Classification and Regression Tree Analysis in Public Health: Methodological Review and Comparison With Logistic Regression

Stephenie C. Lemon, Ph.D., Jason Roy, Ph.D., and Melissa A. Clark, Ph.D.
Brown University School of Medicine

Peter D. Friedmann, M.D., M.P.H.
Brown University School of Medicine and
Rhode Island Hospital, Providence, RI

William Rakowski, Ph.D.
Brown University School of Medicine

ABSTRACT

Background: *Audience segmentation strategies are of increasing interest to public health professionals who wish to identify easily defined, mutually exclusive population subgroups whose members share similar characteristics that help determine participation in a health-related behavior as a basis for targeted interventions. Classification and regression tree (C&RT) analysis is a nonparametric decision tree methodology that has the ability to efficiently segment populations into meaningful subgroups. However, it is not commonly used in public health.* ***Purpose:*** *This study provides a methodological overview of C&RT analysis for persons unfamiliar with the procedure.* ***Methods and Results:*** *An example of a C&RT analysis is provided and interpretation of results is discussed. Results are validated with those obtained from a logistic regression model that was created to replicate the C&RT findings. Results obtained from the example C&RT analysis are also compared to those obtained from a common approach to logistic regression, the stepwise selection procedure. Issues to consider when deciding whether to use C&RT are discussed, and situations in which C&RT may and may not be beneficial are described.* ***Conclusions:*** *C&RT is a promising research tool for the identification of at-risk populations in public health research and outreach.*

(**Ann Behav Med** 2003, *26*(3):172–181)

INTRODUCTION

Decision tree methods, also called *recursive partitioning*, are analytic strategies that were developed as a tool to classify or segment target audiences for the purposes of product marketing. Classification and regression tree (C&RT) analysis is a type of decision tree methodology that was first developed by Breiman and colleagues (1). Since its development, C&RT analysis has been the focus of a recent textbook (2) and several methodological monographs and manuscripts, including the examples listed (3–25).

This article was prepared as part of Stephenie Lemon's doctoral dissertation at Brown University.

Reprint Address: S. C. Lemon, Ph.D., Division of Preventive and Behavioral Medicine, University of Massachusetts Medical School, Worcester, MA 01655. E-mail: Stephenie.Lemon@Umassmed.edu

Audience Segmentation in Public Health

The purpose of audience segmentation strategies in public health and health behavior research is to identify easily defined, mutually exclusive population subgroups whose members share characteristics that are important barriers to or facilitators of the health-related behavior of interest. Each population subgroup should also be reachable through similar outreach and intervention strategies (26). C&RT analysis has the ability to efficiently segment populations into meaningful subsets. This allows researchers to easily identify segments of a population that are most likely to engage or not engage in a particular health behavior and to efficiently target and maximize the distribution of public health resources.

C&RT Analysis in Public Health

C&RT methodology has increasingly been applied to health sciences and clinical research (27–45) and, to a much lesser extent, in public health. Recent examples of public health analyses using C&RT include epidemiologic studies assessing risk factors for mortality and morbidity from specific diseases (46–53), comparisons of the cost-effectiveness of colorectal cancer screening technologies (54) and influenza treatment strategies (55), the development of screening and diagnostic tools (56), and the assessment of predictors of utilization of medical procedures such as caesarian sections (57). Studies using C&RT as a basis for designing targeted behavioral interventions are lacking. Possible reasons for the infrequent use of C&RT by behavioral researchers may be a general lack of awareness of the utility of C&RT procedures and, among those who are aware of these procedures, an uncertainty concerning their statistical properties.

Traditional Statistical Methods

Traditionally, health behavior research has relied on two analytic approaches to determine high-risk segments of the population. Both are limited in their ability to easily segment populations into distinct subgroups whose members share common characteristics that influence participation in a particular health-related behavior. The first method involves simply estimating the percentage of the dependent variable among strata of categorical covariates. Although this bivariate method is useful for very basic descriptions of health behaviors, it does not allow

the simultaneous consideration of multiple independent variables. The second method involves regression modeling, usually using logistic regression because outcome variables of interest are often dichotomous (58). Regression models with main effects allow investigators to test whether a given correlate and a dependent measure are associated, while controlling for confounding factors. Regression models are largely used to determine the average effect of an independent variable on a dependent variable. Thus, when interventions are developed from regression model results, they are geared toward the average member of the population, without consideration of special needs of population subgroups (59). Regression modeling does allow for the testing of statistical interactions among independent variables, which assess differences in the effects of one or more independent variable according to levels of another independent variable. However, statistical interactions can be difficult to interpret, particularly when three or more variables are assessed at a time.

Overview and Study Purpose

The purpose of this investigation is twofold. The first aim is to provide an overview of C&RT analysis, both conceptually and statistically. A review of the C&RT technique is provided using a generic example, followed by an in-depth description of the statistical properties of C&RT analysis. We also provide an example of a C&RT analysis and discuss how the results are interpreted. We then validate our results with those obtained from a logistic regression model, which was created to replicate the C&RT findings. The second aim is to compare the results obtained from the example C&RT analysis to those obtained from a common approach to logistic regression, the stepwise selection procedure. We illustrate each of these objectives using influenza vaccination receipt in older individuals (age 65 and older) as an example. Readers who are interested in a comprehensive tutorial on C&RT are referred to the textbooks by Breiman and colleagues (1) and by Zhang and Singer (2).

C&RT ANALYSIS

Overview

C&RT is a nonparametric statistical procedure that identifies mutually exclusive and exhaustive subgroups of a population whose members share common characteristics that influence the dependent variable of interest. C&RT produces a visual output that is a multilevel structure that resembles branches of a tree.

A generic illustration of C&RT output is presented in Figure 1. Classification and regression trees begin with one "node," or group, containing the entire sample, called a *parent node*, which is illustrated in Figure 1 as Node 1. The C&RT procedure examines all possible independent, or splitting, variables and selects the one that results in binary groups that are most different with respect to the dependent variable, according to a predetermined splitting criterion (described later). The parent node then branches into two descendent, or child, nodes according to the independent variable that was selected. C&RT only splits each

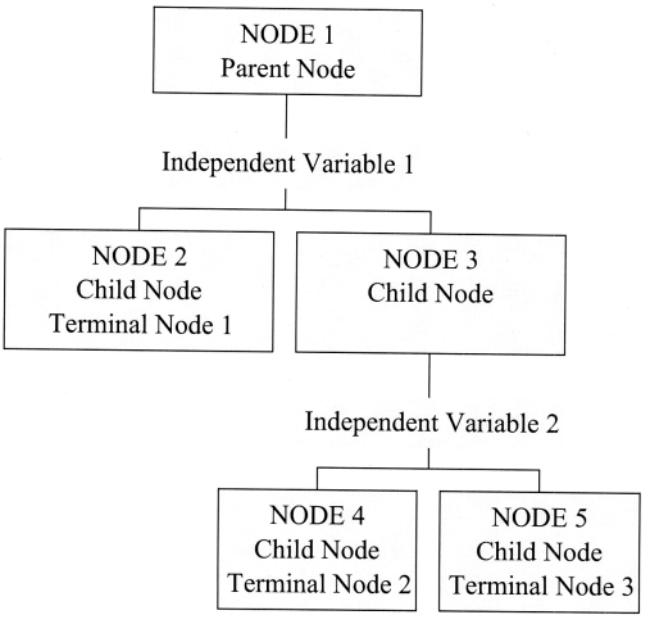

FIGURE 1 Example of classification and regression tree output.

parent node into two child nodes. In Figure 1, Node 1 split into Nodes 2 and 3 according to a level of Independent Variable 1.

Within each of these two child nodes, the tree-growing methodology continues by assessing each of the remaining independent variables to determine which variable results in the best split according to the chosen splitting criterion. Thus, each of the two child nodes becomes a parent node to the two groups into which it splits. In Figure 1, Node 3 splits according to Independent Variable 2 into Nodes 4 and 5. Thus, Node 3 is a child node of Node 1 and a parent node to Nodes 4 and 5. The procedure continues through each branch of the tree until a stopping rule (defined later) is reached. At the point that no further split is made, a terminal node is created. In Figure 1, Nodes 2, 4, and 5 are considered terminal nodes. Terminal nodes are mutually exclusive and exhaustive subgroups of the population.

The dependent, or target variable, can be either categorical (i.e., classification tree) or continuous (i.e., regression tree). In a classification tree, which is the type of analysis used in the example presented in this article, the probability of having the dependent measure is estimated among those within each node. In regression trees, the average value of the dependent measure is estimated among those within each node. Independent variables can be any combination of categorical and continuous variables. However, splits are always binary. In the case of an ordinal or continuous variable, C&RT searches through the full range of values and finds the best combination of categories or cut-point according to the predetermined splitting criterion.

Splitting Criteria

Many different splitting criteria have been proposed. They all begin by defining the impurity of a node. A node that has no impurity would have no variability in the dependent variable within that node (i.e., all 0s or all 1s). The highest amount of im-

purity is when $p_{i/j} = .5$, where $p_{i/j}$ is the probability that the dependent variable is equal to i in Node j, where i can take values 0 or 1. Splitting criteria are therefore based on functions of $p_{i/j}$, also known as *impurity functions*. Impurity functions are symmetrical, concave functions with maximum value at $p_{i/j} = .5$, and value zero when there is no impurity (i.e., when $p_{i/j} = 0$ or 1). Examples of impurity functions include Gini, entropy, and minimum error (2). Regardless of which impurity function is chosen, the splitting criterion selects the split that has the largest difference between the impurity of the parent node and a weighted average of the impurity of the two child nodes.

In the current example, the splitting criterion used was a statistical technique called the Gini improvement measure, which is the most commonly used method for determining the optimal split of a parent node into two child nodes when the dependent variable is categorical (1,2). The Gini impurity function for Node j is $2p_{1/j}(1 - p_{1/j})$. Gini can be calculated as follows:

1. The first step involves calculating Gini impurity function for the parent node, which is sometimes referred to as the Gini diversity index. This can be defined as follows:

$$\text{Diversity index} = 1 - \Sigma\, p_{i/j}^2 = 2p_{1/j}(1 - p_{1/j}).$$

2. The second step involves calculating the Gini diversity index for each of the two child nodes into which the parent node splits.
3. The third step involves calculating the weighted average, according to the proportion of the parent node that is included in each child node, of the Gini diversity indexes for each of the child nodes. This can be obtained by solving the following equation:

$$\text{Weighted diversity index} = [(p_1)(\text{diversity index}_1)] + [(p_2)(\text{diversity index}_2)],$$

where p_1 and p_2 refer to the proportions of the parent node that are included in the respective child nodes.

4. The last step requires calculating the Gini improvement measure, which is equal to the following:

$$\text{Improvement measure} = \text{diversity index of parent node} - \text{weighted diversity index}.$$

Larger values of the Gini improvement measure indicate greater difference with respect to the prevalence of the dependent measure in the two child nodes. The independent variable whose split provides the largest value of the improvement measure is selected for splitting at each step.

To illustrate, consider an extreme hypothetical example in which the parent node includes 1,000 people with a 50% prevalence of dependent variable *Y*. The parent node splits into two child nodes according to independent variable *X*. The first includes 500 people with a 10% prevalence of dependent variable *Y*. The second includes 500 people with a 90% prevalence of dependent variable *Y*. In this instance, the diversity index equals .50 for the parent node (Step 1) and .18 for each of the child nodes (Step 2). For each child node, the proportion of participants included from the parent node is .50. Thus, the weighted diversity index for the child nodes also equals .18 (Step 3). The Gini improvement measure (Step 4) would be equal to .50 minus .18, or .32.

Table 1 demonstrates a comparison of results that would be obtained from Gini improvement measures using C&RT and probability values of chi-square statistics in several hypothetical situations. The purpose of this table is to allow the reader who is comfortable interpreting probability values to compare values of the Gini improvement measure obtained in the same situations. For each test, we assume the null hypothesis of equal proportions of a dependent variable across binary strata of an independent variable. Consider three possible total sample sizes: $N = 100$, $N = 1{,}000$, and $N = 10{,}000$. In each of these samples the association between a binary dependent variable and a binary independent variable is being assessed. Thus, the data are split into two groups, Child Node 1 and Child Node 2, according to the two levels of the independent variable. The prevalence of the dependent variable is assumed to be 25% in Child Node 1 and 35% in Child Node 2 in each of these three samples. Next con-

TABLE 1

Comparison of Gini Improvement Measures Obtained From Classification and Root Tree Analysis With Probability Values Obtained From Chi-square Statistics for Different Sample Sizes and Proportions of Total Sample

Child Node 1—Proportion with DV = .25/ Proportion of Total Sample	*Child Node 2 —Proportion With DV = .35*																	
	Proportion of Total Sample (N = 100)						*Proportion of Total Sample (N = 1,000)*						*Proportion of Total Sample (N = 10,000)*					
	.10		*.25*		*.50*		*.10*		*.25*		*.50*		*.10*		*.25*		*.50*	
	Gini	*p*	*Gini*	*p*	*Gini*	*p*	*Gini*	*p*	*Gini*	*p*	*Gini*	*p*	*Gini*	*p*	*Gini*	*p*	*Gini*	*p*
.10	.0050	.63	.0053	.55	.0059	.53	.0050	.13	.0053	.06	.0059	.04	.0050	< .001	.0053	< .001	.0059	< .001
.25	.0031	.57	.0050	.44	.0015	.40	.0031	.07	.0050	.01	.0015	.004	.0031	< .001	.0050	< .001	.0015	< .001
.50	.0005	.54	.0069	.38	.0050	.27	.0005	.05	.0069	.005	.0050	.0005	.0005	< .001	.0069	< .001	.0050	< .001

Note. DV = dependent variable.

sider that the association of the independent variable and dependent variable in each of these three samples is observed within a subset of the entire data. In the case of the C&RT analysis, this is the equivalent of a node that is at a lower level of the tree (i.e., not the first parent node, which contains the entire sample). In the case of the chi-square analysis, it is the equivalent of an analysis that is conducted in a subset of the population. In Table 1, the columns and row labeled *Proportion of Total Sample* refer to the proportion of the total sample included in the given Child Node in the stratified analysis.

The Gini improvement measure is largely determined by the difference in the proportions of the dependent variable in the child nodes. It is also influenced by the difference between the proportions of participants from the parent node that are contained in the child nodes. It is also slightly affected by where the proportions of the dependent variable lie between 0 and 1. However, it is not affected by sample size. For example, consider the case in Table 1 in which the total sample is equal to 1,000. When the proportion of the total sample in Child Node 1 is .10, this is equal to a cell size of 100, and when the proportion of the total sample for Child Node 2 is .50, this is equal to a cell size of 500. The Gini improvement measure in this situation is equal to .0059, and the probability value for the chi-square statistic in this situation is equal to .04. Under these same circumstances with total sample sizes of 100 or 10,000, the Gini improvement measure is identical, whereas the chi-square probability value, which is dependent on sample size, is different. Now suppose that we assume the prevalence of the dependent variable is 45% in Child Node 1 and 55% in Child Node 2 (data not shown). The Gini improvement measure in this situation is .0032, which differs somewhat from the situation in which the prevalence estimates were 25% and 35%.

Classification Error

As with any form of statistical inference, it is important to understand the uncertainty in the inference. In regression models this takes the form of standard errors of the parameter estimates. Measures of variability are more complicated in C&RT analysis. Clearly, different random samples from the same population would produce different trees. There are several common ways for estimating how different those trees would be. One way is to use a subset of the data as a test sample. A tree T is generated for the test sample and then is applied to the other data. A measure of misclassification cost $R(\mathrm{T})$ can be estimated on the remaining data. An alternative to using a test sample is to use *k*-fold cross-validation. In *k*-fold cross-validation the data are broken into *k* subsets. A tree is calculated using all of the data except from one of the subsets. The tree is applied to the remaining subset, and the misclassification cost is calculated. The process is then repeated for the remaining subsets. Several stopping and pruning procedures utilize information from the corresponding estimate of classification error.

Stopping Rules

C&RT allows the investigator to a priori define criteria for stopping the growing procedure, called *stopping rules*. These are specified so the investigator can control how large the tree becomes, and they establish the minimal degree of statistical difference between groups that is considered meaningful. The first stopping rule requires defining the minimum number of individuals in the child nodes or in the terminal nodes. The second stopping rule requires defining the maximum number of levels to which the tree can grow, which allows the investigator to decide the maximum number of independent variables that can define a single terminal node. The third stopping rule involves predetermining the minimum value of the splitting criterion.

Pruning

A concern with growing trees based entirely on the use of stopping rules is that important associations may be missed because of stopping too soon. Another approach to growing a tree, pruning, addresses this. The first step is to grow a drastically large tree that includes many levels and nodes, possibly to the point where there are just a few cases per terminal node. The next step is to prune the tree back. There are a number of different criteria that could be used for selecting among all possible pruned subtrees, but most involve some measure of misclassification cost (i.e., goodness of fit). Misclassification cost is analogous to residual sums of squares in regression. A number of other pruning rules have been proposed, including minimum cost-complexity pruning (1) and least absolute shrinkage and selection operator (60). We now briefly describe one such pruning criteria: the one standard error (*SE*) rule. Consider a sequence of candidate trees, T_1, T_2, T_3, …, that have a progressively smaller number of terminal nodes. The one SE rule will select the smallest tree that satisfies $R(\mathrm{T}_k) \leq R(\mathrm{T}_j) + SE[R(\mathrm{T}_j)]$, where T_j is the tree with the minimum misclassification cost. In other words, the smallest tree whose cost is within one *SE* of the tree with minimum cost is selected. It is important to note that stopping rules and pruning strategies can be used simultaneously.

Example

Data. Data from the 1999 Behavioral Risk Factor Surveillance System (BRFSS) (61), which is conducted annually by U.S. states' Departments of Health in collaboration with the Centers for Disease Control and Prevention, were used. The BRFSS is a cross-sectional, random-digit dialed telephone survey of noninstitutionalized U.S. adults age 18 and older. The objective of the BRFSS is to collect uniform data on preventive health behaviors and utilization practices and on risk behaviors that are associated with chronic diseases, preventable infectious diseases, and injuries. The median state response rate was 68.4%. Our analyses were restricted to persons age 65 and older (n = 30,668) because guidelines from several professional organizations recommend annual influenza vaccination for all persons in this age group (62–64).

Measures. The example dependent variable under investigation was receipt of an annual influenza vaccination. Participants were asked if they had received an influenza vaccination within the past year. Influenza vaccination receipt in the previ-

ous year (VAC) was analyzed as a dichotomous variable (*yes–no*). Drawing from research on correlates of influenza vaccination receipt in older individuals, three example binary independent variables were selected: ever receiving a pneumonia vaccination (PN), receiving a checkup within the past year (CH), and Black race (BL), compared to all other races (65–71).

Analytic approach. We a priori determined two stopping rules: First, the minimum number of cases in each terminal node could be no smaller than 300, or approximately 1% of the total sample, and second, the minimum value of the Gini improvement measure could not be smaller than .001, which indicates modest differences between the two nodes. Because only three independent variables were investigated in this example, we did not use a stopping rule defining the maximum number of levels of the tree. The one SE pruning technique was used. The C&RT analysis was conducted using Answer Tree Version 2, a product of SPSS (72). In addition to the probabilities (expressed as percentages) generated by the C&RT analysis for each of the nodes generated, we also computed 95% confidence intervals (CIs) for each of the terminal nodes (73).

Results. The C&RT procedure generated a tree containing five terminal nodes (see Figure 2). The percentage of persons who received an influenza vaccination within the past year ranged from 31% to 86% in these five groups. The first variable selected for splitting was ever had a pneumonia vaccine (Gini = .0853). Among those who had a pneumonia vaccine, whether or not a checkup was received in the past year provided the most significant split (Gini = .0044). No further split was observed for those who had not received a checkup within the past year (Group 1; 31.2% received influenza vaccination). The group that had received a checkup within the past year was further split (Gini = .0010) according to Black race (Group 2; 38.5% received influenza vaccination) or non-Black race (Group 3; 48.1% received influenza vaccination). Those who had received a pneumonia vaccine were further differentiated by race (Gini = .0012), separating those who were Black (Group 4; 70.9% received influenza vaccination) from those who were not (Group 5; 86.3% received influenza vaccination). Neither of these groups further split according to checkup status.

For each of these five groups, 95% CIs were then calculated. Group 2 is used as an illustration. Among those in this group, 38.5% had received an influenza vaccine within the past year. The standard error was obtained ($SE = 1.63$), and the upper and lower bounds of the 95% CI were calculated by solving the standard equation:

$$95\%\ \text{CI} = 38.5 \pm (1.63)^*(1.96)$$

The resulting 95% CI was 35.3% to 41.6%.

Interpretation of results. If these results were to be used as the basis for public health program planning or decision making, one would conclude that large differences exist in influenza vaccination receipt among certain subgroups of the older population. For example, if investigators intended to target only those for whom vaccination appeared least likely, according to these results, resources could be targeted to older individuals who did not utilize other preventive health care services, as measured by never receiving a pneumonia vaccination and not having a checkup within the previous year.

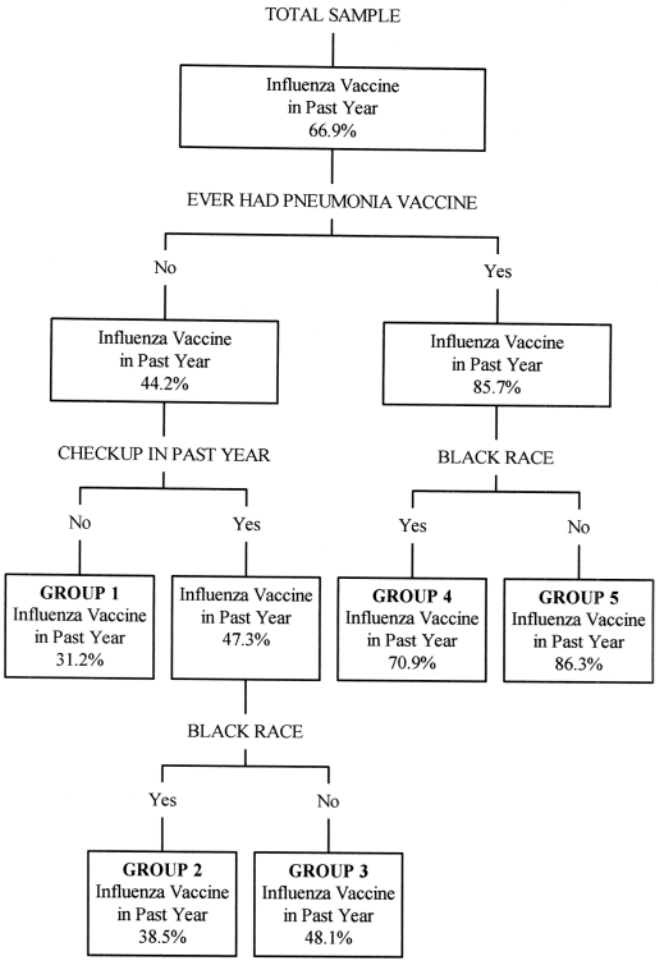

FIGURE 2 Prevalence (percentages) of influenza vaccination in the past year among classification and regression tree subgroups.

Replication Model

Analytic approach. Once the C&RT was created, a logistic regression model was computed to replicate it. This was done to allow comparisons of the prevalence estimates and associated CIs obtained in C&RT with the predicted probabilities of influenza vaccination receipt obtained from the model for each population subgroup identified as a terminal node by C&RT. Because the purpose of this model was to replicate the C&RT results, the model was designed to produce unique probability estimates only for each of the terminal nodes. It is important to note that this approach to the logistic regression model differs from traditional methods. Specifically, interactions that involve the reference category of a binary variable are typically not included in traditional regression models because the interactions have value 0 for every participant in either reference category by definition. We avoided

this limitation by creating a new term for the variable not having a pneumonia vaccine (1 – PN), which would usually be dropped from the model because its value equals zero. This allowed us to enter terms that corresponded to each of the four splits identified in the C&RT analysis into a single model.

To develop the replication model, we created variables that reflected each of the splits made by the tree. Four terms then were entered into the model:

1. The first split made by the tree was based on pneumonia vaccine status. Therefore, the first term entered into the replication model was whether a pneumonia vaccine had ever been received.
2. Among those who had not had a pneumonia vaccine, the next split was according to checkup status. Therefore an interaction term indicating not having a pneumonia vaccine and having a checkup within the past year, (1 – PN)*CH, was entered into the model.
3. Among those who had not had a pneumonia vaccine and who had received a checkup within the past year, a further split was made according to Black race. The variable (1 – PN)*CH*BL was entered into the model.
4. Among those who had a pneumonia vaccine, the next split was according to Black race. Therefore, an interaction term between having a pneumonia vaccine and Black race, PN*BL, was entered into the model to coincide with this split.

To directly compare the results from the replication logistic regression model to those obtained in the C&RT analysis, we calculated predicted probabilities (expressed as percentages) and associated 95% CIs from the resulting replication model for each of the terminal nodes identified from the C&RT analysis. SAS (Version 8) was used (73).

Results. The resulting Replication Model with estimates plugged in was as follows:

$$\log(\pi VAC/1 - \pi\, VAC) = -.85 + 2.71*(PN) + .81*[(1 - PN)*(CH)] + -.52*[(1- PN)*(CH)*(BL)] + -.99*(PN*BL)$$

where πVAC is the probability of receiving an influenza vaccination. As an illustration, using Group 2, for those who did not have a pneumonia vaccine in the past year, had a checkup within the past year, and were Black, the predicted probability of receiving an influenza vaccine in the past year is equal to .385, or 38.5%, and the CI is 35.4% to 41.7%.

Comparisons with C&RT results. As indicated in Table 2, the probabilities and associated 95% CIs of receipt of influenza vaccination within the past year were remarkably similar for the

TABLE 2
Comparison of Percentages and Associated 95% Confidence Intervals (CIs) of Population Subgroups That Received Influenza Vaccination Obtained From Three Statistical Methods

		Statistical Methods					
		Classification and Regression Tree		*Logistic Regression Replication Model*		*Logistic Regression Stepwise Selection Model*	
Group Characteristics	*n*	*%*	*95% CI*	*%*	*95% CI*	*%*	*95% CI*
Group 1: Never had pneumonia vaccine, no checkup within past year	2,638	31.2	29.5–33.0	31.1	29.4–32.8	30.5[a]	28.5–32.6[a]
Black race	109	—	—	—	—	15.5	10.6–22.6
Non-Black race	2,529	—	—	—	—	31.1	29.3–33.0
Group 2: Never had pneumonia vaccine, checkup within past year, Black race	910	38.5	35.3–41.6	38.5	35.4–41.7	38.4	35.3–41.6
Group 3: Never had pneumonia vaccine, checkup within past year, non-Black race	10,081	48.1	47.1–49.1	48.5	47.5–49.5	48.5	47.5–49.5
Group 4: Ever had pneumonia vaccine, Black race	421	71.0	67.2–74.8	70.1	67.2–74.7	70.6[b]	66.2–76.9[b]
No checkup within past year	27	—	—	—	—	52.1	40.5–63.4
Checkup within past year	394	—	—	—	—	71.9	68.0–77.8
Group 5: Ever had pneumonia vaccine, non-Black race	15,863	86.3	85.3–86.8	86.6	86.1–87.2	86.7[c]	86.0–87.4[c]
No checkup within past year	1,415	—	—	—	—	82.4	80.4–84.3
Checkup within past year	14,448	—	—	—	—	87.2	86.6–87.7

Note. A dash indicates that the classification and regression tree procedure did not further split according the respective third variable.

[a]Weighted average of probabilities of those who never had pneumonia vaccine, had no checkup within the past year, and were Black and those who never had pneumonia vaccine, had no checkup within the past year, and were non-Black. [b]Weighted average of probabilities of those who had pneumonia vaccine, were Black, and had no checkup within the past year and those who had pneumonia vaccine, were Black, and had a checkup within the past year. [c]Weighted average of probabilities of those who had pneumonia vaccine, were non-Black, and had no checkup within the past year and those who had pneumonia vaccine, were non-Black, and had a checkup within the past year.

C&RT method and the replication model. As an illustration, again consider Group 2. Results from the C&RT analysis indicated that among those who had never received a pneumonia vaccine, had a checkup within the past year, and were Black, prevalence of receipt of influenza vaccine in the past year was 38.5% (95% CI = 35.3%– 41.6%). The predicted probability (expressed as a percentage) and associated 95% CI obtained for Group 2 from the replication model was 38.5% (95% CI = 35.4%– 41.7%). Results for each of the other four groups were also very similar.

COMPARISON OF C&RT TO LOGISTIC REGRESSION: AUTOMATED STEPWISE SELECTION MODEL

Analytic Approach

We generated a computer-automated stepwise selection logistic regression model. This was done because selection procedures are commonly used techniques, and we wanted to compare results from the C&RT procedure with a method likely to be used by other investigators. Stepwise selection is similar to C&RT in that it is a computer-automated procedure in which the investigator is essentially "blinded" to the model-building process. It builds a regression model by starting with a model that includes no independent variables. The entire set of independent variables is then considered, and the one with the greatest level of statistical significance, as indicated by the smallest probability value, is added into the model. In subsequent steps, the procedure searches the remaining independent variables and enters the next most significant variable into the model. After a new independent variable has been entered, the procedure examines all previously entered independent variables and removes those that are no longer statistically significant. The procedure continues until no further statistically significant independent variables remain. In our analysis, possible independent variables were each of the three main effect terms, each of the three possible two-way interaction terms, and a three-way interaction term.

Once the final model was derived, predicted probabilities (expressed as percentages) and associated 95% CIs were also calculated to directly compare the results with those obtained in the C&RT analysis. We calculated statistics for each possible terminal node (n = 8) that could be created with the three independent variables. SAS (Version 8) was used (73).

Results

The resulting stepwise selection model, in the order in which variables were selected, was as follows:

$$\log(\pi VAC/1 - \pi VAC) = -.79 + 2.34*(PN) + .74*(CH) + -.56*(PN*BL) + -.90*(BL) + -.36*(PN*CH) + 48*(CH*BL)$$

Thus, each of the three main effect terms and each of the three two-way interaction terms were retained in this computer-generated model. The three-way interaction term was not.

Comparisons With C&RT Results

As indicated in Table 2, the C&RT and the stepwise selection model results were similar for Groups 2 and 3, which were defined by combinations of all three independent variables. For example, the predicted probability (expressed as a percentage) and associated 95% CI obtained for Group 2 from the stepwise selection model was 38.4% (95% CI = 35.3%–41.6%), compared with 38.5% (95% CI = 35.3%–41.6%) from C&RT.

Differences, however, were observed between the C&RT and the stepwise selection model for Groups 1, 4, and 5, which were not defined by combinations of all three independent variables. For Group 1, C&RT did not further split according to race, despite the fact that the probability estimates obtained in the selection model were very different for the two race groups. This was because in the C&RT analysis we established an a priori stopping rule that terminal nodes could not contain fewer than 300 people. Among those who had never received a pneumonia vaccine and had not received a checkup within the past year, only 109 were Black.

Similarly, for Group 4, among those who had received a pneumonia vaccine and were Black, only 27 people had not received a checkup within the past year. Thus, despite apparent differences in the probability of having an influenza vaccination, as indicated by the stepwise selection model, the C&RT procedure did not further split this group.

For Group 5, no further split based on checkup status was observed in the C&RT model because among those who had received a pneumonia vaccine and were non-Black, there was no significant difference, based on the Gini improvement measure, between those who received a checkup and those who did not. It should be noted, however, that the 95% CI obtained from the stepwise selection model for those in Group 5 (had pneumonia vaccine, non-Black race) who had a checkup within the past year (86.6%–87.7%) does not overlap with the CI for estimates made by the stepwise selection model for those in Group 5 (had pneumonia vaccine, non-Black race) who did not have a checkup within the past year (80.4%–84.3%), which is often considered to indicate that the groups do differ statistically.

For each of the three groups that did not split according to the third variable in C&RT (Groups 1, 4 and 5), we also calculated a weighted average of the two predicted probabilities and 95% CIs obtained from the stepwise selection model, according to the third variable. This was done to create a single statistic that could be compared with the C&RT results. For example, in Group 1, we initially calculated two predicted probabilities (expressed as percentages) and associated 95% CIs. Among those who had never received a pneumonia vaccine, had not received a checkup within the past year, and were Black (n = 109), 15.5% (95% CI = 10.6%–22.6%) were estimated to have had an influenza vaccine within the past year. Among those who had never received a pneumonia vaccine, had not received a checkup within the past year, and were non-Black (n = 2,529), 31.1% (95% CI = 29.3%–33.0%) were estimated to have had an influenza vaccine within the

past year. A weighted average, based on sample size, of these estimates resulted in a predicted probability (expressed as a percentage) of 30.5% (95% CI = 28.5%–32.6%). This estimate is similar to that obtained from C&RT.

DISCUSSION

C&RT is a decision tree method that can identify mutually exclusive and exhaustive subgroups of a population whose members share common characteristics that influence participation in a particular health-related behavior. It is important to note that there are other types of decision tree methods in addition to C&RT. These include Quick, Unbiased, Efficient Statistical Trees (QUEST), which is another binary decision tree method, and Chi-square-Automatic-Interaction-Detection (CHAID) and Automated CHAID, which use chi-square statistics to grow decision trees and can result in splits of more than two groups. C&RT, however, is considered to be the best decision tree method of the commonly used currently available methods because it is more likely to select the independent variable that is most different with respect to the target variable (74). New techniques, however, are showing considerable promise as data-mining tools. These include bagging and bootstrapping methods that aggregate classifiers and can reduce misclassification error (75–77).

There are several issues to consider when deciding whether to use C&RT analysis. It is often important to estimate the overall impact of a single independent variable on the outcome of interest. This is especially true in studies with specific hypotheses about the effect of an independent variable or group of variables on the outcome. Because C&RT analysis is intended to identify distinct population subgroups, its hierarchical nature does not allow the estimation of net effects of a single independent variable (78). Regression techniques, however, are largely used to estimate the "average" effect of an independent variable on the probability of having a dependent variable, while accounting for other factors. Thus, C&RT analysis would not be used as a substitute for proven regression techniques in this type of situation. Readers interested in an in-depth comparison of the statistical properties of C&RT analysis and regression are referred to a monograph by Michie, Spiegelhalter, and Taylor (79).

Because of the ease with which C&RT analysis can be used, it may be easy to fall into the trap of "data dredging" when using it by simply entering all possible independent variables into the tree and seeing what results. Therefore, C&RT analysis should not be used without a priori consideration of which independent variables to consider. As described by Marshall (78), it is recommended that the direction of expected relationships of each independent variable with the dependent variable also be hypothesized a priori.

It is also important to note that the example provided in this article was quite simple. In fact, classification and regression trees can become very complex and difficult to interpret. Trees can grow into multiple levels and can result in splits that are not practically important. Thus, it is important to set a priori stopping rules that place limits on the size of the tree.

Despite these drawbacks, C&RT techniques have the potential to be a useful analytic tool in public health research. C&RT analysis is particularly beneficial in situations in which the objective is not to test the hypothesis of a single independent variable or set of variables on an outcome measure but rather to describe associations in the data. The appeal of using C&RT analysis is the ease with which high-risk population subgroups can be identified. Using audience segmentation techniques to specifically target certain groups for public health intervention is becoming increasingly important, particularly with the growing focus on identifying and eliminating health disparities (80,81). However, resource limitations often limit the ability of public health personnel to do this effectively. C&RT software is easy to use for persons without strong statistical backgrounds, and the results are straightforward to interpret. Therefore, it can be much less resource intensive than other statistical methods. C&RT analysis can also play an important role in the analysis of data collected for surveillance purposes. Typically, surveillance data are analyzed only in a bivariate fashion, comparing prevalence rates of a dependent variable across strata of an independent variable. C&RT analysis allows this type of surveillance analysis to be extended by investigating more than one independent variable at a time, which allows for more effective use of the available surveillance data. In addition, C&RT analysis has the statistical advantage of being a nonparametric technique that does not invoke assumptions about the functional form of the data. Therefore, C&RT can be used without constraints on the distributions of the variables being investigated.

In conclusion, there is a growing need to be able to identify at-risk populations in public health, but limitations of commonly used statistical methods and resource constraints have made this difficult. C&RT analysis has promising potential to alleviate these barriers and become a useful tool in the analysis of public health data.

REFERENCES

(1) Breiman L, Friedman JH, Olshen RA, Stone CJ: *Classification and Regression Trees* (2nd Ed.). Pacific Grove, CA: Wadsworth, 1984.

(2) Zhang H, Singer B: *Recursive Partitioning in the Health Sciences.* New York: Springer-Verlag, 1999.

(3) Buntine W: Learning classification trees. *Statistics and Computing.* 1992, *2:*63–73.

(4) Chipman H, George EI, McCulloch RE: Bayesian CART model search (with discussion). *Journal of the American Statistical Association.* 1998, *93:*935–960.

(5) Ciampi A, Thiffault J, Nakache JP, Asselain B: Stratification by stepwise regression, correspondence analysis and recursive partitioning: A comparison of three methods of analysis for survival data with covariates. *Computational Statistics and Data Analysis.* 1986, *4:*185–204.

(6) Ciampi A, Negassa A, Lou Z: Tree-structured prediction for censored survival data and the Cox model. *Journal of Clinical Epidemiology.* 1995, *48:*675–689.

(7) Chou PA, Lookabaugh T, Gray RM: Optimal pruning with applications to tree-structured source coding and modeling. *IEEE Trans Information Theory.* 1989, *35:*299–315.

(8) Denison DT, Mallick BK, Smith AM: A Bayesian CART algorithm. *Biometrika.* 1998, *85:*363–378.
(9) Fienberg SE, Kim S-H: Calibration and refinement for classification trees. *Journal of Statistical Planning and Inference.* 1998, *70:*241–254.
(10) Gordon L, Olshen RA: Almost surely consistent nonparametic regression from recursive partitioning schemes. *Journal of Multivariate Analysis.* 1984, *15:*147–163.
(11) LeBlanc M, Crowley J: Relative risk trees for censored survival data. *Biometrics.* 1992, *48:*411–425.
(12) Liu WZ, White AP: A comparison of nearest neighbour and tree-based methods of non-parametric discriminant analysis. *Journal of Statistical Computation and Simulation.* 1995, *5:*341–350.
(13) Loh WY, Vanichestakul M: Tree-structured classification via generalized discriminant analysis. *Journal of the American Statistical Association.* 1988, *83:*715–725.
(14) Long WJ, Griffith JL, Selker HP, D'Agostino RB: A comparison of logistic regression to decision-tree induction in a medical domain. *Computers and Biomedical Research.* 1993, *26*:74–97.
(15) McConnochie KM, Roghmann KJ, Pasternack J: Developing prediction rules and evaluating observation patterns using categorical clinical markers: Two complementary procedures. *Medical Decision Making.* 1993, *13:*30–42.
(16) Oliver JJ, Hand D: Averaging over decision trees. *Journal of Classification.* 1996, *13:*281–297.
(17) Pallara A: A binary decision trees approach to classification: A review of CART and other methods with some applications in real data. *Statistica Applicata.* 1992, *4:*255–286.
(18) Segal MR, Bloch DA: A comparison of estimated proportional hazards models and regression trees. *Statistics in Medicine.* 1989, *8:*539–550.
(19) Segal MR: Tree-structured methods of logitudinal data. *Journal of the American Statistical Association.* 1992, *87:*407–418.
(20) Segal MR: Extending the elements of tree-structured regression. *Statistical Methods in Medical Research.* 1995, *4:*219–236.
(21) Selker HP, Griffith JL, Patil S, Long WJ, D'Agostino RB: A comparison of performance of mathematical predictive methods for medical diagnosis: Identifying acute cardiac ischemia among emergency department patients. *Journal of Investigative Medicine.* 1995, *43:*468–476.
(22) Shannon WD, Banks D: Combining classification trees using MLE. *Statistics in Medicine.* 1999, *18:*727–740.
(23) Zhang H, Holford T, Bracken MB: A tree-based method of analysis for prospective studies. *Statistics in Medicine.* 1996, *15:*37–49.
(24) Zhang HP: Comments on Bayesian CART model search. *Journal of the American Statistical Association.* 1998, *93:*948–950.
(25) Zhang H: Classification trees for multiple binary responses. *Journal of the American Statistical Association.* 1998, *93:*180–193.
(26) Slater MD: Choosing audience segmentation strategies and methods. In Maibach E, Parrott RL (eds), *Designing Health Messages: Approaches From Communication Theory and Public Health Practice.* Thousand Oaks, CA: Sage, 1995, 186–198.
(27) Barriga KJ, Hamman RF, Hoag S, Marshall JA, Shetterly SM: Population screening for glucose intolerant subjects using decision tree analyses. *Diabetes Research and Clinical Practice.* 1996, *34*(Suppl.):S17–S29.
(28) Curran Jr. WJ, Scott CB, Horton J, et al.: Recursive partitioning analysis of prognostic factors in three Radiation Therapy Oncology Group malignant glioma trials. *Journal of National Cancer Institute.* 1993, *85:*704–710.
(29) El-Serag HB, Graham DY, Richardson P, Inadomi JM: Prevention of complicated ulcer disease among chronic users of nonsteroidal anti-inflammatory drugs: The use of a nomogram in cost-effectiveness analysis. *Archives of Internal Medicine.* 2002, *162:*2105–2110.
(30) Falconer JA, Naughton BJ, Dunlop DD, et al.: Predicting stroke inpatient rehabilitation outcome using a classification tree approach. *Archives of Physical Medicine and Rehabilitation.* 1994, *75:*619–625.
(31) Gabriel SE, Crowson CS, O'Fallon WM: A mathematical model that improves the validity of osteoarthritis diagnoses obtained from a computerized diagnostic database. *Journal of Clinical Epidemiology.* 1996, *49:*1025–1029.
(32) Germanson T, Lanzino G, Kassell NF: CART for prediction of function after head trauma. *Journal of Neurosurgery.* 1995, *83:*941–942.
(33) Goldman L, Cook EF, Johnson PA, et al.: Prediction of the need for intensive care in patients who come to the emergency departments with acute chest pain. *New England Journal of Medicine.* 1996, *334:*1498–1504.
(34) Guccione AA, Anderson JJ, Anthony JM, Meenan RF: The correlates of health perceptions in rheumatoid arthritis. *Journal of Rheumatology.* 1995, *22:*432–439.
(35) Haukoos JS, Witt MD, Zeumer CM, et al.: Emergency department triage of patients infected with HIV. *Academic Emergency Medicine.* 2002, *9:*880–888.
(36) Hess KR, Abbruzzese MC, Lenzi R, Raber MN, Abbruzzese JL: Classification and regression tree analysis of 1000 consecutive patients with unknown primary carcinoma. *Clinical Cancer Research.* 1999, *5:*3403–3410.
(37) Kaufman KE, Bailit JL, Grobman W: Elective induction: An analysis of economic and health consequences. *American Journal of Obstetrics and Gynecology.* 2002, *187:*858–863.
(38) Pilote L, Miller DP, Califf RM, et al.: Determinants of the use of coronary angiography and revascularization after thrombolysis for acute myocardial infarction. *New England Journal of Medicine.* 1996, *335:*1198–1205.
(39) Podgorelec V, Kokol P, Stiglic B, Rozman I: Decision trees: An overview and their use in medicine. *Journal of Medical Systems.* 2002, *26:*445–463.
(40) Rainer TH, Lam PK, Wong EM, Cocks RA: Derivation of a prediction rule for post-traumatic acute lung injury. *Resuscitation.* 1999, *42:*187–196.
(41) Roehrborn CG, Malice M, Cook TJ, Girman CJ: Clinical predictors of spontaneous acute urinary retention in men with LUTS and clinical BPH: A comprehensive analysis of the pooled placebo groups of several large clinical trials. *Urology.* 2001, *58:*210–216.
(42) Rudolfer SM, Paliouras G, Peers IS: A comparison of logistic regression to decision tree induction in the diagnosis of carpal tunnel syndrome. *Computers and Biomedical Research.* 1999, *32:*391–414.
(43) Temkin NR, Holubkov R, Machamer JE, Winn HR, Dikmen SS: Classification and regression trees (CART) for prediction of function at 1 year following head trauma. *Journal of Neurosurgery.* 1995, *82:*764–771.
(44) Travis SP, Farrant JM, Ricketts C, et al.: Predicting outcome in severe ulcerative colitis. *Gut.* 1996, *38:*905–910.
(45) Wietlisbach V, Vader JP, Porchet F, Costanza MC, Burnand B: Statistical approaches in the development of clinical practice guidelines from expert panels: The case of laminectomy in sciatica patients. *Medical Care.* 1999, *37:*785–797.

(46) Bachur RG, Harper MB: Predictive model for serious bacterial infections among infants younger than 3 months of age. *Pediatrics.* 2001, *108:*311–316.

(47) Camp NJ, Slattery ML: Classification tree analysis: A statistical tool to investigate risk factor interactions with an example for colon cancer (United States). *Cancer Causes Control.* 2002, *13:*813–823.

(48) Carmelli D, Zhang H, Swan GE: Obesity and 33-year follow-up for coronary heart disease and cancer mortality. *Epidemiology.* 1997, *8:*378–383.

(49) Choi SC, Muizelaar JP, Barnes TY, et al.: Prediction tree for severely head-injured patients. *Journal of Neurosurgery.* 1991, *75:*251–255.

(50) El-Solh AA, Sikka P, Ramadan F: Outcome of older patients with severe pneumonia predicted by recursive partitioning. *Journal of the American Geriatrics Society.* 2001, *49:*1614–1621.

(51) Kuchibhatla M, Fillenbaum GG: Assessing risk factors for mortality in elderly White and African American people: Implications of alternative analyses. *Gerontologist.* 2002, *42:*826–834.

(52) Mehta RH, Eagle KA, Coombs LP, et al.: Influence of age on outcomes in patients undergoing mitral valve replacement. *Annals of Thoracic Surgery.* 2002, *74:*1459–1467.

(53) Nelson LM, Bloch DA, Longstreth Jr. WT, Shi H: Recursive partitioning for the identification of disease risk subgroups: A case-control study of subarachnoid hemorrhage. *Journal of Clinical Epidemiology.* 1998, *51:*199–209.

(54) McGrath JS, Ponich TP, Gregor JC: Screening for colorectal cancer: The cost to find an advanced adenoma. *American Journal of Gastroenterology.* 2002, *97:*2902–2907.

(55) Smith KJ, Roberts MS: Cost-effectiveness of newer treatment strategies for influenza. *American Journal of Medicine.* 2002, *113:*300–307.

(56) LaValley M, McAlindon TE, Evans S, Chaisson CE, Felson DT: Problems in the development and validation of questionnaire-based screening instruments for ascertaining cases with symptomatic knee osteoarthritis: The Framingham Study. *Arthritis and Rheumatism.* 2001, *44:*1105–1113.

(57) Gregory KD, Korst LM, Platt LD: Variation in elective primary cesarean delivery by patient and hospital factors. *American Journal of Obstetrics and Gynecology.* 2001, *184:*1521–1532; discussion 1532–1534.

(58) Hosmer DW, Lemeshow S: *Applied Logistic Regression* (2nd Ed.). NewYork: Wiley, 2000.

(59) Forthofer MS, Bryant CA: Using audience-segmentation techniques to tailor health behavior change strategies. *American Journal of Health Behavior.* 2000, *24:*36–43.

(60) LeBlanc M, Tibshirani R: Monotone shrinkage of trees. *Journal of Computational and Graphical Statistics.* 1998, *7:*417–433.

(61) Centers for Disease Control and Prevention: *Behavioral Risk Factory Surveillance System User's Guide.* Atlanta: U.S. Department of Health and Human Services, 1998.

(62) U.S. Preventive Services Task Force: *Guide to Clinical Preventive Health Care: Report of the U.S. Preventive Services Task Force.* Baltimore, MD: Williams & Wilkins, 1996.

(63) Centers for Disease Control and Prevention: Prevention and control of influenza: Recommendations of the Advisory Committee on Immunization Practices (ACIP). *Morbidity and Morality Weekly Report.* 2001, *50*(RR04):1–46.

(64) American College of Physicians Task Force on Adult Immunization, Infectious Diseases Society of America: *Guide for Adult Immunization* (3rd Ed.). Philadelphia: American College of Physicians, 1994.

(65) Centers for Disease Control and Prevention: From the Centers for Disease Control and Prevention: Influenza and pneumococcal vaccination levels among persons aged >/=65 years, United States, 1999. *Journal of the American Medical Association,* 2001, *286:*413–414.

(66) Centers for Disease Control and Prevention: Leads from the Morbidity and Mortality Weekly Report, Atlanta, GA: Race-specific differences in influenza vaccination levels among Medicare beneficiaries—United States, 1993. *Journal of the American Medical Association.* 1995, *273:*449–451.

(67) Schneider EC, Cleary PD, Zaslavsky AM, Epstein AM: Racial disparity in influenza vaccination: Does managed care narrow the gap between African Americans and Whites? *Journal of the American Medical Association.* 2001, *286:*1455–1460.

(68) Petersen RL, Saag K, Wallace RB, Doebbeling BN: Influenza and pneumococcal vaccine receipt in older persons with chronic disease: A population-based study. *Medical Care.* 1999, *37:*502–509.

(69) Fiebach NH, Viscoli CM: Patient acceptance of influenza vaccination. *American Journal of Medicine.* 1991, *91:*393–400.

(70) Marin MG, Johanson Jr. WG, Salas-Lopez D: Influenza vaccination among minority populations in the United States. *Preventive Medicine.* 2002, *34:*235–241.

(71) Fiscella K, Franks P, Doescher MP, Saver BG: Disparities in health care by race, ethnicity, and language among the insured: Findings from a national sample. *Medical Care.* 2002, *40:*52–59.

(72) SPSS: *AnswerTree 2.0 User's Guide.* Chicago: SPSS, Inc., 1998.

(73) SAS Institute: *SAS/STAT User's Guide, Version 8, Volumes 1, 2 and 3.* Cary, NC: SAS Institute, 2000.

(74) Steinberg D, Colla P: *CART: Tree-structured non-parametric data analysis.* San Diego, CA: Salford Systems, 1995.

(75) Breiman L: Bagging predictors. *Machine Learning*. 1996, *24:*123–140.

(76) Breiman L: Arcing classifiers. *Annals of Statistics.* 1998, *26:*801–824.

(77) Hothorn T, Lausen B: Bagging tree classifiers for laser scanning images: A data- and simulation-based strategy. *Artificial Intelligence in Medicine.* 2003, *27*:65–79.

(78) Marshall RJ: The use of classification and regression trees in clinical epidemiology. *Journal of Clinical Epidemiology.* 2001, *54:*603–609.

(79) Michie D, Spiegelhalter DJ, Taylor CC: *Machine Learning, Neural and Statistical Classification.* New York: Ellis Horwood, 1994.

(80) U.S. Department of Health and Human Services, Office of Disease Prevention and Health Promotion: *Healthy People 2010: Understanding and Improving Health and Objectives for Improving Health* (2nd Ed.). Washington, DC: U.S. Department of Health and Human Services, 2000.

(81) Smedley BD, Stith AY, Nelson AR: The Institute of Medicine Report: Unequal treatment: Confronting racial and ethnic disparaties in health care. Washington, DC: National Academy Press, 2002.