.642</td>
</tr>
</tbody>
</table>

Permutation $t = 2.113$

$T^2$ $p$ values were computed as in Eq. 15 and following. $\chi^2$ $p$ values were computed as in Testing Association Genotype by Disease Status. +, positive; $R$, risk factor; $P$, protective factor. The notation of permutation $t$ is explained in the text after Eq. 19.

A comparison of FlexTree and several other well known classification techniques on fivefold cross-validation is summarized in Table 5. Like FlexTree, CART and QUEST are two classification tools based on recursive partitioning. The main difference lies in how each approach defines a binary split: we use a variable selection procedure whereas most applications of CART use a single predictor; QUEST uses a linear inequality involving all candidate predictors to define the split (8, 9). Logic Regression is an additive model developed recently (10). It deals

Table 4. Premenopausal Chinese women, split three (SNP for nonprevalent type)

<table>
<thead>
<tr>
<th>Variable</th>
<th>$T^2$</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insulin Resistance (+, R)</td>
<td>0.679</td>
<td>0.141</td>
</tr>
<tr>
<td>AGT2R1A1166C (R)</td>
<td>0.702</td>
<td>0.211</td>
</tr>
<tr>
<td>AVPR2G12E (P)</td>
<td>0.759</td>
<td>0.797</td>
</tr>
<tr>
<td>PTP114INV (R)</td>
<td>0.930</td>
<td>0.460</td>
</tr>
</tbody>
</table>

Permutation $t = 4.739$
with complicated interactions among multiple binary predictors and serves as an alternative approach for studying associations with complicated interactions among multiple binary predictors.

Table 5. Comparison among seven methods on Chinese women

| True/Classified | FlexTree | Hyper | Hypo | Sensitivity = 233/357 = 0.653 | Specificity = 112/206 = 0.544 | Misclass = 124 + 2 × 94 = 312 | $R^2 = 0.554$, $SE(R^2) = 0.026$
|-----------------|----------|-------|------|-------------------------------|-------------------------------|-------------------------------|-----------------------------|
| True/Classified | Classification (Salford System) | Hyper | Hypo | Sensitivity = 0 | Specificity = 0 | Misclass = 357 | $R^2 = 0.634$, $SE(R^2) = 0.034$
| True/Classified | Logic Regression | Hyper | Hypo | Sensitivity = 203/357 = 0.569 | Specificity = 128/206 = 0.621 | Misclass = 154 + 2 × 78 = 310 | $R^2 = 0.551$, $SE(R^2) = 0.023$
| True/Classified | Bagging (50 bootstrap samples) | Hyper | Hypo | Sensitivity = 155/357 = 0.434 | Specificity = 142/206 = 0.689 | Misclass = 202 + 2 × 64 = 330 | $R^2 = 0.586$, $SE(R^2) = 0.024$
| True/Classified | MART | Hyper | Hypo | Sensitivity = 42/357 = 0.118 | Specificity = 202/206 = 0.987 | Misclass = 315 + 2 × 32 = 323 | $R^2 = 0.574$, $SE(R^2) = 0.025$
| True/Classified | Random Forest | Hyper | Hypo | Sensitivity = 230/357 = 0.644 | Specificity = 95/206 = 0.461 | Misclass = 127 + 2 × 111 = 349 | $R^2 = 0.620$, $SE(R^2) = 0.026$

Most often, by using such a “committee” method, the accuracy of classification can be improved substantially, yet the price paid for such improved prediction is interpretability because there is no longer a single model. We need to bear in mind that comparisons of Bayes risk for different estimators cannot be done precisely on the basis of summary statistics presented here. Data were partitioned into five exhaustive, disjoint groups for purposes of fivefold cross-validation. Five times, as 20% in succession were held out as “test sample,” all procedures were computed for the 80% that comprised the “learning sample.” Results were then averaged. Concerns like those that applied to comparing simulated data apply nearly verbatim here. FlexTree was superior to CART, with attained significance 0.06. The same applied to QUEST. In neither case did the competitive procedure partition the data at all, so comparisons are really with the “no data Bayes rule” in the sense of ref. 1. No other comparisons with FlexTree were “significant,” although FlexTree and Logic Regression seem better than the others.

We once thought that the impact of our SNP genotypes, if they bear on hypertensive status, should predispose Chinese and Japanese women in the same way. We have described difficulties with the SAPPHIRE sampling scheme already. Recruitment in Taiwan focused far more on hypertensive sibs of hypertensive probands than did recruitment in Hawaii. The majority of our Chinese women were from Taiwan and the majority of Japanese women were from Hawaii. No matter the “true” fractions of hype and hypo within families in the two groups, the prevalence of hypertensives in our sample was far higher for the Japanese than for the Chinese. There are in total 161 samples, 23 hypertensive and 138 hypertensive, which is unlikely to be the distribution in the general population. Moreover, classifying with the same products of priors and misclassification costs (2:1 in favor of hypertensives in the Chinese) produced nothing of interest in the smaller Japanese group. When the ratio was changed to 6:1, a story emerged. The final tree has two splits. The classification was reasonably good, with cross-validation showing 30 misclassified among 161. Given that the Japanese were older than the Chinese, it is not surprising that by and large, the SNPs that figure in our two-split tree are nearly all SNPs that figure in the split of postmenopausal Chinese women. As one might expect, age did not figure in either split for the Japanese. As with the Chinese, there was a wide difference between significance as judged by Hotelling’s $T^2$ and by bootstrap/permutation $\chi^2$, implying that any impact of genes is additive. For details see Supporting Elaboration.

We thank the subjects for participating in this study; Stephen Mockrin and Susan Old of the National Heart, Lung, and Blood Institute; and other members of the SAPPHIRE program. This work was supported by National Institutes of Health Grant U01 HL54527-0151 and National Institute for Biomedical Imaging and Bioengineering/National Institutes of Health Grant R01 EB002764-8.