Assessing Risk Factors for Mortality in Elderly White and African American People: Implications of Alternative Analyses

Maragatha Kuchibhatla, PhD, Gerda G. Fillenbaum, PhD
Assessing Risk Factors for Mortality in Elderly White and African American People: Implications of Alternative Analyses

Maragatha Kuchibhatla, PhD\textsuperscript{1,2} and Gerda G. Fillenbaum, PhD\textsuperscript{2}

\textbf{Purpose:} The aim of this study was to ascertain whether the determinants of death differ as a function of type of analysis in a representative sample of older African American and White people with comparable mortality rates. 

\textbf{Design and Methods:} Participants included all African American (\(n = 2,261\)) and White (\(n = 1,875\)) people at the Duke site of the Established Populations for Epidemiological Studies of the Elderly. Baseline information used to predict mortality 12 years later included demographic, health, and functional characteristics. Mortality (55\% for African American people and 54\% for White people) was determined through the National Death Index. Cox proportional hazards models, logistic regression, and tree-based classification analysis were used (separately for African American and White people) to identify risk factors for mortality.

\textbf{Results:} Risk factors for mortality were comparable, but the constellation of characteristics indicating higher risk for death differed between African American and White people.

\textbf{Implications:} Proportional hazards and logistic regression identified risk factors in general; tree-based classification models identified the characteristics of groups at risk. The analysis used may influence the type and manner of intervention.

\textbf{Key Words:} Proportional hazards models, Logistic regression, Classification tree analysis, Racial differences

Nationally, even among those who reach age 65, the remaining life span of African American individuals is less than that of White individuals (2 years less for African American men than for White men, 1.7 years less for African American women than for White women [U.S. Bureau of the Census, 2000; Table 116]). Only among those 85 years of age and older is the situation reversed, the death rate being less, and the remaining years of African American people being marginally greater than for White people (Horner, 2001; U.S. Bureau of the Census, 2000; Table 118).

This situation, however, does not appear to hold for the African American and White participants at the Duke site of the Established Populations for Epidemiologic Studies of the Elderly (EPESE). In this major 10-year longitudinal study of a sample of more than 4,000 people 65 years of age and older (who are representative of more than 28,000 community residents the same age in a five-county area), the proportions dying within 12 years of entry into the study were highly similar—55\% for African American participants and 54\% for White participants. The age composition of the two groups at entry into the study was similar. A summary measure of medical status indicated that disease burden did not differ significantly between the two groups, but there were differences in the relative presence of the chronic conditions examined. A question therefore arises as to the risk factors for death in each group and whether these risk factors were the same for the African American and White participants.

Alternative statistical procedures are available to address this question. Here, we compare three such methods: Cox proportional hazards (PHs) modeling, logistic regression (LR), and tree-based classification analysis using recursive partitioning methods. These procedures depend on different statistical assumptions. Although the information they produce is compatible, the presentation of that information leads to different ways of understanding the findings and may have different practical implications, as results from the present study illustrate.

\textbf{Methods}

\textbf{Sample}

Data for the present study came from the participants of the Duke EPESE project (Cornoni-Huntley et
Independent Variables

Data Gathering

The participants from other races were deleted were either African American (2,261) or White (1,875). Sons contacted, 4,162 (80%) participated. All but 26 (they constituted 54% of the sample, compared with race, with African American participants oversampled to improve statistical precision for this group (they constituted 54% of the sample, compared with an actual representation of 35%). Of the 5,223 persons contacted, 4,162 (80%) participated. All but 26 were either African American (2,261) or White (1,875). The participants from the other races were deleted from the analysis. Race/ethnicity was self-declared. The participants were fairly evenly divided between residence in urban and rural areas. All levels of socio-economic status were represented in both racial/ethnic groups. The age and gender composition of the African American and White groups was similar. This study was approved by the Institutional Review Board of Duke University Medical Center.

Data Gathering

At baseline (1986/1987), data were gathered by means of structured questionnaires administered in person in the homes of participants by trained interviewers. The information obtained included demographic characteristics, health conditions, and functional status.

Dependent Variable

Survival status and date of death were determined by search of the National Death Index. Information on report of death should have been equally available for all decedents. The final National Death Index search was carried out in March 2000, capturing deaths through December 1998. The present analyses focus on determinants of death occurring within 12 years of baseline. During this time interval, 1,248 African American (55%) and 1,017 White (54%) participants died. Time-to-death was defined as the time between the date-of-death and baseline date-of-interview, with censoring at Year 12 for those alive at that time.

Independent Variables

The demographic variables selected for the present analyses included gender, age, and education (Campbell, Diep, Reinken, & McCosh, 1985; Drever, Whitehead & Roden, 1996; Fried et al., 1998; Howard, Anderson, Russell, Howard, & Burke, 2000; Kitagawa & Hauser, 1973). Physical health conditions included the self-reported presence of five chronic health conditions known to be risk factors for death—heart attack, stroke, diabetes, hypertension, and cancer (excluding skin cancer; Campbell et al., 1985; Fillenbaum, Pieper, Cohen, Corno-Huntley, & Guralnik, 2000; Fried et al., 1998). The functional status scales included basic activities of daily living (ADLs; bathing, dressing, eating, transferring, using the toilet; Katz & Akpom, 1976), instrumental ADLs (IADLs; using the telephone, traveling, shopping, preparing meals, doing housework, taking medicine, handling everyday finances; Fillenbaum, 1985), and mobility (heavy housework, climbing stairs, walking half-a-mile; Rosow & Breslau, 1966). Each of these three measures was summed separately to indicate the number of activities that participants reported being unable to carry out independently. Each scale was included in the analysis as a continuous measure. These variables were chosen based on the correlations among a set of potential risk factors and based on the empirical evidence found in the literature.

Analytic Approach

Three analytic approaches were used: Cox PH modeling (Lee, 1980), LR (Hosmer & Lemeshow, 1989), and tree-based classification analysis (Breiman, Friedman, Olsen, & Stone, 1998) using recursive partitioning methods. The PH model, also called survival model, determines the risk factors for time to event (here survival status, dead or alive) and not just the event. Hence, it is more accurate than LR. The Cox PH method is one of the most popular methods for analyzing survival data, because the method does not require knowledge of the probability distribution of the survival times and can incorporate time-varying covariates. The LR approach indicates which, among the variables included in analysis, is a significant predictor of the dependent variable and gives the relative net predictive strength of the variables in the model. Tree-based classification analysis is an exploratory technique to understand the pattern of course of outcome, which here is mortality. In the tree-based classification analysis, a set of classes into which the subjects fall, such as dead or alive as in the analysis here, is specified. The technique then aims to find splits of the predictor variables that result in the accurate assignment of subjects into the class they belong. The tree-based classification approach using recursive partitioning methods (Breiman et al., 1998) identifies successively the best variable at its best split that produces the classification into outcome groups (here alive/dead) that is more accurate than any other split and the best combination of variables by their cutpoints. That is, the algorithm for this method considers all variables and all possible splits at each stage and chooses the best split among all variables based on misclassification error. [The error representing the percentage of participants incorrectly classified into the outcome group (here dead/alive) is the misclassification error (e.g., in a tree structure, a participant could be classified to belong to the outcome group of those “alive” when actually the participant is dead)]. Its terminals are groups with specific characteristics, for which the proportion meeting the endpoint in question is known.

Tree-based models differ from regression models in that they do not require the linearity and additive assumptions inherent in the regression models, and can detect the interactions not routinely detected by other methods. Interactions between variables need not be
prespecified as in the PH and logistic models, but are revealed through tree structure. Interactions of higher order are harder to interpret in both PH and LR models. The tree-based models have a built-in method for handling missing data (Breiman et al., 1998), whereas the PH and logistic models are based on observations with complete information. PH and logistic models can be and have also been used with tree-based models to supplement the results of each other.

For the current dichotomous outcome (dead/alive), the following three steps were used to construct the final classification tree: (1) recursive partitioning to choose the predictor variables, (2) pruning the tree obtained above to snip off the least important splits, and (3) selecting an optimum-sized tree from the above pruned tree using a technique called cross-validation. Cross-validation provides estimates (Breiman et al., 1998, pages 10–12, & 306–309) of the misclassification rates of the subtrees, and identifies the most complex subtree that minimizes the cross-validated misclassification rate of the outcome (dichotomized values of death). To apply this technique, the data set is randomly partitioned into a number of subsamples (i.e., “v” subsamples), with each subsample containing $v\%$ randomly selected cases. The first sample, consisting of $v\%$ of the data, is set aside and a classification tree is generated using the remaining data. The first $v\%$ sample is then used to validate the new tree and the misclassification (error) rate is obtained. The above method is repeated on all the remaining samples and a “$v\%$”-fold cross-validated and unbiased estimate of the overall misclassification rate is obtained. For the analysis here, the data set was randomly partitioned into 10 subsamples, each containing 10$\%$ of the sample. The 10-fold cross-validated estimates were used to select the trees that were optimal with respect to minimizing the misclassification (error) rates. In the classification analysis that follows, the final trees have six or seven terminal nodes (as shown in the tree diagrams of Figures 1 through 4).

PH models and LR models for African American and White participants separately were executed in SAS software (SAS Institute, 1994) using baseline demographic characteristics: gender, age, education (in years), and the presence or absence of selected health conditions (heart attack, cancer, stroke, hypertension, and diabetes) as predictors of mortality. Another set of models was obtained using demographic variables and functional status variables (basic ADLs, IADLs, and mobility) measured in terms of problems reported. For all models, tests indicated nonsignificant collinearity among variables.

S-plus software (Statistical Sciences, 1995; Venables & Ripley, 1999) was used to construct the classification trees. One set of classification trees was constructed for African American and White participants separately using baseline demographic characteristics and the presence or absence of the selected health
Figure 2. Classification tree for risk factors for mortality among African American participants: demographic and specific health conditions. The first number in each cell gives the number dead, and the second number is the number alive. Bolded third number on all the terminal nodes is the proportion dead at that node and indicates mortality groups.

Figure 3. Classification tree for risk factors for mortality among White participants: demographic characteristics and activities of daily living. The first number in each cell gives the number dead, and the second number is the number alive. Bolded third number on all the terminal nodes is the proportion dead at that node and indicates mortality groups. IADL = instrumental activities of daily living.
conditions, as predictors of mortality. Another set of classification trees was constructed using demographic variables and the functional status variables. The purpose of the classification tree is to identify the specific set of variables that identify mortality groups—separately for older White groups and for older African American groups—and to determine whether there are racial differences in these groupings.

**Results**

Table 1 gives the baseline characteristics of African American and White participants. The mean ages of the two racial groups were similar, as was their gender distribution; but, the mean education level of White participants (9.1 years) was higher than that of African American groups (7.3 years). A larger percentage of African American participants were hypertensive, had had a stroke, and were diabetic than were White participants; White participants were more likely to report heart attacks and cancer. Also, African American participants had more problems with basic ADLs, IADLs, and mobility than White participants. In both the White and African American participants there was less than 1% missing on variables relating to health conditions and less than 2.5% missing on physical function variables.

The PH model of demographic and specific health conditions (Table 2) indicated that older age, male gender, less education, presence of heart attack, hypertension, stroke, cancer, and diabetes were all risk factors for mortality for White participants. With the exception of hypertension (which was not identified as a risk factor), findings were comparable for African American participants. Increased age and heart attack, however, appeared to be stronger risk factors for White participants than for African American participants. Results based on LR were comparable with those from the PH model, except that no significant difference between White and African American participants in risk from heart attack was found. In the models with demographic and functional status variables (Table 3), for both White and African American participants, the same demographic characteristics were risk factors for mortality. Increased age,

---

**Table 1. Baseline Characteristics by Race**

<table>
<thead>
<tr>
<th>Variable</th>
<th>African American</th>
<th>White</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age (Mean, SD)</strong></td>
<td>73.6 (6.9)</td>
<td>73.5 (6.6)</td>
</tr>
<tr>
<td><strong>Gender (% Male)</strong></td>
<td>34.9</td>
<td>35.0</td>
</tr>
<tr>
<td><strong>Education (Mean, SD), yr</strong></td>
<td>7.3 (4.0)</td>
<td>9.1 (3.7)</td>
</tr>
<tr>
<td><strong>Heart attack (%)</strong></td>
<td>13.1</td>
<td>17.3</td>
</tr>
<tr>
<td><strong>Hypertension (%)</strong></td>
<td>63.2</td>
<td>50.4</td>
</tr>
<tr>
<td><strong>Stroke (%)</strong></td>
<td>9.6</td>
<td>8.1</td>
</tr>
<tr>
<td><strong>Cancer (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(excluding skin cancer)</td>
<td>7.2</td>
<td>11.5</td>
</tr>
<tr>
<td><strong>Diabetes (%)</strong></td>
<td>24.2</td>
<td>15.3</td>
</tr>
<tr>
<td><strong>Katz ADL (Mean, SD)</strong></td>
<td>0.29 (.89)</td>
<td>0.24 (.83)</td>
</tr>
<tr>
<td><strong>Instrumental ADL (Mean, SD)</strong></td>
<td>0.82 (1.47)</td>
<td>0.65 (1.35)</td>
</tr>
<tr>
<td><strong>Rosow-Breslau (Mean, SD)</strong></td>
<td>1.01 (1.16)</td>
<td>0.88 (1.14)</td>
</tr>
</tbody>
</table>

*Note: ADL = activities of daily living.*
however, was a greater risk factor for White participants, as was male gender in the LR model only. An increasing number of IADL problems and mobility problems were risk factors for mortality. Basic ADL problems were not a risk factor for mortality for White participants, and education just misses being protective for both groups.

The classification tree for White participants based on demographic and health conditions is shown in the tree diagram in Figure 1. The first number in each of the nodes (circle or rectangle) is the number dead at that node, and the second number is the number alive at that node. Here, the overall proportion dead was 0.54, and a terminal node with 54% or more dead was considered a mortality group. The optimal split in the sample of White participants was from the variable age (split at 77.5 years), indicating that higher age was an important predictor of mortality among White participants. The group on the right (age >77.5 years) contained a larger percentage of those who die (404/488 = 81.7%) than did the one on the left (44.1%; age ≤77.5 years). The group on the right was further classified into those with age greater than 82.5 years and less than 82.5 years. Those with age less than 82.5 years form a terminal node, with a probability of death = .76. In the final classification tree, the mortality groups were terminal subsets 1, 3, 5, and 6. These four subgroups among White participants can be described as follows: (1) In the group on the left, the terminal subset 1 consisted of persons less than 77.5 years of age who had had a heart attack (probability of death = .66); (2) terminal subset 3 consisted of men between ages 69.5 and 77.5 years who had had no heart attack (probability of death = .63); (3) terminal subset 5 consisted of persons 77.5 through 82.5 years of age (probability of death = .76); and (4) terminal subset 6, the rightmost subgroup, was the oldest, with age > 82.5 years (probability of death = .93).

The classification tree based on demographic and health conditions, for African American participants, is given in the tree diagram in Figure 2. The optimal split for African American participants was also age, but at a slightly older age (79.5 years) than the 77.5 years for White participants. The terminal nodes were 2, 4, and 7. The classification tree with significant subgroups for mortality were: (1) terminal node 2, which consisted of men between 72.5 and 79.5 years old who had less than 9.5 years of education (probability of death = .74); (2) terminal node 4, which consisted of women less than 79.5 years of age, who were diabetic (probability of death = .58); and (3) terminal node 7, which consisted of persons older than 79.5 years (probability of death = .82).

### Table 2. Demographic Characteristics and Specific Health Conditions: Risks of Mortality Based on Cox Proportional Hazards Models (Risk Ratios) and Logistic Regression Models (Odds Ratios)

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>White Risk Ratio (95% CI)</th>
<th>White Odds Ratio (95% CI)</th>
<th>African American Risk Ratio (95% CI)</th>
<th>African American Odds Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age 1.10 (1.09–1.12)</td>
<td>1.18 (1.15–1.20)</td>
<td>1.07 (1.06–1.08)</td>
<td>1.12 (1.10–1.13)</td>
<td></td>
</tr>
<tr>
<td>Gender (Ref: female)</td>
<td>1.87 (1.63–2.13)</td>
<td>2.46 (1.94–3.11)</td>
<td>1.69 (1.50–1.90)</td>
<td>2.19 (1.79–2.68)</td>
</tr>
<tr>
<td>Education 0.97 (0.95–0.98)</td>
<td>0.95 (0.93–0.98)</td>
<td>0.97 (0.95–0.98)</td>
<td>1.20 (1.03–1.41)</td>
<td>1.37 (1.04–1.80)</td>
</tr>
<tr>
<td>Heart attack 1.64 (1.41–1.90)</td>
<td>2.28 (1.69–3.08)</td>
<td>1.01 (0.90–1.14)</td>
<td>1.74 (1.46–2.07)</td>
<td>2.17 (1.54–3.04)</td>
</tr>
<tr>
<td>Hypertension 1.24 (1.09–1.41)</td>
<td>1.35 (1.09–1.67)</td>
<td>1.74 (1.99–5.25)</td>
<td>1.30 (1.06–1.59)</td>
<td>1.67 (1.16–2.41)</td>
</tr>
<tr>
<td>Stroke 1.92 (1.59–2.33)</td>
<td>3.23 (1.99–5.25)</td>
<td>1.50 (1.32–1.70)</td>
<td>1.30 (1.06–1.59)</td>
<td>1.67 (1.16–2.41)</td>
</tr>
<tr>
<td>Cancer 1.39 (1.17–1.66)</td>
<td>1.94 (1.38–2.73)</td>
<td>1.50 (1.32–1.70)</td>
<td>1.50 (1.32–1.70)</td>
<td>1.85 (1.48–2.30)</td>
</tr>
<tr>
<td>Diabetes 1.43 (1.21–1.68)</td>
<td>1.82 (1.35–2.46)</td>
<td>1.50 (1.32–1.70)</td>
<td>1.50 (1.32–1.70)</td>
<td>1.85 (1.48–2.30)</td>
</tr>
</tbody>
</table>

Notes: CI = confidence interval.

### Table 3. Demographic Characteristics and Functional Status: Risks of Mortality Based on Cox Proportional Hazards Models (Risk Ratios) and Logistic Regression Models (Odds Ratios)

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>White Risk Ratio (95% CI)</th>
<th>White Odds Ratio (95% CI)</th>
<th>African American Risk Ratio (95% CI)</th>
<th>African American Odds Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age 1.08 (1.07–1.09)</td>
<td>1.15 (1.13–1.17)</td>
<td>1.04 (1.03–1.05)</td>
<td>1.05 (1.04–1.06)</td>
<td></td>
</tr>
<tr>
<td>Gender (Ref: female)</td>
<td>2.26 (1.96–2.59)</td>
<td>3.04 (2.40–3.85)</td>
<td>1.77 (1.56–2.00)</td>
<td>1.84 (1.62–2.08)</td>
</tr>
<tr>
<td>Education 0.98 (0.96–1.00)</td>
<td>0.97 (0.94–1.00)</td>
<td>0.98 (0.97–1.00)</td>
<td>1.07 (1.00–1.15)</td>
<td>1.08 (1.01–1.15)</td>
</tr>
<tr>
<td>Basic ADL problems (0–5) 1.08 (0.99–1.17)</td>
<td>1.07 (0.85–1.34)</td>
<td>1.11 (1.05–1.16)</td>
<td>1.11 (1.05–1.16)</td>
<td>1.11 (1.05–1.16)</td>
</tr>
<tr>
<td>IADL problems (0–7) 1.06 (1.00–1.13)</td>
<td>1.19 (1.03–1.38)</td>
<td>1.27 (1.19–1.36)</td>
<td>1.27 (1.19–1.36)</td>
<td>1.27 (1.19–1.36)</td>
</tr>
<tr>
<td>Mobility problems (0–3) 1.34 (1.24–1.44)</td>
<td>1.50 (1.31–1.71)</td>
<td>1.27 (1.19–1.36)</td>
<td>1.27 (1.19–1.36)</td>
<td>1.27 (1.19–1.36)</td>
</tr>
</tbody>
</table>

Notes: CI = confidence interval; ADL = activities of daily living; IADL = instrumental activities of daily living.
The classification tree for White participants based on demographic characteristics and functional status is given in the tree diagram in Figure 3. The significant terminal nodes were 3, 5, and 6. The mortality groups based on these subsets were: (1) terminal node 3—men less than 77.5 years old with one or more mobility problems as measured by the Rosow-Breslau scale; (2) terminal node 5—persons greater than 77.5 years of age with at most one IADL problem (probability of death = 0.75); and (3) terminal node 6—persons greater than 77.5 years with at least two IADL problems (probability of death = 0.96).

The classification tree for African American participants based on demographic characteristics and activities of daily living is given in the tree diagram in Figure 4. Unlike the previous analyses, where age was the primary characteristic for identifying those at greatest risk for death, here the primary classification was by the presence of mobility problem. The mortality groups consisted of terminal nodes 3, 4, and 6: (1) terminal node 3 consisted of persons more than 79.5 years old with no mobility problems as measured by the Rosow-Breslau scale (probability of death = 0.74); (2) terminal node 4 consisted of men with one or more mobility problems and with at most one IADL problem (probability of death = 0.77); and (3) terminal node 6 consisted of persons with one or more mobility problems and two or more IADL problems (probability of death = 0.81). Thus, whereas increased age continued to be a risk factor for mortality in both White and African American participants, problems related to functional status alone were risk factors for African American participants.

Discussion

By comparing three different analytic approaches (PH modeling, LR, and tree-based classification analysis using recursive tree partitioning methods), we wish to ascertain whether the primary determinants of mortality are the same for African American participants as for White participants in a sample in which the rates of death over a period of 12 years differ minimally between the two groups. In all analyses, we include age, gender, and education, which are established risk factors for mortality (Campbell et al., 1985; Drever et al., 1996; Fried et al., 1998; Howard et al., 2000; Kitagawa & Hauser, 1973). In addition, in one set of analyses, we include health conditions that are known risk factors for death (U.S. Department of Health and Human Services, 2000, Table 33), whereas in a second set of analyses we include information on basic ADL and IADL functioning and mobility. The former analyses allow us to examine the effects of specific health conditions, the latter, by focusing on functional status that crosscuts health conditions and allows us to look at the impact of decline in independent functioning.

The PH and LR models provide similar information whereas results from the tree-based classification presented results rather differently from that of the other two approaches, but which is to some extent confirmatory. It is important to be aware of the differences because they influence our understanding of the relevance of selected risk factors, and they may influence policy decisions regarding intervention. The PH analysis takes into account time to death while examining the extent to which the included variables predict death. LR analysis focuses on whether or not death occurred, indicates the extent to which each of the included variables are net of those entered into analysis, and is a statistically significant determinant of the outcome. Thus, LR asks to what extent the presence of say, heart attack is associated with an increase in risk of death consistently across all subgroups or other variables, and estimates a summary value of that. Based on the PH and LR analyses, findings for African American and White participants are similar. Age, gender, and education have a comparable impact in both groups, although the effect of age is stronger in White than in African American participants. With the exception of hypertension (a risk factor for White participants), the same health conditions are risk factors for mortality for both groups. Although heart attack, stroke, and cancer have higher risk odds ratios for White participants, the confidence intervals for stroke and cancer overlap with those for African American participants. Findings were comparable in analyses that included information on functional status, except that here the point estimates for the two groups were even closer. These findings suggest that, in the present sample, African American and White participants are susceptible to comparable risk factors for mortality, because the PH and LR analyses do not differ greatly in their findings.

The tree-based classification analyses using recursive partitioning methods indicate a slightly different story. This classification method is used to understand the data structure by identifying the groups differing in mortality and the role of the predictor variables in identifying them. Now comparing African American and White participants we find that in analyses that included health factors in addition to demographic characteristics, older age, alone, is a notable risk factor for both groups. As might be anticipated from both the PH and LR models, the age split that examines identification of mortality is lower for White participants (where relative risk and odds ratios for White participants are greater) than for African American participants. After age, heart attack is the most important variable among White participants, whereas gender is the important variable among African American participants. The key health variables identified from this analysis are heart attack among White participants and diabetes for African American participants. PH analysis (but not LR) indicated that heart attack was a significantly stronger risk factor for White participants than for African American participants. Identification of heart attack as a risk factor distinguishing White participants from African American participants might have been anticipated. One could not, however, have anticipated that it would be the first health condition selected; stroke, for instance, was found to offer a higher risk. Neither could diabetes in
African American participants have been anticipated as the first division point, because neither PH nor logistic analyses found diabetes to offer the greatest risk. It is possible that such associations could have been identified using two-way or higher interactions terms. Such terms, however, are notoriously difficult to interpret. Tree-based classification essentially yields the relevant findings in a readily understandable manner.

Looking at functional status, the start point for White participants is again age, cut at 77.5 years as in the health conditions analysis; but, for African American participants, the start point is mobility (mobility becomes important at the second stage for younger White participants). Among White participants, older age is a major determinant of mortality, with the risk of mortality increased in those who have difficulty performing multiple IADL tasks or, among the younger men, those with mobility problems. Findings for African American participants involve combinations of problems with mobility, IADLs, and mobility among men. Although the lack of involvement of basic ADL tasks is to be expected given the results of the PH and LR analyses, the order of variables used for classification, and the differences in combinations of variables between White and African American participants could not be readily anticipated from PH and LR analyses. The tree-based models essentially identify the importance of the variables in the different groups of subjects who are at risk for mortality, whereas the variables may differ between the groups.

PH models and LR allow us to identify multiple characteristics that are risk factors. Depending on the modifiability of these characteristics, interventions can then be planned. Classification tree analysis identifies those combinations of characteristics that identify groups at particular risk. This permits very specific targeting of resources and very specific determination of interventions. Each of these approaches has its advantages and disadvantages. Here, we find that, whereas the proportion dying and the risk factors for mortality are comparable in African American and White groups, the characteristics of the high-risk groups may differ. Classification tree encourages potential targeting of groups, whereas PH and LR are better suited for targeting specific conditions. Essentially, the commonly used multivariable methods, PH and LR models, help determine the multiple risk factors for an outcome, whereas tree-based models can identify the joint effects of risk factors on the outcomes. Tree-based models have been used in many clinical applications: Goldman and colleagues (1981) have used these models to classify patients with chest pain into homogeneous groups that will help physicians plan for appropriate medical care in emergencies; Levy and colleagues (1985) have used them to predict the outcome of coma from cerebral hypoxia-ischemia using 13 predictors. They have been used to help improve the medical diagnosis of systemic lupus erythematosus (Edworthy, Zatarian, McShane, & Bloch, 1982), osteoarthritis (Gabriel, Crowson, & O’Fallon, 1996), and tension and migraine headache (Diehr et al., 1981); Zhang and Bracken (1995) have used the tree-based models to predict the risk factors for preterm delivery.

The form of classification or prediction rule in tree-based models is very different from logistic and linear regression analyses in which linear combinations are the primary mode of expressing relationships between variables (Clark & Pregibon, 1992). Consequently, the results of these analyses are not expected to be the same. Certain consistencies are, however, expected. One may have a variable that is not selected by logistic regression because of its apparent insignificance. By examining this variable in the presence of other variables or in a different functional form (logistic regression assumes a logit–linear relationship), the relation might become more significant. Tree-based classification methods, and LR and PH models, can be and have been used in combination. For example, in one approach, the linear equation derived from LR can be an additional predictor in tree-based classification analysis. Another approach is to obtain a LR model after the tree is grown; for example, we can create dummy variables that correspond to each of the terminal nodes and include these as additional to predictors in the LR models (Zhang & Singer, 1999). In the classification tree with health conditions for African American participants, education is one of the predictors in the group of persons aged 77.5 years, and education is not a significant predictor in either the logistic or PH model. In PH and LR models, tests of interaction of education with sex and education with age indicate that interactions are not significant.

PH and LR models are based on observations with complete information on all the predictors. Tree-based classification models can be based either on observations with complete information, or on all the observations, including those with missing data. In the analysis we use for comparison (as shown in the tree diagrams of Figures 1 through 4), we use all of the data, including the observations with missing values. We also conduct tree-based classification analysis on the data with only complete information. The final nodes of trees are essentially the same, the difference being with respect to inclusion of either age or gender or both along their path. One significant difference is in the classification model with health conditions for African American participants, in which stroke is substituted for diabetes in the tree.

The present study has certain limitations. Our data are drawn from a representative sample of older community residents living in the South, they may not apply to other areas of the country, to situations where mortality rates are different, or to persons with different levels of impairment and disability, such as nursing home residents. These data apply to the years 1986–1999 and to people 65 years of age and older in 1986; they may not apply to more recent cohorts aged 65 and older whose level of education may be greater than that of those evaluated here and whose health status may be different.

Over time, as the demographic, health, and functional status characteristics of the older population
change, analyses such as these will need to be repeated, because both risk factors for mortality and groups at highest risk of death may change with the general aging of the population, an increase in their level of education, an improvement in their health habits, and greater efficacy of pharmaceutical and other interventions.

References


Received November 20, 2001
Accepted May 31, 2002
Decision Editor: Terrie T. Wetle, PhD