27-YEAR MORTALITY IN THE WESTERN COLLABORATIVE GROUP STUDY:
CONSTRUCTION OF RISK GROUPS BY RECURSIVE PARTITIONING

BY

D. CARMELLI, J. HALPERN, G. E. SWAN, A. DAME
M. McELROY, A. B. GELB, and R. H. ROSENMAN

TECHNICAL REPORT NO. 142

February 1991

PREPARED UNDER THE AUSPICES

OF

PUBLIC HEALTH SERVICE GRANT 2 R01 GM21215–15

DIVISION OF BIOSTATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA
27-Year Mortality in the Western Collaborative Group Study: Construction of Risk Groups by Recursive Partitioning

By

D. Carmelli, J. Halpern, G. E. Swan, A. Dame
M. McElroy, A. B. Gelb, and R. H. Rosenman

Technical Report No. 142

February 1991

Prepared Under the Auspices

Of

Public Health Service Grant 2 R01 GM21215-15

Division of Biostatistics

Stanford University

Stanford, California
27-Year Mortality in the Western Collaborative Group Study: 
Construction of Risk Groups by Recursive Partitioning

D. Carmelli,1 J. Halpern,2 G. E. Swan,1 A. Dame,1 
M. McElroy,1 A. B. Gelb,2 and R. H. Rosenman1

1Health Sciences Program, SRI International 
333 Ravenswood Avenue 
Menlo Park, California 94025

2Stanford University Medical Center 
Stanford, CA 94305

Address reprint requests to Dr. D. Carmelli, 
SRI International, 333 Ravenswood Avenue, 
Menlo Park, CA 94025
Abstract

The relationship of selected biological and behavioral characteristics measured at baseline examination to 27-year mortality due to coronary heart disease (CHD), cancers of all sites, and total mortality in the 3,154 men that form the Western Collaborative Group Study was investigated using tree-structured survival analysis or recursive partitioning (RP). Intake (1960-61) characteristics included in the present analyses were age, serum cholesterol, systolic blood pressure (SBP), cigarette smoking, body mass index, Type A/B behavior, and behavioral hostility. Tree-structured survival analysis for CHD mortality partitioned the cohort into six groups and identified five groups with distinct survival experience. Exceptionally high CHD mortality rates (17.3 and 14.6 per thousand) were experienced by 89 older men with elevated hostility ratings and SBP ≤ 150, and 238 men whose initial SBP was greater than 150 mm Hg. Younger men (age ≤ 48) with SBP ≤ 150 and with serum cholesterol levels > 227 had a death rate of 4.8 per thousand, compared with a rate of 1.7 in similar men with lower cholesterol levels. Applied to 27-year cancer mortality, the RP algorithm partitioned the cohort into four distinct survival groups. Younger (age < 45) Type B men had superior survival compared with Type A men of similar ages, and the proportion of ever cigarette smokers in these two groups was not statistically different. The results obtained by tree-structured survival analyses were compared with results obtained by Cox regression survival analyses.

Behavioral hostility, Mortality, Recursive partitioning, Interaction of risk factors
1. INTRODUCTION

Biological, physical, and lifestyle characteristics have been the primary focus of coronary heart disease (CHD) epidemiology over the past four decades. A consistent pattern of results has emerged from these studies with age, blood pressure, serum cholesterol, and cigarette smoking among the strongest and most consistent predictors [1, 2]. However, motivated by clinical observations [3, 4] and the fact that the traditional risk factors account for only a portion of the variance in CHD, considerable interest has developed in psychological and social characteristics as potential risk factors for CHD. One of the factors receiving major attention was Type A behavior pattern (TABP) [5-7], characterized by attributes such as hard driving, time urgency, and increased potential for hostility/anger dimensions [8].

The Western Collaborative Group Study (WCGS) was the first major prospective study to report a twofold excess in CHD among Type A subjects in its first-phase, 8.5-year incidence study from 1960 to 1969 [9, 10]. These findings stimulated many other studies in different populations of the U.S. and abroad [11-13]. However, the study results were not always consistent [14,15]. Moreover, a recent 22-year mortality follow-up of the WCGS [16,17] failed to support the positive association between TABP and long-term CHD mortality in this cohort. Despite this finding, there remained much interest in the relationship between components of TABP and long-term mortality in the WCGS. This interest is based on the recognition that the term "Type A behavior pattern" is a global descriptor of a complex set of behaviors, attitudes, and physiological characteristics, among which only a subset may be predictive of coronary heart disease [18]. A component of TABP receiving
much current attention as a predictor of CHD outcomes is hostility, defined broadly—depending on the method of assessment—either as a predisposition to respond directly or indirectly to challenge with various intensities of anger and related mood states, including behavioral ratings of irritation, annoyance, and resentment, [19] or as an attitude of cynical mistrust [20] and hostility toward others [21-22] in the use of questionnaire assessments. Both methods of assessment have been found to predict both the severity of coronary atherosclerosis and CHD incidence [23-26].

Although the investigation of hostility in the WCGS is not new, its relationship to mortality has not yet been studied. Previous investigations determined that: (1) individuals with "silent MI" at intake into the WCGS were more likely to exhibit manifest hostility during the Structured Interview than (SI) matched controls [27]; (2) the 4.5-year incidence of CHD was associated with a rated potential for hostility and anger expressiveness [28]; and (3) the 8.5-year incidence of CHD was associated with the behaviorally rated hostility component of the SI after controlling for other risk factors, including TABP, smoking, serum cholesterol, and systolic blood pressure [29, 30]. This latter study was a case-control study based on a subset of only 750 subjects from the WCGS.

In this report, data are presented on the long-term relationship of personal characteristics, including behavioral hostility assessed from the TABP Structured Interview [31, 32] and 27-year mortality due to CHD, cancer of all sites, and total mortality among the 3,154 men that form the Western Collaborative Group Study. Another important aspect of our study is the use of a novel statistical methodology, recursive partitioning (RP), also known as tree-structured survival analysis. The primary goal of an RP algorithm is
to make the finest possible partition of subjects' characteristics measured at an initial examination, so that the subsequent survival experience of a group of subjects associated with a partition can be reliably predicted. The RP algorithm makes no assumption about the shape of the underlying survival curve and can reveal important information about interactions among risk factors as they relate to an outcome of interest. Since limited experience with these methods [33, 35] found that the RP and regression approaches are often comparable in terms of predictions and classification ability, we applied both methods for a comparison of the results.

2. SUBJECTS AND METHODS

The current 27-year follow-up is the third phase of the Western Collaborative Group Study. The first phase, beginning in 1960-61, was a study of coronary heart disease incidence in 3,154 men followed for an average of 8.5 years. Male subjects free from coronary heart disease and other major health problems and 39 to 59 years of age at entry were recruited from 10 California companies. Methods and results for the 8.5-year follow-up have been presented previously [9, 10, 36, 37]. The second phase of the study, conducted during 1982-83, was a mailed questionnaire that determined the vital status of 99% of the cohort. During the follow-up period, 584 subjects died (2,532 subjects were alive at the date of contact in 1982-83, and 38 were not located) [16, 17]. The third phase of the study, beginning in October 1986 and completed in March 1989, included further monitoring of mortality and a reexamination of the surviving subjects.

2.1 Data collection

Using a series of follow-up steps, we determined the vital status of 97% of the cohort. Among the 3,154 healthy men who entered the study in 1960-61,
1,033 were deceased, 2,025 were alive at the date of contact, and 96 were not located and were therefore excluded from further analyses. Death certificates were obtained for all deceased subjects and coded by a trained physician using the ninth revision of the International Classification of Diseases (ICD9). Coronary heart disease encompassed ICD9 codes 410-414, 427-428, 429.2, and 518.4; 85% of these were due to ischemic heart disease (IHD) D (codes 410-414). Validation of cause of death was attempted using reports from next of kin, review of hospital records, and autopsy report if available. Such information was available for 60% of deaths occurring since 1982-83. A total of 353 deaths (34% of total deaths) were due to cardiovascular disease, 329 deaths (32%) were due to cancer of combined sites, and 351 (34%) were due to other causes. Subjects were designated as "died of cancer" if the underlying cause of death was a malignant neoplasm (ICD-9 codes 140-239). Death rates were based on age at death and person-years of observation. The average follow-up was 27.2 years for living subjects.

Intake variables included in the present analysis were: serum cholesterol, systolic blood pressure, cigarette smoking, and body mass index (weight in kilograms divided by the square of height in meters). Type A/B behavior was assessed by the Structured Interview, and behavioral hostility was recently scored from the intake interviews by two raters who were completely blinded to outcomes [29, 30]. Operationally, the definition of hostility used in this study differs from pencil-and-paper questionnaire definitions used by other investigators, in that specific behavioral indicators of hostility were scored. These include direct and indirect expressions of hostility toward the interviewer; self-reports of anger,
irritation, or annoyance involving a third party or a situation; and emotional expressions of hostility conveyed by word choice or tone of voice [29].

2.2 Statistical methods

An important aspect of the current study was the use of "tree-structured" survival analysis. This method, also termed recursive partitioning (RP), is based on extensions by Gordon and Olshen of classification and regression trees [31, 32]. With this method, the entire cohort is divided into two subgroups, as defined by the subject characteristics at intake (e.g., age, blood pressure, cholesterol) that differ most in their survival experience over the 27 years of follow-up. These two subgroups are again partitioned (each subgroup being split on the same or other risk factors that provide the greatest difference in survival), thus creating a tree structure. This process is continued until no further subdividing is worthwhile for prediction purposes. In sum, the tree-growing paradigm for the cohort examines every allowable split on each of the intake characteristics and selects (creates) subgroups that maximize a specified dissimilarity measure of the survival curves at each split. The procedure essentially generates the minimum number of cells needed to capture all the information in the covariates as they relate to survival. Best splits in the current algorithms were computed using the log rank statistic, which compares the Kaplan-Meier (KM) survival curves of the two resulting groups at each node [38].

Tree size is determined in three steps. First, a very large tree is constructed. The splitting criterion is used to create subgroups that could be expected to have the same survival experience as those of subjects
in future cohorts placed into the same nodes. Since initially such subgroups are small and statistically unstable, a second step uses a pruning algorithm to collapse this large tree upward, thereby creating a nested sequence of trees. The third step uses cross-validation to pick an "optimal" tree from this sequence. (See appendix for more details).

A further pertinent feature of an RP analysis is the notion of "surrogate" variables. The creation of surrogates is primarily for dealing with missing value problems and secondarily with finding the best substitute for a splitting covariate. As opposed to conventional regression procedures, which will discard cases with any missing covariate, or replace a missing covariate with its mean value, in the RP algorithm cases with missing values on the best splitting variable are sent to the left or right daughter node in accordance with the first surrogate on which they have a nonmissing value.

Another use of surrogates is in the ranking of variables in terms of their potential effect on forming subgroups. For a given nonterminal node, the first surrogate is the variable that places the highest percentage of subjects in the same daughter nodes as the variable creating the best split of that node. The second surrogate is the variable that places the second-highest percentage of such subjects in the same nodes, and so on. Thus at every nonterminal node, an ordered list of surrogates or substitute variables is compiled.

3. RESULTS

3.1 Tree-structured survival analysis for CHD mortality

Covariates selected for this analysis were: age, systolic blood pressure, total serum cholesterol, cigarette smoking, highest education level, global Type A behavior, and the behaviorally rated hostility component
of the SI administered in 1960-61. Each of these covariates is a recognized CHD risk factor found to be prospectively related to CHD incidence and survival in previous studies of the WCGS [10, 17]. Other than Type A/B behavior pattern and smoking (ever smoked, yes or no) all variables were treated as continuous variables.

The results of the RP analysis are presented in Figure 1. Terminal nodes are indicated by squares and intermediate nodes by circles. The numbers in each square or circle are the number of survivors and the number of CHD deaths. At any split, left branches are associated with better survival. The Kaplan-Meier survival curves for the final six terminal nodes are shown in Figure 2.

Tree-structured survival analysis identified systolic blood pressure > 150 mm Hg as the first split point (Figure 1). This divided the cohort into two groups of 2,820 and 238 subjects; better survival was indicated for subjects whose systolic blood pressure was 150 mm Hg or lower (Figure 1). The group of subjects whose intake SBP was higher than 150 mm Hg formed the first terminal node (group I). First and second surrogate covariates at this node were total serum cholesterol and smoking.

Subjects with SBP ≤ 150 mm Hg were further subdivided by age into those above and below age 48; surrogate covariates for this node were smoking and education. Subjects 49 years of age and older then were subdivided by the rated hostility scores into two terminal subgroups of 89 and 812 subjects. Older subjects whose hostility ratings were > 1.3 standard deviations above the mean had the worst survival experience in this cohort (see Figure 2).

The younger group of subjects with SBP ≤ 150 mm Hg was further subdivided by serum cholesterol into those above and below 227 mg/100 ml. As
shown in Figure 2, younger men with cholesterol levels above 227 mg/100 ml had poorer survival than their counterparts with lower levels. Surrogate covariates for this node were smoking and education. Finally, young subjects whose cholesterol level was 227 mg/100 ml or lower were further subdivided on SBP. As shown in Figure 2, this last split marked the most favorable survival subgroup (VI) of this cohort. The right-branch subgroup (V) of 462 men whose blood pressure levels were intermediate, (124 mm Hg ≤ SBP ≤ 150 mm Hg) had poorer survival than their counterparts with the lowest SBP levels. Surrogate covariates at this node were cholesterol and smoking.

The resulting six terminal nodes divide the long-term CHD mortality experience of this cohort into five distinct groups. The survival curves of all six terminal nodes are shown in Figure 2. The favorable group (27-year survival 98%) consists of young men (intake age under 49) whose intake SBP was under 125 mm Hg and their cholesterol levels were below 228 mg/100 ml. The groups with poorest survival consisted of the older subset of 89 men whose hostility scores were > 1.3 standard deviations above the mean (27-year survival 66%) and the 238 men whose initial SBP was higher than 150 mm Hg (27 year survival 71%). Mortality rates in these two subgroups were 3.0 to 3.6 times as high as those of the intermediate subgroup of younger subjects whose cholesterol levels were above 227 mg/100 ml (see Table 1). The remaining group of men with cholesterol levels ≤ 227 mg/100 ml and SBP > 124 mm Hg had an overall 27-year survival rate of 93%. Gehan's generalized Wilcoxon test indicated nominally significant differences in survival (p < 0.01) between all pairwise comparisons of groups III, IV, V, and VI.

To examine the profile of subjects in the different risk groups, we tabulated the characteristics used for classification at each final node. In
addition, mean levels and corresponding percentages of these characteristics in the whole cohort are also presented. As seen from Table 1, the two groups with the poorest survival differ markedly on age, SBP, and the rated hostility components. However, both show a high prevalence of Type A (72% and 63% compared with 50% in the whole cohort) and smoking (83% and 88% ever smoked, compared with 77%). The intermediate survival group of younger men with elevated cholesterol levels (terminal node IV) is not noticeably different from the group with the highest CHD mortality (II) in mean SBP and percent ever smoked. Groups II and IV do differ markedly on hostility scores and the percentage of Type A subjects (49% Type A in group IV, compared with 72% in group II). Groups V and VI show minor differences in the prevalence of Type A rated subjects and ever smokers.

3.2 Tree-structured survival analysis of cancer mortality

Tree-structured survival analysis of the 329 deaths due to cancer identified age as the best first and second predictor variables of cancer survival in this cohort (Figure 3). This analysis divided the cohort into three age groups: 39 to 44, 45 to 51, and 52 to 59 years old when first examined. Survival in the oldest age group was the poorest, with subjects in this group forming the first terminal node (group I). At the first split on age, surrogate variables were smoking and education; at the second split on age, surrogates were rated hostility scores and smoking. The remaining 1,027 subjects age 45 to 51 years at intake were further subdivided by smoking history into two terminal nodes: those who never smoked (236 subjects) and the remaining group of 791 subjects who reported ever smoking cigarettes. The subgroup of never smokers had the most favorable survival experience due to cancer (27-year survival of 98%, compared with 87% in smokers). Total
cholesterol and hostility were, decreasing in order of significance, the two surrogate variables at this node. The youngest group of subjects, age 39 to 44 at intake, were further subdivided by SI behavior pattern in Type A and Type B. Type A had poorer survival than Type B. Surrogate variables at this node were hostility and total cholesterol.

The resulting five terminal nodes divide cancer mortality in this cohort into four distinct groups. Kaplan-Meier survival curves of all five terminal nodes are shown in Figure 4. Poorest survival was in the oldest group of subjects, age 52 to 59 at intake. The intermediate survival group of ever smokers, age 45 to 51, had a mortality rate of 6.0 per thousand, compared with 9.0 per thousand in the older group of subjects forming the first terminal node. The remaining younger Type B subjects had superior survival compared with their Type A counterparts of similar ages (mortality rate of 1.9, compared with 3.9). This group of Type B subjects did not differ in mortality rate from the group of older subjects who never smoked.

Table 2 summarizes intake characteristics of subjects at each terminal node and in the whole cohort. The two groups with the poorest survival (I and II) show the highest percentages of ever smokers, and education levels that are somewhat lower than those of the whole cohort. Highest education levels of subjects at the third terminal node (Type A subjects age 39 to 44) were not different from those of never smokers, who formed the most favorable survival group (V). Among subjects who formed group III, 100% of whom were Type A, 78% reported ever smoking cigarettes, compared with 74% among the Type B subjects in group IV, who had superior survival.

3.3 Multivariate regression analyses of survival

The same covariates used in the RP analyses were included in the separate multivariate Cox regression analyses of CHD, cancer, and total
mortality. Table 3 describes the results of Cox regression analyses stratified by age at intake. The stratification by age allows the testing of the interaction effect with age identified in the RP algorithm of both CHD and cancer mortality. In addition, when total mortality was subjected to RP analysis, a single split at age 48 formed the best survival "tree" of these data. This RP algorithm for total mortality divided the whole cohort into a group of 1,017 older subjects (age > 48), whose mortality rate was 25.1 per thousand, and a group of 2,041 younger subjects (age ≤ 48), whose rate was 10.0 per thousand. Surrogate covariates at this single node were SBP and smoking.

Examining the results in Table 3, we observe that all the variables found to be significant in the tree-structured analyses of CHD mortality retained their significance in the Cox regression analyses. Only education (in the separate analyses of both age groups) and smoking (in the subgroup of older subjects) retained a significant association to CHD mortality. We can infer the interaction between age and behavioral hostility, age and total cholesterol, and age and systolic blood pressure from the significant differences in the corresponding beta coefficients of the two age groups.

The results of the Cox regression analyses of cancer mortality are in general agreement with those of the RP analysis. We do observe, however, that behavioral hostility in the Cox analysis replaced the global TABP identified by the RP algorithm. Finally, comparison of the beta coefficients of predictors of total mortality in the two age groups demonstrates again the interaction between age and risk factors in predicting 27-year survival in this cohort. Specifically, hostility emerged as a significant independent predictor of mortality in older subjects. The relationship of both total
cholesterol and smoking to mortality of all causes was stronger in younger subjects than older subjects. Tests for differences in the corresponding beta coefficients were statistically significant.

4. DISCUSSION

In the present 27-year mortality analysis of the WCGS cohort, we have found that age, systolic blood pressure, serum cholesterol, and behavioral hostility measured at the intake examination are contributing factors to a differential pattern of CHD mortality in this cohort. Tree-structured survival analysis identified age, systolic blood pressure, and behaviorally rated hostility as the main characteristics that generated the two groups of subjects with exceptional high CHD mortality rates. Multivariate Cox regression analysis confirmed the independent contribution of behavioral hostility to long-term mortality from CHD in the older group of subjects.

Findings of these analyses suggest that only a relatively small group of older hostile individuals (n = 89) carried the exceptional risk for CHD mortality. We also observe that this group had the highest prevalence of Type A and ever-smoking subjects. In the current study, the relationship of hostility to smoking was weak (r = 0.07). However, the two subgroups showing the highest CHD mortality rates also had the highest percentage of ever smokers.

That the group with high hostility ratings had a mortality experience equal to, if not slightly greater than, the high blood pressure group suggests similar mechanisms contributing to damage of the arterial endothelium. Williams and Barefoot [39] hypothesize several pathways by which hostility can cause CHD, including increased testosterone, catecholamine, and cortisol secretion and reduced parasympathetic antagonism
of sympathetic nervous system effects that could lead, over time, to endothelial injury, atherosclerotic progression, and the onset of CHD. It is interesting to note that although research has yet to demonstrate the differences between the underlying pathophysiology associated with elevated blood pressure and that associated with elevated hostility, the end result shows a striking similarity in eventual CHD mortality rates.

Two types of hostility measurements have been commonly used: (1) questionnaires such as the MMPI and the Ho Scale, and (2) behavioral assessments such as potential for hostility (19,28) and the Hostility ratings used in the current study (24,30). Costa et al. [40] point out that different components of hostility, such as antagonistic hostility, and neurotic hostility, may be differentially related to CHD. Paper-and-pencil questionnaires like the Ho Scale may be picking up a neurotic type of hostility related to other negative affects such as anxiety and depression, while behavioral ratings measure an antagonistic interpersonal style marked by disagreeableness or being uncooperative. In a study using the Buss-Durkee Hostility Inventory (BDHI) [41], scores on subscales measuring neurotic hostility were negatively associated with CHD and those measuring antagonistic hostility were positively associated with CHD. Dembroski (23) found his measure of antagonistic hostility to predict incidence of CHD, whereas the measure reflecting neuroticism did not.

Previous studies have also found hostility to be more predictive of CHD at younger ages [23, 24, 42], except for the Western Electric Study [43], in which subjects had a mean age of 47 years at intake. In this later study significant associations between Ho Scale scores and age at the 20-year follow-up were found, and scale scores of hostility predicted both CHD
incidence and total mortality. In the current study, we have found hostility/age interaction effects for both CHD and total mortality.

It has also been hypothesized that, since Type A behavior and its components are causally related to CHD by a physiological mechanism, it is possible that other disease processes, such as cancer, may be influenced by such behaviors [44, 45]. There is recent support [43, 45] for the hypothesis that hostile Type A individuals have an increased risk for cancer (presumably through the effects of chronic stress on a reduced immune capacity).

Shekelle [43] found a significant relationship between hostility and crude 20-year mortality from malignant neoplasms in a large sample of employed, middle-aged men. Results from the current study lend support to these findings. Both the tree-structured survival analysis and the Cox regression analyses suggest an independent relationship between Type A hostility and cancer mortality in this cohort.

Tree-structured survival analysis identified a group of 661 younger men (mean age at intake = 41.4 years), all of whom were rated Type A, who experienced a cancer mortality rate of 3.3 per thousand, a rate significantly higher than that experienced by a similarly aged group of 737 men (mortality rate 1.9 per thousand), all of whom were rated as Type B. These two groups of subjects were similar with respect to smoking history, education, and total serum cholesterol. However, the group of Type A men with higher mortality rates was also rated higher on behavioral hostility than the group that was all Type B. The role of behavioral hostility in cancer mortality was supported by Cox regression analyses, where hostility was a significant independent risk factor for cancer mortality in the younger age group (below age 49 years at intake). Taken together, these findings support the
hypothesis forwarded by Suls [47, 48] that TABP may be associated with multiple disease processes.

One of our primary goals in evaluating the relationship between risk factors and long-term mortality in this cohort was to identify subgroups of individuals with distinct characteristics and distinct survival experiences. The identification of these subgroups underscores the utility of the novel statistical approach of recursive partitioning or tree-structured survival analysis. The analysis presented here illustrates the prospective use of this methodology for further classification of subjects and its implications for intervention strategies that are tailored to the individual. Other advantages of this approach include the fact that no modeling assumptions are necessary in order to use it (and hence it has broad applicability) and its automatic identification of interactions (i.e., synergistic effects among risk factors).

Finally, some limitations on the use of the classification-tree approach should be mentioned. Constructing a tree is a stepwise procedure. As with stepwise regression, the tree procedure should not be used mechanically to determine either the "best" variables or the "relative importance" of the variables. If a variable is never split on, then one might infer that it has little association with the outcome variable, although its effect may be only masked by other variables. For example, smoking, evaluated in the current study as having ever smoked cigarettes, was not a major discriminator of CHD mortality in this cohort. Since a significant number of subjects, 77% of the cohort, reported ever cigarette smoking, the relatively small number of never smokers may explain, in part, the lack of impact of smoking in the RP analysis of CHD mortality. Nonetheless, smoking did emerge as a surrogate covariate at all split points in the tree-structured survival analysis.
REFERENCES


Captions

Figure 1. Results of multivariate recursive partitioning of 27-year mortality from CHD in the WCGS. Roman numerals designate terminal subgroups. Numbers in circles and squares indicate number of alive and deceased subjects in each subgroup.

Figure 2. Actuarial survival curves for the six terminal subgroups in the CHD mortality "tree" shown in Figure 1.

Figure 3. Results of multivariate recursive partitioning of 27-year mortality from cancer in the WCGS.

Figure 4. Actuarial survival curves for the five terminal subgroups in the cancer mortality "tree" shown in Figure 3.

Acknowledgements

This study was supported by grant HL 32795 from NHLBI, grant GM21215-15 from PHS, and a grant from R. J. Reynolds-Nabisco. The authors thank Richard Olshen for his help with the application of tree survival analysis to these data and Michael Hecker for the development and training of raters in the assessment of the components of the SI.
Fig. 1.
Table 1.  CHD mortality rates per 1,000; means and ranges of subjects' intake characteristics at terminal nodes

<table>
<thead>
<tr>
<th>Terminal node groups</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>238</td>
<td>89</td>
<td>812</td>
<td>805</td>
<td>462</td>
<td>652</td>
<td>3,058</td>
</tr>
<tr>
<td>ID deaths</td>
<td>68</td>
<td>30</td>
<td>117</td>
<td>92</td>
<td>33</td>
<td>13</td>
<td>353</td>
</tr>
<tr>
<td>Person-years (1000)</td>
<td>4,664</td>
<td>1,731</td>
<td>17,676</td>
<td>19,175</td>
<td>11,201</td>
<td>16,326</td>
<td>70,773</td>
</tr>
<tr>
<td>Age</td>
<td>14.6</td>
<td>17.3</td>
<td>6.6</td>
<td>4.8</td>
<td>2.9</td>
<td>0.8</td>
<td>5.0</td>
</tr>
<tr>
<td>Age (years)</td>
<td>48.7</td>
<td>53.1</td>
<td>52.9</td>
<td>43.1</td>
<td>43.1</td>
<td>42.6</td>
<td>46.3</td>
</tr>
<tr>
<td>[39 - 59]</td>
<td>[49 - 59]</td>
<td>[49 - 59]</td>
<td>[39 - 48]</td>
<td>[39 - 48]</td>
<td>[39 - 48]</td>
<td>[39 - 59]</td>
<td></td>
</tr>
<tr>
<td>SBP (mm Hg)</td>
<td>163.6</td>
<td>128.7</td>
<td>127.0</td>
<td>126.1</td>
<td>135.0</td>
<td>116.5</td>
<td>128.6</td>
</tr>
<tr>
<td>[152 - 230]</td>
<td>[106 - 150]</td>
<td>[100 - 150]</td>
<td>[100 - 150]</td>
<td>[126 - 150]</td>
<td>[98 - 124]</td>
<td>[98 - 230]</td>
<td></td>
</tr>
<tr>
<td>Total cholesterol</td>
<td>238.6</td>
<td>237.7</td>
<td>230.8</td>
<td>262.1</td>
<td>195.4</td>
<td>193.2</td>
<td>226.5</td>
</tr>
<tr>
<td>(mg / 100 ml)</td>
<td>[112 - 645]</td>
<td>[133 - 386]</td>
<td>[103 - 390]</td>
<td>[228 - 414]</td>
<td>[110 - 227]</td>
<td>[111 - 227]</td>
<td>[103 - 645]</td>
</tr>
<tr>
<td>Hostility a</td>
<td>-0.04</td>
<td>2.4</td>
<td>-0.24</td>
<td>-0.01</td>
<td>-0.03</td>
<td>0.004</td>
<td>0.00</td>
</tr>
<tr>
<td>[-1.2 - 3.6]</td>
<td>[1.5 - 5.2]</td>
<td>[-1.2 - 1.3]</td>
<td>[-1.2 - 3.9]</td>
<td>[-1.2 - 3.9]</td>
<td>[-1.2 - 5.2]</td>
<td>[-1.2 - 5.2]</td>
<td></td>
</tr>
<tr>
<td>Ever smoked</td>
<td>83%</td>
<td>88%</td>
<td>78%</td>
<td>79%</td>
<td>73%</td>
<td>71%</td>
<td>77%</td>
</tr>
<tr>
<td>&quot;pe A</td>
<td>63%</td>
<td>72%</td>
<td>54%</td>
<td>49%</td>
<td>47%</td>
<td>42%</td>
<td>50%</td>
</tr>
</tbody>
</table>

Individual ratings scored 1 - 5 were standardized to mean zero and S.D. = 1.
Table 2. Cancer mortality rates per 1,000; means and ranges of subjects' intake characteristics at terminal nodes

<table>
<thead>
<tr>
<th>Terminal node groups</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>633</td>
<td>791</td>
<td>661</td>
<td>737</td>
<td>236</td>
<td>3,058</td>
</tr>
<tr>
<td>Cancer deaths</td>
<td>127</td>
<td>106</td>
<td>53</td>
<td>35</td>
<td>8</td>
<td>329</td>
</tr>
<tr>
<td>Person-years</td>
<td>13,009</td>
<td>17,606</td>
<td>16,107</td>
<td>18,223</td>
<td>5,827</td>
<td>70,772</td>
</tr>
<tr>
<td>Rate / 1000</td>
<td>9.8</td>
<td>6.0</td>
<td>3.3</td>
<td>1.9</td>
<td>1.4</td>
<td>4.6</td>
</tr>
<tr>
<td>Lung cancer a</td>
<td>43 (34%)</td>
<td>39 (37%)</td>
<td>15 (28%)</td>
<td>13 (37%)</td>
<td>1 (13%)</td>
<td>111 (34%)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>54.8</td>
<td>47.7</td>
<td>41.4</td>
<td>41.4</td>
<td>47.7</td>
<td>46.3</td>
</tr>
<tr>
<td></td>
<td>[52 - 59]</td>
<td>[45 - 51]</td>
<td>[39 - 44]</td>
<td>[39 - 44]</td>
<td>[45 - 51]</td>
<td>[39 - 59]</td>
</tr>
<tr>
<td>Ever smoked</td>
<td>81%</td>
<td>100%</td>
<td>78%</td>
<td>74%</td>
<td>0%</td>
<td>77%</td>
</tr>
<tr>
<td>Type A</td>
<td>59%</td>
<td>52%</td>
<td>100%</td>
<td>0%</td>
<td>40%</td>
<td>50%</td>
</tr>
<tr>
<td>Education rank</td>
<td>1.82</td>
<td>1.82</td>
<td>2.11</td>
<td>2.05</td>
<td>2.11</td>
<td>1.96</td>
</tr>
<tr>
<td></td>
<td>[1 - 3]</td>
<td>[1 - 3]</td>
<td>[1 - 3]</td>
<td>[1 - 3]</td>
<td>[1 - 3]</td>
<td>[1 - 3]</td>
</tr>
<tr>
<td>Hostility b</td>
<td>0.02</td>
<td>0.04</td>
<td>0.15</td>
<td>-0.15</td>
<td>-0.17</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>[-1.2 - 5.2]</td>
<td>[-1.2 - 5.2]</td>
<td>[-1.2 - 5.2]</td>
<td>[-1.2 - 4.1]</td>
<td>[-1.2 - 4.1]</td>
<td>[-1.2 - 5.2]</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>232.5</td>
<td>229.3</td>
<td>225.9</td>
<td>220.0</td>
<td>223.0</td>
<td>226.5</td>
</tr>
<tr>
<td>(mg / 100 ml)</td>
<td>[103 - 386]</td>
<td>[111 - 400]</td>
<td>[111 - 645]</td>
<td>[110 - 390]</td>
<td>[130 - 324]</td>
<td>[103 - 645]</td>
</tr>
</tbody>
</table>

a Number and percent of lung cancer based on total cancer deaths at each terminal node.

b Individual ratings scored 1 - 5 were standardized to mean zero and S.D. = 1.
Table 3. Cox regression analyses by age at intake, relating behavioral hostility, biologic, and sociodemographic characteristics at baseline examination (1960-61) to CHD, cancer, and total mortality during 27 years of follow-up in the WCGS

<table>
<thead>
<tr>
<th></th>
<th>CHD mortality</th>
<th>Cancer mortality</th>
<th>Total mortality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Younger</td>
<td>Older</td>
<td>Younger</td>
</tr>
<tr>
<td></td>
<td>Age ≤ 48</td>
<td>Age &gt; 48</td>
<td>Age ≤ 48</td>
</tr>
<tr>
<td></td>
<td>n=166</td>
<td>n=187</td>
<td>n=155</td>
</tr>
<tr>
<td>Dependents</td>
<td>β</td>
<td>Z</td>
<td>β</td>
</tr>
<tr>
<td>σ</td>
<td>0.045</td>
<td>1.57</td>
<td>0.054</td>
</tr>
<tr>
<td>P</td>
<td>0.031</td>
<td>6.69  ***</td>
<td>0.026</td>
</tr>
<tr>
<td>cholesterol</td>
<td>0.009</td>
<td>6.98  ***</td>
<td>0.007</td>
</tr>
<tr>
<td>H</td>
<td>-0.027</td>
<td>-0.84</td>
<td>0.034</td>
</tr>
<tr>
<td>education</td>
<td>-0.267</td>
<td>-3.02  **</td>
<td>-0.171</td>
</tr>
<tr>
<td>ever smoked</td>
<td>0.512</td>
<td>2.24</td>
<td>0.322</td>
</tr>
<tr>
<td>BP</td>
<td>-0.162</td>
<td>-1.01</td>
<td>-0.027</td>
</tr>
<tr>
<td>nitility</td>
<td>0.000</td>
<td>0.00</td>
<td>0.230</td>
</tr>
</tbody>
</table>

* < 0.05, **p < 0.01, ***p < 0.001 (one tail test).

1. Cox regression coefficient
2. β / standard error of β
3. BP = Type A behavior pattern
4. Number of deaths
Appendix

A tree begins with the single node containing all N subjects. For each possible value, $a$, of covariate $v_i$, $i, \ldots, I$, a trial split of this node is made. If subject $\lambda$'s value of $v_i \leq a$, then $\lambda$ goes into the left group; if $v_i > a$, then $\lambda$ goes into the right one. Some measure, $l_{ia}$, of the split is made; we used the value of the logrank statistic which compares the Kaplan-Meier (KM) survival curves of the two resulting groups. For each $i$ this determines a candidate split, $CS_i = (i, a^*)$, where $l_{ia^*} = \max_a l_{ia}$. The first split, $(i_0, a_0)$, is made with the covariate $v_{i_0}$ corresponding to $\max D(CS_i)$ where $D$ is a second measure of a split; for $D(CS_i)$ we used the decrease in $R(T)$ when the tree is grown with $CS_i$. Here $R(T)$ is the sum (weighted by the proportion of subjects in the nodes) over the terminal nodes of $T$ of the 4th power Wasserstein distance between the KM curve of the subjects in a node and the survival curve that decreases at a single point which is closest to that KM curve in Wasserstein distance. The 4th power Wasserstein distance between two survival curves is the 4th root of the integral of the 4th power of the horizontal distance between the curves. In the case of the first split, $T$ is the whole data set.

We will call covariate $v_{i'}$ the first surrogate for $v_i$ at a particular node if the two groups it produces for some value $a'$ are, subject by subject, more like the groups produced by $(v_i, a)$ than those for any other $(v_j, a''; j \neq i, i')$ at that node; $v_{i'}$ is not necessarily the covariate corresponding to $\max_{j \neq i} l_{jb}$. If a subject has a missing value for $v_{i_0}$ then the surrogate, $v_{i_0}'$, is used to categorize that subject, see BFOS(1984), pp. 142-146, for details. This process is iterated until some criterion, such as a minimum node size for each final node is met. The final tree is $T^*$.

$T^*$ has been grown to optimally predict survival for the data it is being used to analyze, so initially $T^*$ may be too big in that future data to be classified by $T^*$ can not be expected to be sorted out as well as $T^*$ sorts the data from which it is grown. This is like using too many covariates in a regression analysis.

One remedy for this "overfitting" is to report a pruned back subtree, $T^*_5$, determined by cross-validation. A measure of the cost, $c(T, \alpha) = R'(T) + \alpha |T|$, of a tree is adopted.
\(\alpha\) is a constant cost per node and \(|T|\) is the number of final nodes in the tree, \(T\). Thus, the specificity given by a complex tree with many final nodes is balanced by a cost for that complexity. For example \(R'(T)\) could be a global measure of how homogeneous the final nodes in a tree are. If each final node in \(T\) is composed of subjects with very different expected survivals, then \(R'(T)\) should be large; but if there is only a single subject in each final node then \(R'(T)\) should be small, perhaps 0. We used \(R'(T) = R(T)\). Under very modest conditions, (BFOS(1984), Sec 10.2 and Butler et al. (1990)), for each value of \(\alpha\) there is a unique smallest optimally pruned subtree \(T^*_\alpha\) as measured by the minimum value of \(R'(T^*_\alpha) + \alpha |T^*_\alpha|\) over all of the subtrees, \(\{T^*_k\}\), of \(T^*\). Furthermore, increasing \(\alpha\) from 0, it is possible to find \(\{\alpha^*_k\}\), \(k = 1, \ldots, K\) for which \(T^*_\alpha = \) (the root node with all \(N\) subjects) and \(T^*_\alpha\) is strictly contained in \(T^*_\alpha\) for all \(\alpha < \alpha^*_k\).

Now randomly partition the original data into two groups, \(g_{11}\), with \((1/M)^{th}\) of the data and \(g_{12}\) with the remaining \((1-1/M)^{th}\). Grow a new tree, \(T_1\), using only the subjects in \(g_{12}\). Then use \(T_1\) to classify the subjects in \(g_{11}\). This process of dividing the data and growing a new tree \(T\) is repeated \(M\) times, such that \(\{g_{11}, g_{12}, \ldots, g_{M1}\}\) is a random partition of the original \(N\) subjects. For each \(\alpha^*_k = \sqrt{\alpha^*_k \alpha^*_{k+1}}\) and each partition, \(j\), let \(C_{kj}\) be the sum, node by node, of the 4\(^{th}\) power Wasserstein distance between the KM survival curve of the subjects in a node of \(T^*_{j\alpha^*_k}\) applied to \(g_{j1}\) and the KM survival curve of the subjects in the corresponding node of \(T^*_{j\alpha^*_k}\) applied to \(g_{j2}\), weighting each term of the sum by the proportion of the subjects of \(g_{j1}\) in that term.

The tree \(T^*_{\alpha^*_k}\) is finally reported out, where mean \(\text{mean } C_{mk} = \min_k \text{mean } C_{mk}\).