AUTOMATED CLASSIFICATION OF FATAL CARDIOVASCULAR END POINTS IN LARGE SCALE MULTI-CENTER CLINICAL TRIALS

by

Michael D. Thorn

Department of Biostatistics, University of North Carolina at Chapel Hill, NC.

Institute of Statistics Mimeo Series No. 2104T

July 1992

# Automated Classification of Fatal Cardiovascular End Points in Large Scale Multi-Center Clinical Trials

by

Michael D. Thorn

A Dissertation submitted to the faculty of The University of North Carolina at Chapel Hill in partial fulfillment of the requirement for the degree of Doctor of Public Health in the Department of Biostatistics in the School of Public Health.

Chapel Hill

1992

Approved by:

Adviser

Reader

Reader

Reader

Reader

MICHAEL D. THORN. Automated Classification of Fatal Cardiovascular End Points in Large Scale Multi-Center Clinical Trials (Under the direction of C.E. Davis)

## ABSTRACT

The validity and reproducibility of the primary end points in clinical trials is vital to it's scientific integrity. Classification of fatal events for cause-specific mortality can be an expensive and time-consuming activity. Two statistical discriminant procedures (logistic discriminant analysis and recursive partitioning) and an expert system were examined and compared amoung themselves and with three methods used to classify fatal events in the Lipid Research Clinics (LRC) Coronary Primary Prevention Trial (CPPT). The primary objective of this research was to emulate the human classification, not to validate the LRC classification of fatal events. Of 3,806 participants in the LRC CPPT, 139 fatal events occurred. Only three events could not be used in this research. Of these methods, recursive partitioning also known as binary tree-structure classifiers, using the CART software had the highest overall agreement with the LRC classification. Only three percent of the events were misclassified. Only the clinic cardiologist's classification was more accurate. Expert systems, and logistic discriminant were less accurate although both methods had difficultly with observations with missing data. In the absence of missing data, expert systems had an overall accuracy equal to an independent cardiologist's review of the medical records; both methods misclassified six percent of these fatal events. Logistic discrimination and a nosologist's classification based only on death certificate had the lowest overall accuracy. Recursive partitioning is recommended for automated classification of end points in a clinical trial. The use of expert systems is also recommended, if previous

data for the development of a classification tree is unavailable and missing data can be minimized.

## Table of Contents

List of Tables

## List of Figures

List of Abbreviations

| | |
|---|---|
| **ABENZ** | abnormal enzymes present [a variable on the LRC CPPT dataset] |
| **ARIC** | Atherosclerosis Risk In Communities |
| **CARDPN** | ischemic cardiac pain present [a variable on the LRC CPPT dataset] |
| **CART** | Classification and Regression Trees |
| **CHD** | coronary heart disease |
| **CPPT** | Coronary Primary Prevention Trial |
| **CPR** | Central Patient Registry and Coordinating Center |
| **CVD** | Cerebrovascular Disease |
| **DARTRUP** | has death certificate with consistent underlying or immediate cause (atherosclerotic arterial aneurysm with rupture) [a variable on the LRC CPPT dataset] |
| **DATHCVD** | has death certificate consistent with underlying immediate cause (atherosclerotic cerebrovascular disease) [a variable on the LRC CPPT dataset] |
| **DCDCHF** | has death certificate consistent with underlying immediate cause (atherosclerotic coronary heart disease) [a variable on the LRC CPPT dataset] |
| **DCSCHF** | has death certificate consistent with underlying immediate cause (atherosclerotic coronary heart disease), but neither adequate preterminal documentation of the event nor previous atherosclerotic coronary heart disease [a variable on the LRC CPPT dataset] |
| **DIAGECG** | diagnostic ECGs present [a variable on the LRC CPPT dataset] |
| **DPREFGAN** | has death certificate with consistent underlying or immediate cause (atherosclerosis of peripheral arteries with gangrene) [a variable on the LRC CPPT dataset] |
| **DWIN1** | death occurring within one hour after the onset of severe symptoms or having last been seen without symptoms [a variable on the LRC CPPT dataset] |
| **EQECG** | equivocal electrocardiogram present [a variable on the LRC CPPT dataset] |

| | |
|---|---|
| **EQENZ** | equivocal enzymes present [a variable on the LRC CPPT dataset] |
| **FACT** | Fast Algorithm for Classification Trees |
| **FUS** | Follow-Up Study of the North American Prevalence Study |
| **HOSPMI** | pre-terminal hospitalization with definite or suspect myocardial infarction [a variable on the LRC CPPT dataset] |
| **HPDP** | Hypertension Detection and Follow-UP Program |
| **ICDA** | International Classification of Diseases, Adapted |
| **LDF** | linear discriminant function |
| **LRC** | Lipids Research Clinics |
| **MI** | myocardial infarction |
| **MLE** | maximum likelihood estimator |
| **NHLBI** | National Heart, Lung, and Blood Institute |
| **NOCHRN1** | no known non-atherosclerotic acute or chronic process or event that could have been potentially lethal [a variable on the LRC CPPT dataset] |
| **NOCHRN2** | no known non-atherosclerotic acute or chronic process or event that could have been potentially lethal [a variable on the LRC CPPT dataset] |
| **ONLY1** | suspect transient cerebral ischemic attack has occurred only one time (non-fatal end point item) [a variable on the LRC CPPT dataset] |
| **PREVANG** | previous definite angina, or suspect or definite myocardial infarction when no cause other than atherosclerotic coronary heart disease could be ascribed as cause of death [a variable on the LRC CPPT dataset] |
| **QDF** | quadratic discriminant function |
| **RECPAM** | recursive partition and amalgamation [a computer program for recursive partition for censored survival data] |
| **SDMB** | Safety Data Monitoring Board |
| **SOLVD** | Studies of Left Ventricular Dysfunction |
| **TIA** | transient cerebral ischemic attack |
| **UNC** | University of North Carolina |

**VPXDEATH** expert system death classification [a variable on the VP-Expert expert system knowledge base]

## I. Introduction and Literature Review

### A. Importance

In epidemiologic studies, identification and classification of events, or end points are often essential to assessing the primary hypothesis. Observations will generally be classified into one of n categories, of which n-1 categories will be considered "events." An observation incorrectly classified as an event is defined as a false positive, and an event which is incorrectly considered to be a non-event is defined as a false negative. Sensitivity is defined as the proportion of events correctly classified as events, and specificity is the proportion of non-events correctly classified as non-events. The misclassification ratio is the proportion of observations which are false negative or false positive. The reliability is one minus the misclassification ratio.

A critical element of any epidemiologic study or clinical trial in medical/epidemiological cardiovascular research is the validity and reproducibility of the classification of these end points. This can be a time-consuming and expensive activity in long term clinical trials. The cost of this process can easily run into hundreds of thousands of dollars. Its duration may be also be a threat to the consistency of this process over time. One suggestion for the standardization and cost efficiency of this process is to develop an automated system such as an expert system. Such a system could incorporate the knowledge and expertise of *human experts* which classify these morbidity or mortality end points. If this expert system were reliable, it could decrease the time and expense of this process. Another option is to use traditional statistical classification techniques, such as discriminant analysis or more recently developed nonparametric methods such as classification and regression trees (also known as recursive partitioning) to *classify* an event as an end point. This research develops such an expert system and uses it on end points previously classified by *human experts*. It then evaluates these results, comparing them to various standard expert system and statistical classification techniques. The objective of this research is not to validate the correctness of the classifications made by human experts. Only the agreement of these automated methods with human classification methods is evaluated.

## B. Review of Past Classifications and Suggested Improvements

There has been a substantial amount of research regarding the classification of mortality, specifically the quality of death certificate data for epidemiologic research. Remington (1984) and Kuller et al. (1980) have studied and suggested improvements in the mortality classification of large scale multicenter prevalence studies and clinical trials.

### 1. Review of Mortality

Many investigators have studied death certificate data looking at all cause mortality [Alderson and Meade (1967), Andersen et al. (1991a, 1991b), Curb et al. (1983), Gittlesohn and Senning (1979), Glasser (1981), James et al. (1955), Kohn (1982), Kuller et al. (1967), Rigdon (1981), Wingrave et al. (1981)]. Others have studied selected cause-specific mortality Andersen et al. (1991a, 1991b), Beadenkopf et al. (1963), Florey et al. (1969), Hook et al. (1977), Kuller et al. (1967), Paton (1957), Percy and Dolman (1978), Percy et al. (1981), Schoenberg and Powell (1968)]. The primary aim of much of this work has been to address the quality of available vital statistics data for epidemiologic purposes. Very little research has been directed towards the quality of these sources for end point identification and classification.

There is evidence to suggest that the cause of death as noted by the underlying cause of death on the death certificate is inaccurate in those deaths in which an autopsy was not performed, or the physician completing the death certificate did not use the results of an autopsy [Andersen et al. (1991a, 1991b), Beadenkopf et al. (1963), Florey et al. (1969), James et al. (1955), Kohn (1982), Paton (1957)]. Evidence also exists to support the claim that death certificate cause of death often differs from that recorded in hospital discharge records, or other clinical records [Alderson and Meade (1967), Florey et al. (1969), Gittlesohn and Senning (1979), Hook et al. (1977), Kuller et al. (1967), Percy et al. (1981), Rigdon (1981), Schoenberg and Powell (1968), Wingrave et al. (1981)]. Since the 1950's, the percentages of accuracy in cause of death, either using the autopsy or clinical record, has slowly improved. In the 1950's the accuracy ratings were between 44 and 73 percent [James et al. (1955), Paton (1957), Beadenkopt et al. (1963)]. In the 1960's, it ranged from approximately 60 percent to 79 percent [Alderson and Meade (1967), Florey et al. (1969), Kuller et al. (1967), Schoenberg and Powell (1968)]. In the 1970's, it ranged from 65 to 69 percent [Gittlesohn and Senning (1979), Percy et al. (1981), Hook et al. (1977)]. In the 1980's, accuracy ranged between 50 and 72 percent [Andersen et al. (1991a, 1991b), Wingrave et al. (1981)]. In an editorial, Glasser (1981) suggested that the accuracy of vital statistics is dependent on the performance requirement determined by the specific research. Kohn (1982), and Rigdon (1981) suggest that cause of death data obtained from the death certificate is of no value, even in some broad categories of mortality. Others recommend the cautious use of cause of death data obtained from death certificates in epidemiology for following general trends on broad disease classes. However use of these data to investigate specific causes of death is of limited value due to their quality [Alderson and Meade (1967), Gittlesohn and Senning (1979), Hook et al. (1977), James et al. (1955), Kuller et al. (1967), Percy et al. (1981)].

2. Classification Procedures in Large Multi-Center Research Programs

A few investigators have studied the quality of the cause of death as reported on death certificates as a potential method for classify end points for cause-specific mortality. Curb et al. (1983) studied the nosological coding of cause of death from one of the fourteen centers of the Hypertension Detection and Follow-up Program (HDFP). During this seven year study, 1973-1979, the death certificate for each death was obtained from each state's bureau of vital statistics. Using three highly trained nosologists, they found all three nosologists agreed on the three digit International Classification of Diseases, Adapted (ICDA) in 90.2 percent (691 out of 766) of the deaths and at least two of the three agreed in 99.7 percent (764 out of 766) of the death certificates examined. Looking at particular disease classifications, three out of three agreement was highest for neoplasms (97.8 percent) and lowest for myocardial infarction (88.0 percent). A 20 percent recoding (i.e., reclassifing) scheme was also instituted. The intranosologist agreement rates for three digit ICDA coding ranged from 94.8 percent to 96.1 percent.

Based on these results, Remington (1984) made some suggestions for future classification of cause-specific mortality end points in future large multi-center clinical trials. He observed that most death certificates are completed by physicians. They are highly trained clinically but often less than precise in the completing of these documents. These documents are then subjected to the nosologist for coding. He or she carefully applies defined and precise rules in the classification of the underlying cause of death. Although these nosological rules often appear to make little sense to the clinician, they have been shown to be highly reproducible. To take advantage of the expertise of both groups, Remington suggested that study physicians complete a *study death certificate* based on the clinical record using specific rules determined by the study investigators. These study death certificates would then be subjected to nosological classification. This recommendation was used in Studies of Left Ventricular Dysfunction (SOLVD), a National Heart, Lung, and Blood Institute (NHLBI) clinical trial which is currently in the field [SOLVD (1990)].

Recently, several large scale multi-center collaborative research programs with cause-specific mortality as one of the end points have tended to use a panel of clinicians to classify end points. For example, panels of cardiologists in a large scale cardiovascular research programs have evaluated the cardiovascular mortality classification [Beta Blocker Heart Attack Trial Research Group (1982), Hjermann et al. (1981), Jacobs et al. (1990), Kuller et al. (1980), Lipid Research Clinics Program (1984)]. Bangdiwala et al. (1989), and O'Connell et al. (1985) have investigated the efficacy of these mortality classification panels, comparing them to the information obtained entirely on the death certificate and coded by a trained nosologist.

Bangdiwala et al. (1989) examined the Lipid Research Clinics (LRC) Follow-up Study (FUS) of the North American Prevalence Study mortality classification procedure including the reproducibility of the nosological coding. This mortality classification procedure uses nosological coding of all death certificates. In addition, if there was any mention of cardiovascular disease, local clinic staff obtained hospital and physician records. These records were blinded, and sent independently to two cardiologists for classification. The definitions of cardiovascular deaths in this study was more specific than can be obtained using the ICDA codes alone. If the two cardiologists did not agree, the blinded records went to a full panel of cardiologists for adjudication. Using the 423 cardiovascular deaths of the first six contact years, they found an average agreement between the two independent cardiologists of 0.56 measured by the Kappa statistic. For individual pairs of cardiologists, the Kappa statistics ranged from 0.36 to 0.79. The LRC definition of coronary heart disease (CHD) requires information too specific to be contained on a death certificate. If the definitions of CHD are collapsed to definite CHD and suspect CHD death, the nosological coding and mortality classification panel had an agreement rating represented by a Kappa statistic of 0.58. This suggests only moderate agreement. All deaths from the first three contact years (365) were also resubmitted to the same nosologist for recoding. The Kappa statistic of the first and second coding was 0.96 representing extremely high agreement.

O'Connell et al. (1985) examined the LRC Coronary Primary Prevention Trial (CPPT) mortality classification procedures. The LRC CPPT used the same mortality classification scheme as the LRC FUS. They found a 95 and 100 percent agreement between nosological and mortality classification panel with ICDA codes 410-420, and 425-427 respectively. [ICDA codes 410-420 are the range for ischemic heart disease, including acute myocardial infarction, other acute and subacute forms of ischemic heart disease, chronic ischemic heart disease, angina pectoris, and asymptomatic ischemic heart disease; and for acute non-rheumatic pericarditis. ICDA codes 425-427 include cardiomyopathy, pulmonary heart disease, and symptomatic heart disease, such as congestive heart failure, left ventricular failure, and heart block.] If only the definite CHD deaths are considered, there is a 87 and 67 percent agreement. The nosological coding was not specific enough to differentiate between definite and suspect CHD deaths, in either the LRC Follow-up Study of the North American Prevalence Study or the LRC CPPT.

### C. Possible Approaches to the Automation of End Points

Several approaches to the automation of end point classification are possible. These include classical discriminant analysis, logistic regression, binary structured tree classification (also known as classification and regression trees and recursive partitioning analysis), and the development of an expert system.

#### 1. Classical Discriminant Analysis

The goal of discriminant analysis is to assign an observation on the basis of p explanatory variables $x = (x_1, x_2, \ldots, x_p)^T$, of unknown origin to one of two or more groups on the basis of the values of these explanatory observations. Under the assumptions of joint multivariate normality, equal covariance matrices, and known explanatory parameters x, a class of optimal classification rules is based on Fisher's linear discriminant function (LDF) [Anderson (1984)]. For the two class problem, this leads to the rule:

$$[x - (\mu_1 + \mu_2) / 2]^T \Sigma^{-1} (\mu_1 - \mu_2) > \log_e (q2/q1)$$

where $\mu_i$ the mean vector of the $i^{th}$ population,

$\Sigma$ the common covariance matrix, and,

$q_i$ the prior probability of the $i^{th}$ population.

Under the assumptions of joint normality and known explanatory variables **x**, but unequal covariance matrices, a class of optimal classification rules is based on the quadratic discriminant function (QDF) [Lachenbruch (1975)]. In standard analyses, replacing the mean and covariance matrices with observed sample mean and pooled covariance matrices yields the Wald-Anderson classification rule or sample LDF. This has been found to be relatively robust in a number of different situations. It can be expected to perform reasonably well, especially when the true discriminant function is at least approximately linear [Choi (1986)]. A further discussion of discriminant analysis in comparison to logistic discriminant analysis can be found in section I.C.2.

Kernel Density Estimation and Kth Nearest Neighbor Methods are two additional methods for classification and discrimination and are considered nonparametric in nature. This is because they make minimal assumptions regarding the form of the underlying distribution. They also have limitations common to both methods [Breiman et al. (1984)]. They are sensitive to the choice of the metric used to group observations which are similar together. There is not a simple way to handle categorical variables or missing values. They also provide little insight into the structure of the data.

## 2. Logistic Discriminant Analysis

Logistic regression was suggested by Walker and Duncan (1967) as an alternative to discriminant analysis for modeling the probability of an individual with a p-dimensional explanatory vector **x** developing a disease. They suggested this as what they call a *least squares approach*. This is, in fact, equivalent to estimation of the parameters by maximum likelihood, giving it desirable asymptotic properties. This approach makes no assumption regarding the distribution of **x**. It only makes the assumption that the appropriate model is the logistic, that is of the form:

$$\log_e (p/1\text{-}p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Because this function includes a wide variety of families of distributions, including mixtured continuous and categorical, this form is particularly appealing with data of this type. Therefore, at least theoretically, a logistic approach is more robust than the LDF [Choi (1986)]. Efron (1975) has shown for the special case of multivariate normality, the efficiency of logistic regression relative to the LDF is less than one. However, for mixed continuous, and categorical data, the sample logistic regression approach has been shown to out-perform the sample LDF, although often only slightly [Choi (1986), Gordon (1974), Halperin et al. (1971), Press and Wilson (1987), Schmitz et al. (1983)]. Much, if not all of the data of interest in clinical trials datasets are qualitative, and would be considered to be of the mixed continuous and categorical type. Therefore, the assumption of multivariate normality is often questionable.

3. Tree Structured Methodologies

Regression trees, also known as recursive partitioning analysis, is one of the newer data analytic tools for classification and regression. A tree methodology repeatedly stratifies the observations **x** into smaller and smaller subgroups based on values of **x** that *predict* or *classify* these observations. It does not depend on models which postulate relationships with unknown values between the predictive factors, such as would be true with either discriminant analysis, or logistic regression. A more detailed description of this method can be found in chapter III.

4. Expert Systems

Artificial intelligence is defined as the software (and development of such software) which allows computers to simulate humans and human problem solving. Expert systems is one area of artificial intelligence. Specifically, it is the area in which computer algorithms aid individuals analyzing problems and making decisions in areas where there is little or no expertise or experience. Expert system software differs from conventional programming in that the software processes data using rules-of-thumb and logical operators such as *and*, *or*, and *not*. It is highly interactive, and can be halted at any time to ask why a particular line of reasoning is being pursued. A more detailed description of expert systems can be found in chapter II.

## II. Artificial Intelligence and Expert Systems

### A. Background

Since the 1940's, there has been an intensive effort to develop software which would simulate humans and human problem solving. This work is usually defined as artificial intelligence. One area of artificial intelligence is the development of software which assists individuals in analyzing problems and making decisions in areas where they have little or no experience. This is known as knowledge-based expert systems, or simply expert systems. There are currently many available texts which review expert systems [Goodall (1985), Hayes-Roth et al. (1983), Harmon and King (1985), Jackson (1985), Jackson (1986), Rich (1983), Winston (1984)], in addition to a rapidly proliferating primary literature [Bobrow and Stefik (1986), Buchanan (1986), Davis (1986), Denning (1986), Duda and Shortliffe (1983), Friedman and Frank (1983), Lehner and Barth (1985), Liebowitz (1986), Michaelsen et al. (1985), O'Hare and Bell (1985), Schafer (1985), Shortliffe (1984), Thompson and Thompson (1985)]. Figure 2.1 displays the history and evolution of expert systems [Harmon and King (1985)]. Some of the applications which expert systems have been developed include: assisting managers with complex business planning and scheduling tasks; aiding physicians in diagnosing diseases specifically in areas of medicine: bacterial infections and meningitis [Evans et al. (1985)], dermatology [Vanker and VanStoecker (1984)], oncology [d'Anincourt (1986)], pulmonology [Aikins et al. (1983), Miller (1985)], internal medicine [First et al. (1985), Masarie et al. (1985), Miller et al. (1982), Miller et al. (1984)], and neurology [Jagannathan et al. (1985)]; aiding geologists in locating mineral deposits [Goodall (1985), Harmon and King (1985)]; in configuring complex computer hardware for VAX minicomputers [Goodall (1985), Harmon and King (1985)]; in cotton crop management [Lemmon (1986)]; and to aid mechanics in troubleshooting locomotive problems [Goodall (1985)]. They are also being developed for use in "intelligent tutoring systems" [Miller (1984)].

Feigenbaum, a well known researcher in the area of expert systems, more precisely defines expert systems as: "... an intelligent computer program that uses knowledge and inference procedure to solve problems that are difficult enough to require significant human expertise for their solution. Knowledge necessary to perform at such a level, plus the inference procedures used, can be thought of as a model of the expertise of the best practitioners of the field [Harmon and King (1985)]. The knowledge of an expert system consists of facts and heuristics. The 'facts' constitute a body of information that is widely shared, publicly available, and generally agreed upon by experts in a field. The 'heuristics' are mostly private, little-discussed rules of good judgment (rules of plausible reasoning, rules of good guessing) that characterize expert-level decision making in the field. The performance level of an expert system is primarily a function of the size and the quality of a knowledge base it possesses."

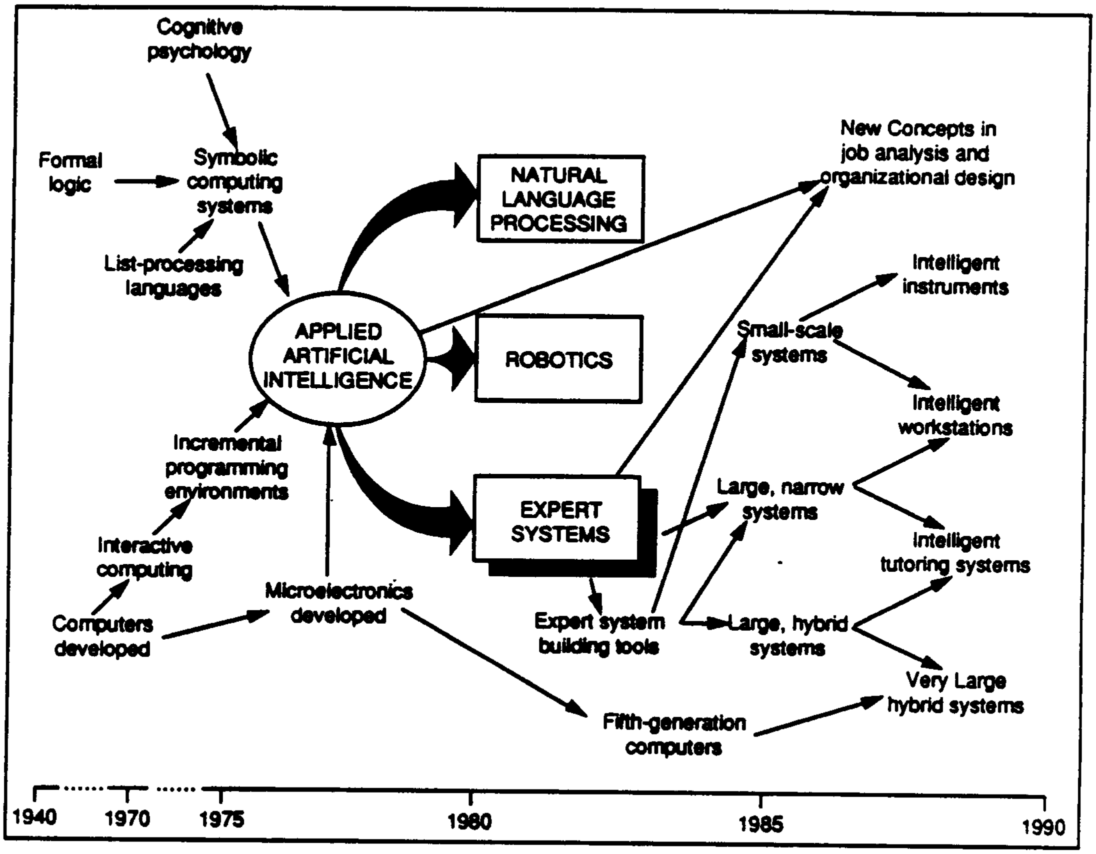


Figure 2.1: The evolution of expert systems [Harmon and King (1985)].

Expert systems differ in several aspects from conventional programming [Goodall(1985)]. Conventional programming has been used to create large data processing systems, capable of collecting and processing large volumes of data. They process these data by complex step-by-step procedures or algorithms. They are usually oriented toward numerical processing, are run in sequential, batch processing, and are maintained by programmers. If a nonprogrammer were to examine the code of a conventional program, it is unlikely anything useful could be learned. Expert systems process data using *heuristics* or rules-of-thumb and are oriented towards symbolic processing. Symbolic processing is the manipulation of logical operators such as *and*, *or*, and *not* rather than numerical. Expert systems are usually highly interactive. The user can halt the processing at any time and ask why a particular line of reasoning is being pursued or how a particular conclusion was reached. Expert systems, unlike conventional programs, are relatively easy to modify, the knowledge base is readable, and thus can and often are maintained by experts and not programmers.

Models of human problem solving using artificial intelligence are based on the principals in cognitive psychology [Harmon and King (1985)]. Knowledge can be represented in a variety of ways. One such way is to describe knowledge on a continuum, from no knowledge to compiled knowledge. Compiled knowledge is readily usable for problem solving. There are two ways to compile knowledge. The first is from domain-independent definition such as formal study. Examples include learning from school, lectures, and reading a textbook. This results in a set of definitions, axioms and laws. This type of knowledge can be described as formal theories but may not be applied in any practical way. It can also be thought of as deep knowledge. The second way to compile knowledge is using domain-dependent facts such as by learning from a mentor, or by experience. In this case, rules-of-thumb or *heuristics* are used to focus on particular facets of a problem and to solve these problems based on previous experience. This can be thought of as surface knowledge. Surface knowledge which is well organized can be easily used to solve problems within the area of the domain-dependent facts, but is usually limited in its application outside a narrow domain.

There are many features of expert systems which make them valuable in many commercial and research applications [Goodall (1985)]. In business applications, they can do a particular job with fewer staff, save money by saving time and help make decisions to work on a task with the least expensive equipment. Expert systems have been described as performing superior to humans in many instances because they make few errors (given the knowledge bases are accurate), do not become tired or bored, will not overlook a particular situation, can respond rapidly, can handle large volumes of data, can function in a hostile environment, and do not sleep or become sick. They also have some qualities which make them superior to conventional computer programs in some situations. They can perform previously unprogrammable tasks, can handle data with uncertainty or data which are incomplete, can make their knowledge explicit, that is display their line of reasoning or logic at any instance for human operators.

## B. Expert System Components

Expert systems can be divided into two basic pieces [Jackson(1986)]. The knowledge base and working memory comprise one part of the system. The knowledge base contains the facts and rules that make up the expert's knowledge. These heuristics must be obtained from the expert, by a knowledge engineer, the individual developing the expert systems. This can be and often is a formidable task. The inference engine constitutes a second part. The inference engine has two main tasks: it examines existing facts and rules attempting to add new facts when possible. It also decides the order in which inferences are made. Thus the inference engine conducts the consultation with the user.

1. Knowledge Bases

Facts and relationships which comprise a knowledge base can be represented in a variety of ways [Harmon and King(1985)]. The most general and also one of the oldest is the semantic network. In a semantic network, a collection of objects and descriptors called *nodes* are connected together or related by arcs called *links*. Another common way to represent factual information is known as object-attribute-value (O-A-V) triplets. In this scheme, a special case of a semantic network, objects may be conceptual entities such as an event or may be physical entities. Attributes are general characteristics or properties associated with the objects. The value is the specific nature of an attribute in a particular situation. For example, one such O-A-V triplet might be:

| | |
|---|---|
| patient x | (object) |
| taking vasodilators | (attribute) |
| yes | (value). |

Heuristic rules are used to represent relationships. Each rule contains two parts: an expression(s) or *if* clause(s) and a conclusion or *then* clause. Certainty factors (cf) can be used to represent the confidence of a particular piece of evidence. They be attached to either O-A-V triplets or rules. For example, a rule from MYCIN, an expert system developed at Sanford University for aiding physicians in the diagnosis and treatment of bacteremia and meningitis, can be represented as follows:

| | Attribute | Object | Value | (cf) |
|---|---|---|---|---|
| If | Site | Culture | Blood | |
| | Morphology | Organism | Rod | |
| | Gram stain | Organism | Gramneg | |
| | Compromised | Patient | True | |
| Then | Identity | Organism | Pseudomonas aeruginosa | (.6) |

A variable rule in the same knowledge base might have the same objects and attributes and differing values could have a different then clause. Frames are another method for representing facts and relationships. A frame is a description of an object that contains slots for all the information associated with the object. Slots like attributes may store values. They may also contain default values, pointers to other frames, sets of rules or procedures by which values may be obtained. They are richer representations of knowledge but more complex and difficult to develop then O-A-V/rule systems.

2. Inference Engines

Inference engines perform two major tasks: inference and control [Harmon and King(1985)]. In simple expert systems, inference is performed with the strict application of logical rules to derive new facts from existing rules and known facts. More complex expert systems also incorporate inexact reasoning, that is, the ability to use certainty factors, both in premises and in rules. Additionally, more complex expert systems' contain the ability of resolution to discover if a new fact is valid, given a set of logical statements. The control portion of the inference engine decides with which rule to begin reasoning, and how conflicts will be resolved. The type of inference engine can also be described by its organization in three areas: the type of chaining, the search strategy, and the type of reasoning. First, the inference engine can be started at the goal and worked backward through subgoals to choose an answer (backward chaining) or begin working with rules until a conclusion is asserted, continuing onto the next step (forward chaining). Control of forward chaining systems is somewhat more complex than that of backward chainers. The shape of the search space (associations and rules) determines whether forward or backward chaining is more efficient. A second distinguishing feature is the search strategy. In a depth-first search, the engine takes every opportunity to produce a subgoal. A breadth-first search engine sweeps across all premises in a rule before digging for greater detail. Breadth-first will be more efficient if one rule succeeds and the goal attribute value is obtained. Most expert systems employ depth-first searching. A final distinction among inference engines is the support of monotonic or nonmonotonic reasoning. In monotonic reasoning, all values concluded for an attribute remain true to the duration of the consultation; in a nonmonotonic reasoning system, facts that are true may be retracted. Most expert systems today support monotonic reasoning and may allow for carefully controlled types of nonmonotonic reasoning.

### C. Expert Systems and Clinical Trials - Current Status

Until recently, statisticians were unaware of the work in expert systems. Expert systems, while gaining greater support in the medical community, has some severe limitations. Heuristic rules-of-thumb are usually developed by a knowledge engineer interviewing the expert(s), without the analysis of data. Therefore, any biases, misconceptions, inaccuracies, or misinterpretations, of the expert, or translated by the knowledge engineer into the knowledge base will be present in the system developed. Data which may not exactly fit the structure required by the expert systems may give misleading results. This may also occur while providing no indication of problems. The use of certainty factors has been criticized as implying a probabilistic value to reasoning which is inexact, heuristically derived and based on observations which may be subjective by its very nature [Speigelhalter and Knill-Jones (1984), Spiegelhalter (1987), Lindley (1987), Shafer (1987)].

Expert systems have been used in clinical medicine since the early 1970's. The expert system MYCIN has been well documented for its' use in the diagnosis and recommendation of therapy of bacteremia and meningitis. Since its' development, applications of expert systems in clinical medicine include: use in internal medicine (INTERNIST), pulmonary medicine (PUFF), oncology (ONCOVIN), dermatology, and in reading and coding EEGs. Other than their use in oncologic study protocol management (OPAL) [d'Anincourt (1986)], most reports of expert systems in the medical literature is of their use as a diagnostic aid [Adams et al. (1986), deDombal et al. (1972), Evans et al. (1985), First et al. (1985), Jagannathan et al. (1982), Knill-Jones (1987), Masarie et al. (1985), Miller et al. (1984), Miller et al. (1982), Miller (1984), Spiegelhalter (1983), Speigelhalter and Knill-Jones (1984), Stern et al. (1974), Vanker and Van Stoecker (1984)]

Speigelhalter and Knill-Jones (1984) have developed a type of hybrid statistical-expert system for use as a clinical decision support tool for general practitioners requiring consultation by a gastroenterologist. This system was developed in response to interest in an expert systems for the general practitioner while maintaining a probabilistic basis. This system uses branching logic which is entirely algorithmic, and uses the *knowledge-based* reasoning of an expert system. It also can handle missing data, and inexact reasoning, that is, predictive uncertainty. It was developed from a probabilistic base, collecting a large amount of data, and developing a scoring scheme based on a logistic model. It was from this base that the rules were obtained for the knowledge base. This system also directly interviewed the patients for data in the form of a medical history. It provides the clinician with output displaying recommendations and the reasoning for these recommendations. Hutchinson et al. (1991) has also developed an expert system for the identification and classification of adverse experiences in pharmaceutically based clinical trials. Speigelhalter and Knill-Jones (1984) are also collaborating on the development of a hybrid system for the interpretation of blood sample data in a leukemic blood laboratory.

## III. Classification and Regression Trees

### A. Background

Tree structured methodologies are one of the more recent and innovative data analytic tools available today for classification and regression analyses. Tree structured classifiers, more specifically ***binary tree structured classifiers***, are constructed by repeatedly splitting (or partitioning) the sample space X into more homogeneous descendant subsets [Breiman et al. (1984), p. 20-21]. The following example best describes this method:

> When a patient is first seen in the emergency room of Yale-New Haven Hospital with symptoms of chest pain, normally many measurements are performed. A study was performed to develop a decision tree to classify or "predict" which of these chest pain patients are likely to be suffering from a myocardial infarction [Cook and Goldman (1984)].

Figure 3.1 was the tree structured classification rule constructed on the basis of these data. Many other examples exist in the medical literature of the use of this method; these include: Altman et al. (1986), Brand et al. (1982), Diehr et al. (1981), Dillman and Koziol (1983), Feinstein et al. (1975), Feinstein et al. (1977), Fleiss et al. (1977), Gammerman and Thatcher (1991), Giampaolo et al. (1991), Gilpin et al. (1983), Goldman et al. (1982), Goldman et al. (1988), Goldman et al. (1988), Harrell et al. (1985), Henning et al. (1976), Honigfeld et al. (1969), Klein et al. (1972), MacEntee et al. (1991), Meyers et al. (1978), Spitzer and Endicott (1968, 1969), Stewart and Stamm (1991), Streitberg et al. (1989), Sutherland et al. (1980), Wheeless et al. (1986).

Tree structured methodologies were originally developed by social scientists in the early 1960's. Leo Breiman of the University of California, Berkeley and Jerome Friedman of Stanford University began the most recent work on trees in the early 1970's. Charles Stone of the University of California, Berkeley and Richard Olshen of the University of California, San Diego joined this group, and both contributed substantially to its development [Breiman et al. (1984), p. ix]. The CART (an acronym for Classification and Regression Trees) software is a powerful and flexible piece of software written by Jerome Friedman. This software implements this technique, and automates a substantial portion of the methodology.

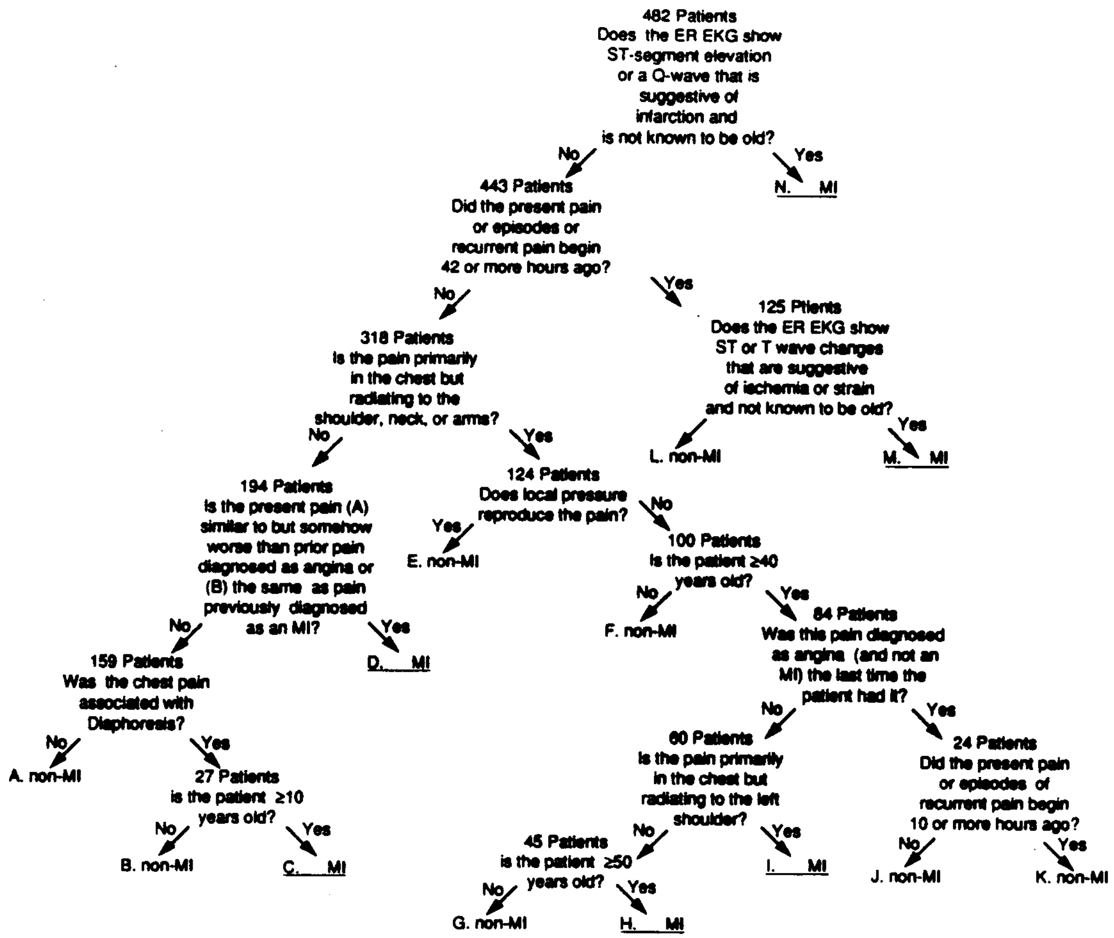


Figure 3.1: Computer-derived decision tree for the classification of patients with acute chest pain [Cook and Goldman (1984)].

B. Data Requirements and Analysis Assumptions

Breiman et al. (1984) use the following terminology in discussing tree structured methodologies. The original unsplit dataset in the sample space is called the root node. Repeatedly splitting the sample space into smaller and smaller subsets is to *split nodes*. A subset or node which has been identified as homogeneous and thus not requiring further splitting is called a terminal node. The entire process of splitting nodes to construct a classification rule (or regression rule) is to grow a tree. There are three important issues involved in the growing of trees. The first consideration is how splits are formulated within the sample space. A second important issue is how far to continue splitting the nodes, i.e., when to stop splitting and declare the node terminal. The last issue in growing trees is defining the assignment of a class to members of the terminal node. This is done by simply assigning the observations in that terminal node to the class with the highest proportion. The first two issues are discussed in the following two sections.

Each split is made at nonterminal nodes into two descendant nodes. In order to split each of the nodes, the data must be able to be standardized into a set of binary questions, based on the sample space. These binary questions can be based on either a single variable, or a combination of variables. Combinations of variables can be either linear or Boolean. Therefore, the data must have a structure of either categorical, ordinal, or continuous. Standardized binary questions would have the form $x_m$ exist in $\{R\}$ (for categorical variables) or $x_m \leq c$ (for ordered or continuous variables). There are no other data requirements or analysis assumptions.

## C. Growing Trees and Splitting Rules

How splits of nodes occur is one of the important issues in the growing of trees. Breiman et al. (1984) [p. 102] use a class of splitting rules, for the two-class problem, as a simple function of node impurity with the form:

$$i(t) = p(1|t)p(2|t)$$

where $i(t)$ is the measure of node impurity, and $p(a|t)$ is the proportion of class a in node t, $a = 1,2$. For pure nodes, $i(t)$ would equal zero, while for a node which was completely heterogeneous (i.e., a 50:50 mixture), $i(t)$ would have a maximum value of 0.25. Several different splits could be constructed. For example, one split may produce two nodes, one pure and one somewhat homogeneous. An alternative split may produce two nodes, neither pure, but both having improved purity and requiring further splits. Breiman et al. (1984) [p. 25] use a function which reduces misclassification costs in the form:

$$i(t) - p_L i(t_L) - p_R i(t_R)$$

where $i(t)$ is defined as above and $p_J i(t_J)$ is the proportion of vectors **x**, splitting left(right) times node impurity in the left(right) node for $J$ = left, right.

The multi-class problem requires some additional strategies. Breiman et al. (1984) [p. 103-106] use two different criteria of node impurity for the multiclass problem. The *Gini diversity index* as the form:

$$i(t) = \sum_{j \neq i} p(j|t)p(i|t) \ .$$

This reduces in the two class problem to $i(t)=2p(1|t)p(2|t)$, which is equivalent to the two-class criteria times the constant two. A second method is the *Twoing Criterion*. This criteria combines the multiple classes into two supersets of classes, and continues computations based on a two-class problem.

With regard to choices between the two splitting rules, Breiman et al. (1984) [p. 111] states:

> "There are usually only slight differences between the Gini and twoing trees. In balance, comparing the two on many data sets, where they differ, the Gini splits generally appear to be better. In fact, one can give examples of two candidate splits of a node, one of which is clearly superior to another in terms of producing pure descendant nodes, in which twoing (but not Gini) selects the poorer split. For these reasons, we usually prefer the use of the Gini criterion."

Two additional parameters are available in the tree structured methods in minimize misclassification. The first is the costs of misclassification. The costs of misclassifying an observation may be different for different classes. Incorporating the costs of misclassification into the estimated expected cost is done by altering the measure of node impurity. Breiman et al. (1984) [p. 113] extend the Gini diversity index to the following estimated expected misclassification cost:

$$i(t) = \sum_{j \neq i} C(i|j)p(i|t)p(j|t)$$

where $C(i|j)$ is the variable misclassification costs of misclassifying a class j observation as class i, and others are as previously defined. In the two class problem, this reduces to:

$$(C(2|1) + C(1|2))p(1|2)p(2|1).$$

The second parameter which can be altered is the prior. Using equal priors tends to distribute the misclassification equally across all the classes; putting a larger prior on a class will distribute misclassifications away from that class [Breiman (1984), p. 114]. Therefore, if the data on which the tree is grown has unequal numbers in each of the classes, the proportion of misclassifications will tend not to be proportionally equal to the sample size within each class. Using priors help equalize the proportion of misclassifications across classes. Breiman et al. (1984) [p. 114] explain that this can be thought of as altering the costs of misclassification. If one class has half as many observations as a second class, and equal priors are used, there would tend to be twice as many misclassifications. In this case, the costs of misclassification for the first class would be twice that of the second. Breiman et al. (1984) [p. 114] implement use of priors within the tree structured methods by altering the costs to a unit priors and grow the trees with the altered misclassification costs.

As was mentioned in the previous section, splits can be made on categorical, ordinal, or continuous variables. They can be single variable splits based on one variable, or combination variable split based on a combination of variables. Combination variable splits can be made on ordinal or continuous data, thus a linear combination. With some data, classes may be separated by hyperplanes not perpendicular to coordinate axes of the sample space [Breiman et al. (1984), p. 39, 132-135]. It is in these situations where linear combination variable splits may grow a tree with less misclassification. A learn search algorithm including a backward elimination procedure is used in the CART software [Breiman et al. (1984), p. 133, 171-173].

Categorical variables can also be combined into Boolean combinations. This type of combination splits has been useful in medical applications [Breiman et al. (1984), p. 39, 136-138]. An example split of this might be: Is lab test A positive and the subject has symptom B. A direct maximization procedure to search on Boolean combinations has not been found, but a stepwise procedure has been used [Breiman et al.(1984), p. 40].

D. Accuracy

Early in the development of tree structured classifiers, large trees were grown with optimistic high estimates of accuracy [Breiman et al. (1984), p. 59-60]. Any tree can be grown so that each terminal node is pure (e.g., each terminal node contains one observation). However not unlike overparameterization of other statistical models, it was found that the predictive accuracy of these classifiers was poor.

Breiman et al. (1984) [p. 61-62] considered four methods for evaluating accuracy. The resubstitution estimate simply computes an error rate based on the same data used to grow the tree. Not surprisingly, using the same data to construct the classifier gave overly optimistic estimates of accuracy [Breiman et al. (1984), p. 60]. Another method which was considered was a bootstrap estimate. For theoretical reasons, this method also did not give accurate estimates of error. The final two methods worked well and are implemented within the CART software. In the test sample method, the dataset is divided into two groups. The first group is used to grow the tree, and the second group is used to estimate error [Breiman et al. (1984), p. 73-75]. The other method used is the V-fold cross-validation method [Breiman et al. (1984), p. 75-78]. Using this method, the dataset is divided into V subsets of the sample of approximately equal sample size. V auxiliary trees grown on the V fractions (of size (V-1)/V ) of the data and each tested on the remaining (1/V) fraction. These are used to estimate error in the main tree, which is grown using all the data.

Accuracy is closely related to the issue of when to stop splitting nodes. Breiman et al. (1984) [p. 78] found it best to grow sufficiently large trees, and then prune them to five "honest" estimates of error. This is implemented in the following manner. Continuing the splitting of notes until at least one of three conditions exist: (1) the node is pure, (2) the members contain identical measurement vectors **x**, or (3) there is some minimum number of observations, often set at five. Once a sufficiently large tree is grown, prune branches of the tree (i.e., nodes and descendant notes) such that this pruning minimizes the misclassification. This pruning is progressively done on each previously pruned tree until only the root node is left.

Because the absolute misclassification costs always increase with smaller trees (i.e., the more splits, the lower the resubstitution estimate of misclassification costs), pruning is done with a method known as minimal cost-complexity pruning [Breiman et al. (1984), p. 66-72]. The cost-complexity measure $R_\alpha(T)$ is defined as:

$$R_\alpha(T) = R(T) + \alpha\ |\tilde{T}|$$

where: $R(T)$ = misclassification cost of the tree

$\alpha$ = complexity cost per terminal node, and

$|\tilde{T}|$ = number of terminal nodes.

The misclassification cost of the tree is the sum over all terminal nodes of the expected misclassification costs given terminal *t*, times the probability of falling into terminal node $t$. $\alpha$ is a value $\geq 0$. For each value of $\alpha$ there is a value which minimized $R_\alpha(T)$ and is unique [Breiman et al. (1984), p. 68-69]. Furthermore, for increasing values $\alpha$, a sequence of tree is developed such that each tree is a subset of the previous tree.

Once a succession of trees result from the progressive pruning of a sufficiently large tree, the problem is to select the best size tree. Using either the test sample or V-fold cross-validation methods for error estimation, means and standard errors of the misclassification costs can be estimated assuming a simple binomial model. Breiman et al. (1984) [p. 78, 303-311] have derived expression for these statistics. If the estimates of the misclassification costs are plotted against the number of terminal nodes from each of the trees, there is commonly a rapid decrease with small trees followed by a long, flat area, and a gradual rise for the largest tree [Breiman et al. (1984), p. 78]. Because of the instability of the minimum value for misclassification costs, Breiman et al. (1984) [p. 78] suggest a one standard error rule. That is, to choose the simplest tree (i.e., the one with the fewest number of terminal nodes) within one standard error of the tree with minimum misclassification costs.

E. Missing Data

Breiman et al. (1984) [p. 40, 140-143] have incorporated a very effective and ingenious method of dealing with missing data. This is done through the use of *surrogate splits*. During the process of identifying the best split, CART keeps track of all potential splits. A surrogate split is a split whose result is very similar to the original best split. Specifically, this rule is obtained by maximizing the sum of the probabilities of cases set to the each of the two descendant nodes by the best and surrogate splits [Breiman et al. (1984), p. 140-142]. This is similar to counting the number of concordant pairs of the surrogate and original splits. A predictive measure of the advantage of a surrogate split is the relative reduction in error of using the surrogate split over assigning the split, left or right, to on the direction, left or right, that the maximum number of observations were directed [Breiman et al. (1984), p. 142]. Surrogate splits are only useful if the split based on the surrogate is highly correlated to the split based on the original variable.

By using this feature of handling missing data, one can make maximumal use of the data. In contrast to many other discriminant functions, observations are not casewise deleted if there are missing data on some of the variables. When the tree is grown, observations falling into a node which is split based on a missing value is split based on the variable which most closely resembles the splitting variable. If that is also missing, the next most highly correlated variable is used in the split, and so on.

Surrogate splits have two other applications [Breiman et al. (1984), p. 41, 146-148]. The first is in variable ranking. Variable ranking is a relative measure of importance of the variable in the classification rule. It is defined as the relative magnitude of all variables of interest of the measure of importance, and is standardized to a range of 0 (low) to 100 (high). The measure of importance is the sum over all nodes of the decrease in impurity by using the surrogate variable as the splitting variable instead of the best splitting variable.

The second use of surrogate splits is in the identification of masking [Breiman et al. (1984), p. 41, 149]. Variables of high importance but not used as splitting variables within the tree are considered masked.

F. Advantages, Disadvantages, and Potential Difficulties

Some difficulties can occur when trees are grown on data where multicollinearity or other complex associations exist between class membership and potential predictive (i.e., splitting) variables [Breiman et al. (1984), p. 156]. When this happens, trees grown on newer data or samples from the original dataset may have very different structures. Breiman et al. (1984) [p. 158] find that in their experience, misclassification ratios remain fairly stable even when tree structures are not.

Often when these complex relationships exist, the use of combinations splits (either linear or Boolean), or the elimination of certain variables from the dataset can improve the misclassifications. Breiman et al. (1984) [p. 160-162] suggest that the growing of exploratory trees, either before or after the primary tree is grown. This can be useful in understanding the data. However, they also recommend to avoid growing too many exploratory trees, as this "data dredging" can bias tree growing.

Tree structured methods are clearly a product of improved computational resources. However, growing of trees involves a large number of sorts and optimizations [Breiman et al. (1984), p. 163]. Particularly with large datasets, computational resources can be strained. To minimize this problem, Breiman et al. (1984) [p. 163] have implemented the use of subsampling for the growing of trees in the CART software. Users can specify upper nodal splitting sample size. For any node to be split, from the root node onward which has a sample size greater then this upper limit, a random sample is selected to determine the split. Once the split is determined, the entire nodal population is used to split the node into it's descendant nodes. If an upper nodal splitting sample size is specified, this sampling process continues until the nodal sample sizes are below this upper limit.

Tree structured methods have some clear advantages and disadvantages over other forms of statistical (parametric and nonparametric) analyses [Breiman et al. (1984), p. 55-57; Cook and Goldman (1984)]. One advantage of this method is the calculation of a graphical picture which can be easily understood by statistically naive researchers. This allows researchers to investigate and understand the predictive relationships between important variables in the classification of observations. Tree structured methods can be applied to any data structure by using a set of binary questions or conditions, making no assumptions regarding the underlying distributions. The final classification rule is a simple form, can be compactly stored and efficiently used to classify new data (for example to develop the classification rules of an expert system). It is designed to identify synergistic interactions among factors and may also be used to identify nonlinear relationships within the classes. The software CART, automatically grows a classification tree, and also provides estimates of misclassifications.

Despite these attractive features, there are some disadvantages. Currently only the CART software can grow classification trees. It is not widely distributed outside of universities, in part due to it's cost. As node splitting continues, the number of observations in each node decreases. This increases the potential to miss additional predictive factors during the later stages of node splitting. Without proper pruning of the tree, it may increase the multiple comparisons problem because it automatically does a stepwise variable selection and reduction process. (Pruning is automatically done within CART, and thus the lack of pruning should not be considered a major problem.) Lastly unless combination splits are used, there is a loss of information because the set of questions are binary and do not use all the information with continuous data.

G. CART and Clinical Trials - Current Status

CART and other earlier versions of recursive partitioning method have been reported in the medical literature in a wide variety of specialities. Meyers et al. (1978) have used CART for analyzing meningitis in children. Fleiss et al. (1977) reported comparing CART to using other computer diagnosis aids in psychiatry. Feinstein et al. (1975) reported using it to aid in the staging of rectal cancer, and Feinstein et al. (1977) used it similarly in cancer of the larynx. Gilpin et al. (1983), Henning et al. (1976), Goldman et al. (1982), Goldman et al. (1988), and Cook and Goldman (1984), reported using CART in diagnosing acute myocardial infarctions. Stewart and Stamm (1991) and MacEntee et al. (1991) have reported it's use in with dental data and compared it to logistic discriminant analysis. Harrell et al. (1985) have compared CART to other discriminant methods with oncology data. Wheeless et al. (1986) have used it as a "prescreening alarm" for gynecologic specimens in pathology. Altman et al. (1986) have used in rheumatalogy to diagnosis osteoarthritis of the knee, and Streitberg et al. (1989) have reported using CART as part of an analysis of circadian blood pressures in hypertensive clinical trials.

## IV. Logistic Discriminant Analysis

Logistic regression is one statistical method in a larger class of methods, specifically log linear models. Walker and Duncan (1967) first suggested logistic regression as an alternative to linear discriminant function (LDF) without the restrictive multivariate Gaussian framework. They gave a maximum conditional likelihood solution based on the product binomial logistic model. Theil (1969) soon extended this model to allow multinomial data.

Anderson (1972) noted the logistic formulation can result from a wide variety of underlying assumptions about the explanatory variables. In particular, it can result from multivariate normally distributed with equal covariance matrices, explanatory variables which are independent and binomial, or some mixture of multivariate normal and binomial. Specifically, if D is a multinomial variable representing r classifications, then the conditional probability

$$pr(D=d \mid x=x_1, x_2, \cdots, x_p) = [1 + \exp(-(\beta_o+\beta_1 X_1+\beta_2 X_2+ \cdots + \beta_p X_p))]^{-1}$$

which is equivalent to the logit of $p_j$, the probability of being classified into the jth disease category, that is:

$$\log_e(p_j/1-p_j)=\beta_{oj}+\beta_{1j}X_{1j}+ \cdots + \beta_{pj}X_{pj} .$$

This model is linear in the logit, and is fit by iterative methods. As these estimates are well known to be maximum likelihood estimates (MLE), the parameters have desirable qualities. Halperin et al. (1971) carefully and extensively proved the results on consistency. MLEs are well-known to be functions of sufficient statistics. Imrey et al. (1981, 1982), Kleinbaum et al. (1982), and Koch et al. (1982) have written excellent reviews of logistic regression, discussing both the theoretical and practical aspects. These provide an in-depth discussion of this method.

Halperin et al. (1971) studied the differences between the LDF and logistic regression for the estimation of parameters. Citing both theoretical, and empirical justifications based on the Framingham 12 year incidence data for CHD, and CHD death or myocardial infarction, they suggest using logistic regression over the LDF if the linear model holds but the normality assumption is violated. They note this may not be practical in instances where computing facilities are limited. The time required for compilation and execution of these two methods were 1.3 to 2 times greater for the maximum likelihood (logistic regression) method. Another benefit is that the expected number of deaths will be equal to the observed number of deaths which is not true for the LDF method [Halperin et al. (1971), Gordon (1974), Press and Wilson (1987)]. Halperin et al. (1971) noted empirically that the expected numbers of deaths per decile of risk was also similar to the observed numbers of deaths.

Schmitz et al. (1983) in a simulation study compared the performance of four classification procedures for mixtures of continuous and discrete variables: Fisher's LDF, logistic discrimination, QDF, and a kernel method. Using four different discrimination evaluation methods to compare these discrimination procedures, they concluded that the LDF and logistic discrimination have an almost similar performance; but recommend using the LDF because the logistic method failed to produce a classification rule in some situations. They also concluded that the kernel method always did at least as well as QDF, sometimes substantially better. For discrimination procedures with mixed continuous and discrete variables, they recommend choosing between LDF and kernel methods based on the equality/inequality of the covariance matrices.

Press and Wilson (1987) have noted several additional arguments which suggest the use of logistic regression instead of a LDF analysis. Use of discriminant function estimates can tend to hide what they define as troublesome cases within some data, by not providing danger signals. For example given an observed dataset:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $z_i$ | -4 | -3 | -2 | -1 | 1 | 2 | 3 | 4 |
| $y_i$ | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |

a perfect fit may be obtained by $y_i = 0$ for $z \leq -1$, $y_i = 1$ for $z \geq +1$. For $-1 < z < +1$, there is no unique solution, although the sample LDF does not provide an indicator. They also note that the estimates of the logistic model are MLEs, while the discriminant function estimators are not MLEs. By the Rao-Blackwell theorem, a smaller mean squared error of the estimated parameters can be achieved using the logistic model.

Logistic discrimination has been frequently used in cardiology with favorable results. These reports have reported accuracy comparable to the results of logistic discriminant analysis in this research. Logistic discriminant analysis has been primarily used in the area of treadmill exercise testing. One recent non-exercise logistic discriminant analysis was reported by Chivot et al. (1990). They selected 74 coronary bypass patients and 78 control patients and were able to correctly classify these patients 86.2 percent of the time. They validated their discriminant function on an additional 40 patients and correctly classified 87.5 percent of these patients. They did not compare logistic discriminant function to any other method. Kansal et al. (1983) used a logistic discriminant function to classify patients with coronary artery disease based on a treadmill exercise test. They were able to develop and validate logistic discriminant function for men using a five or four variable model, and for women using a four or two variable model. They achieved accuracies of 80 to 81 percent in men and 74 to 77 percent in women. These were substantial improvement over previously used prediction using ST-T segment changes alone. Hung et al. (1985) substantiated and expanded these results in men with both coronary artery and multivessel disease. They used models based on clinical, fluoroscopy, exercise, and exercise thallium scans for each of the two diseases in men with discriminant function of two to eight variables. Using clinical variables alone, they found overall accuracies of 71 to 78 percent, while using all types of variables, some models had overall accuracies of 84 to 91 percent. As a general rule, models with more variables did not necessarily result in greater accuracy; higher accuracies were more common in coronary artery disease then multivessel disease. One model with eight variables in multivessel disease had an overall accuracy of 84 percent, while model based on the same type of data with only four variables in coronary artery disease had a 90 percent accuracy. Robert et al. (1991) developed a more specific logistic discriminant function for exercise testing in women. She was able to develop and validate a model which achieved a 70 percent sensitivity and a 93 percent specificity using a three variable model. A recent non-exercise testing logistic discriminant function use was reported by Chivot et al. (1990). In their study, a logistic model was developed and validated using lipids, apolipoproteins and apolipoproteins fractions in a population of coronary bypass patients.

## V. Cardiovascular Mortality Classification from the Lipid Research Clinics (LRC) Program Coronary Primary Prevention Trial (CPPT) - An Example

### A. The LRC CPPT

The Coronary Primary Prevention Trial (CPPT) of the Lipid Research Clinics Program (LRC) was a clinical trial sponsored by the National Heart, Lung and Blood Institute (NHLBI) [Lipid Clinics Research Program (1979, 1984)]. This trial was a multi-center, randomized, double-blind study in 3,806 middle-aged men. The primary objective of the trial was to test the hypothesis that long-term reduction of serum cholesterol in hypercholesterolemic men initially free of clinical coronary heart disease (CHD) will lead to a lower incidence of CHD. The treatment group received cholestyramine, which lowers serum cholesterol by sequestering bile acid; the placebo group received a placebo. The average follow-up was 7.4 years. The primary end point of this trial was definite CHD death or definite nonfatal myocardial infarction. Other fatal end points were suspect CHD death and total mortality. Other non-fatal events included: suspect myocardial infarction, angina pectoris, intermittent claudication, arterial peripheral vascular disease, atherothrombotic brain infarction, transient cerebral ischemic attack, and abnormal exercise electrocardiogram. A Cardiovascular Endpoints Committee, composed of a cardiologist from each site and the central electrocardiographic coding center, was responsible for verifying and classifying end points which occurred during this trial.

When a possible end point was identified, all available clinic records and other information were reviewed by the clinic cardiologist. If he or she and other clinic physicians agreed an end point has occurred, the Central Patient Registry and Coordinating Center (CPR) was notified, and the appropriate data collection forms were completed. After identification, each end point was verified using the following procedure. The CPR collected all available documentation except drug allocation and any laboratory values which might unblind the patient or his treatment randomization. These data are reviewed by two independent members of the Cardiovascular Endpoints Committee. No individual reviewed cases from his/her site. If the clinic cardiologist and the two independent cardiologists review agreed, the case was considered verified. If there was any disagreement between the clinic and the two reviewers, or any uncertainty among the two reviewers, the case was presented to the entire Cardiovascular Endpoints Committee for adjudication.

An objective algorithm was to be used in the classification process. The Cardiovascular Endpoints Committee was empowered with redefining, if necessary, the end points. If an algorithm was changed, all relevant previously classified cases were to be reviewed again and reclassified according to the new algorithm. Because any change to an algorithm was a protocol change, this was subject to approval by the Intervention Committee, the LRC Directorate and the Safety Data Monitoring Board (SDMB).

These data were used to evaluate the three methods for automating the fatal cardiovascular end point classification. To use each of these methods, the data had to have several characteristics. The data had to have variables from at least one review (record) by a single independent cardiologist (i.e., more than just the clinic cardiologist and the final committee classification). Of the 139 fatal end points in the LRC CPPT, 136 had a review by at least one independent reviewing cardiologists. To be useful for the three automated methods, an observation should not have missing values for all of the variables. Of the 136 records, 15 had missing values for all the relevant variables used in the classification algorithms or discriminant functions, although all 136 observations were used in this dissertation. For the nosologist, only a death certificate was necessary. The death certificate data were not available for 2 of the 136 cases, therefore the nosologist's classification was available for 134 of the 136 available fatal end points.

Missing data were a problem for both the logistic and expert system methods. Only the CART methodology handles missing data, using surrogate splitting variables, when they are available. Because of the way the data were collected, there was a large amount of missing data. Therefore, this issue could not be ignored. Since the LRC algorithm is based on the existence or presence of various data, each criteria, characteristic, or question was dichotimized into the *presence/yes* versus *absence/no/unknown/missing* responses. An important point is that the LRC classification was not considered to be *truth* for the purposes of this research. The objective of this study was to simulate the LRC classification, not necessarily to achieve a *correct* classification of these end points, or to validate the LRC classifications.

The LRC algorithm used for classification of fatal events is shown in Table 5.1. This is the algorithm which was used by the LRC clinic and death committee cardiologists as a reference for classifying fatal end points.

Table 5.1
LRC Coronary Primary Prevention Trial
Primary Fatal End Point Definition

I. Definite Atherosclerotic coronary heart disease death - either or both of the following categories:
   A. Death certificate with consistent underlying or immediate cause plus either of the following:
      1. Preterminal hospitalization with definite or suspect myocardial infarction (see below).
      2. Previous definite angina or suspect or definite myocardial infarction when no cause other than atherosclerotic coronary heart disease could be ascribed as the cause of death.
   B. Sudden and unexpected death (requires all three characteristics):
      1. Deaths occurring within one hour after the onset of severe symptoms or having last been seen without them.
      2. No known nonatherosclerotic acute or chronic process or event that could have been potentially lethal.
      3. An "unexpected" death occurs only in a person who is not confined to his home, hospital, or other institution because of illness within 24 hours before death.

II. Suspect atherosclerotic coronary heart disease death - one or both of the following categories:
   A. Death certificate with consistent underlying or immediate cause but neither adequate preterminal documentation of the event nor previous atherosclerotic coronary heart disease diagnosis.
   B. Rapid and unexpected death (requires all three characteristics):
      1. Death occurring between one and 24 hours after the onset of severe symptoms or having last been seen without them.
      2. No known nonatherosclerotic acute or chronic process or event that could have been potentially lethal.
      3. An "unexpected death" occurs only in a person who is not confined to his home, hospital, or other institution because of illness within 24 hours before death.

III. Cardiovascular Death (other than coronary heart disease)
   - Death certificate with consistent, immediate or underlying cause (i.e., atherosclerosis or peripheral arteries with gangrene, atherosclerotic arterial aneurysm with rupture, or atherosclerotic cerebovascular disease)

Table 5.1 (continued)
LRC Coronary Primary Prevention Trial
Primary Fatal End Point Definition

IV. Non Cardiovascular Death
- Death certificate with consistent, immediate or underlying cause (i.e., deaths from all other causes).

Criteria for non-fatal definite or suspect myocardial infarction - any one or more of the following categories:
A. Diagnostic ECG at the time of the event.
B. Ischemic cardiac pain.
C. Diagnostic enzymes.
D. A routine Lipid Research Clinics ECG is diagnostic for myocardial infarction while the previous one was not.
E. Equivocal ECG and equivocal enzymes.
F. Equivocal ECG alone, provided that it is not based on ST or T-wave changes only.

---

## B. Expert Systems Results

The LRC algorithm was used as a basis for the development a knowledge base for the expert system. The rules which comprise the knowledge base are displayed in Appendix A. The engine use in this expert system used a backward chaining engine, and was written using the VP-Expert software (Paperback Software International, Berkeley, CA). This expert system was run on an IBM PC (8086 chip running at 4.77 MHz). This expert system classified all 136 end points in about 90 minutes. The results of this method compared to the LRC final classification are displayed in Table 5.2. Comparing the LRC classification with that of the expert system, the expert system worked reasonably well for the fatal end points definite CHD, suspect CHD, and CVD. Only 40 of the 58 Non-CVD end points were classified correctly by the expert system. Thirteen of these 18 misclassified Non-CVD end points could not be classified by the expert system. Two additional fatal end points could also not be classified by the expert system. Of the 15 deaths which the expert system could not classify, all had missing data on all variables used in the knowledge base of the expert system.

Table 5.2
LRC Coronary Primary Prevention Trial
Classification of Fatal End Points

Comparison of Expert System with LRC Classification

| Expert System Classification | LRC Classification: Definite CHD | Suspect CHD | CVD | Non-CVD | Total |
|---|---|---|---|---|---|
| Definite CHD | 65 | 0 | 0 | 2 | 67 |
| Suspect CHD | 2 | 6 | 0 | 2 | 10 |
| CVD | 0 | 0 | 3 | 1 | 4 |
| Non-CVD | 0 | 0 | 0 | 40 | 40 |
| Unknown | 0 | 1 | 1 | 13 | 15 |
| Total | 67 | 7 | 4 | 58 | 136 |

## C. CART Results

Using CART methods, several elements are necessary. Three of these elements: the splitting rule, the priors, and the costs of misclassification are important in growing classification trees. All possible combinations of the parameters used in growing trees (i.e., splitting rules, priors, and misclassification costs) were used to grow preliminary trees. Two splitting rules were used: Gini and Twoing. Two sets of priors were used: one based on the data, and an equal prior. Two misclassification costs were also used: a unit misclassification cost, and a prespecified cost. The prespecified cost used was to assign a relative cost of one of misclassifying a CVD with a Non-CVD end point, a relative cost of two for misclassifying a definite CHD with a suspect CHD, and a relative cost of five for misclassifying a definite or suspect CHD with a CVD or Non-CVD. The trees were grown and pruned using a 10-fold cross-validation method, and a 1-SE rule for selecting trees. The misclassification results of the growing of these trees are shown in Table 5.3.

Similar results were seen for the two splitting rules. The choice of the priors, and misclassification costs were much more important in selecting an accurate tree. With this dataset, the most important of these two was the priors. Trees grown with priors specified by the data produced a more accurate tree. This was not surprising given the imbalance in the number of classes in this dataset. Of much less importance was the misclassification costs with unit costs producing better trees. Figure 5.1 displays the tree grown with a Gini splitting rule, priors specified by the data, and unit misclassification costs. The tree grown with a Twoing splitting rule, but the same priors and misclassification costs varied only at one split, and produced an identical case by case classification. The split which differs is the DWIN1 is changed to a DCSCHF.

Table 5.3

Misclassification Results of CART
CART Output

MISCLASSIFICATION BY CLASS

Splitting Rule: Gini
Priors: Data
Costs: Unit

| | | CROSS VALIDATION | | | LEARNING SAMPLE | | |
|---|---|---|---|---|---|---|---|
| CLASS | PRIOR PROB. | NO. OF CASES | NO. MIS-CLASSIFIED | COST | NO. OF CASES | NO. MIS-CLASSIFIED | COST |
| 1 | 0.493 | 67 | 3 | 0.45E-01 | 67 | 0 | 0.00E+00 |
| 2 | 0.051 | 7 | 3 | 0.43E+00 | 7 | 1 | 0.14E+00 |
| 3 | 0.029 | 4 | 2 | 0.50E+00 | 4 | 1 | 0.25E+00 |
| 4 | 0.426 | 58 | 5 | 0.86E-01 | 58 | 2 | 0.34E-01 |
| TOTAL | 1.000 | 0 | 13 | | 136 | 4 | |

Splitting Rule: Gini
Priors: Data
Costs: User Specified

| | | CROSS VALIDATION | | | LEARNING SAMPLE | | |
|---|---|---|---|---|---|---|---|
| CLASS | PRIOR PROB. | NO. OF CASES | NO. MIS-CLASSIFIED | COST | NO. OF CASES | NO. MIS-CLASSIFIED | COST |
| 1 | 0.493 | 67 | 1 | 0.75E-01 | 67 | 1 | 0.75E-01 |
| 2 | 0.051 | 7 | 7 | 0.33E+01 | 7 | 7 | 0.29E+01 |
| 3 | 0.029 | 4 | 4 | 0.10E+01 | 4 | 4 | 0.10E+01 |
| 4 | 0.426 | 58 | 3 | 0.26E+00 | 58 | 3 | 0.26E+00 |
| TOTAL | 1.000 | 0 | 15 | | 136 | 15 | |

Splitting Rule: Gini
Priors: Equal
Costs: Unit

| | | CROSS VALIDATION | | | LEARNING SAMPLE | | |
|---|---|---|---|---|---|---|---|
| CLASS | PRIOR PROB. | NO. OF CASES | NO. MIS-CLASSIFIED | COST | NO. OF CASES | NO. MIS-CLASSIFIED | COST |
| 1 | 0.250 | 67 | 7 | 0.10E+00 | 67 | 7 | 0.10E+00 |
| 2 | 0.250 | 7 | 1 | 0.14E+00 | 7 | 1 | 0.14E+00 |
| 3 | 0.250 | 4 | 2 | 0.50E+00 | 4 | 1 | 0.25E+00 |
| 4 | 0.250 | 58 | 7 | 0.12E+00 | 58 | 4 | 0.69E-01 |
| TOTAL | 1.000 | 0 | 17 | | 136 | 13 | |

Table 5.3 (continued)

Misclassification Results of CART
CART Output

MISCLASSIFICATION BY CLASS (continued)

Splitting Rule: Gini
Priors: Equal
Costs: User Specified

| | | CROSS VALIDATION | | | LEARNING SAMPLE | | |
|---|---|---|---|---|---|---|---|
| CLASS | PRIOR PROB. | NO. OF CASES | NO. MIS-CLASSIFIED | COST | NO. OF CASES | NO. MIS-CLASSIFIED | COST |
| 1 | 0.250 | 67 | 7 | 0.34E+00 | 67 | 7 | 0.34E+00 |
| 2 | 0.250 | 7 | 1 | 0.71E+00 | 7 | 1 | 0.71E+00 |
| 3 | 0.250 | 4 | 1 | 0.25E+00 | 4 | 0 | 0.00E+00 |
| 4 | 0.250 | 58 | 53 | 0.12E+01 | 58 | 58 | 0.12E+01 |
| TOTAL | 1.000 | 0 | 62 | | 136 | 66 | |

Splitting Rule: Twoing
Priors: Data
Costs: Unit

| | | CROSS VALIDATION | | | LEARNING SAMPLE | | |
|---|---|---|---|---|---|---|---|
| CLASS | PRIOR PROB. | NO. OF CASES | NO. MIS-CLASSIFIED | COST | NO. OF CASES | NO. MIS-CLASSIFIED | COST |
| 1 | 0.493 | 67 | 3 | 0.45E-01 | 67 | 0 | 0.00E+00 |
| 2 | 0.051 | 7 | 3 | 0.43E+00 | 7 | 1 | 0.14E+00 |
| 3 | 0.029 | 4 | 2 | 0.50E+00 | 4 | 1 | 0.25E+00 |
| 4 | 0.426 | 58 | 5 | 0.86E-01 | 58 | 2 | 0.34E-01 |
| TOTAL | 1.000 | 0 | 13 | | 136 | 4 | |

Splitting Rule: Twoing
Priors: Data
Costs: User Specified

| | | CROSS VALIDATION | | | LEARNING SAMPLE | | |
|---|---|---|---|---|---|---|---|
| CLASS | PRIOR PROB. | NO. OF CASES | NO. MIS-CLASSIFIED | COST | NO. OF CASES | NO. MIS-CLASSIFIED | COST |
| 1 | 0.493 | 67 | 3 | 0.90E-01 | 67 | 2 | 0.60E-01 |
| 2 | 0.051 | 7 | 3 | 0.17E+01 | 7 | 3 | 0.13E+01 |
| 3 | 0.029 | 4 | 4 | 0.10E+01 | 4 | 4 | 0.10E+01 |
| 4 | 0.426 | 58 | 5 | 0.36E+00 | 58 | 3 | 0.26E+00 |
| TOTAL | 1.000 | 0 | 15 | | 136 | 12 | |

Table 5.3 (continued)

Misclassification Results of CART
CART Output

MISCLASSIFICATION BY CLASS (continued)

Splitting Rule: Twoing
Priors: Equal
Costs: Unit

| | | CROSS VALIDATION | | | LEARNING SAMPLE | | |
|---|---|---|---|---|---|---|---|
| CLASS | PRIOR PROB. | NO. OF CASES | NO. MIS-CLASSIFIED | COST | NO. OF CASES | NO. MIS-CLASSIFIED | COST |
| 1 | 0.250 | 67 | 9 | 0.13E+00 | 67 | 4 | 0.60E-01 |
| 2 | 0.250 | 7 | 3 | 0.43E+00 | 7 | 7 | 0.10E+01 |
| 3 | 0.250 | 4 | 2 | 0.50E+00 | 4 | 1 | 0.25E+00 |
| 4 | 0.250 | 58 | 12 | 0.21E+00 | 58 | 4 | 0.69E-01 |
| TOTAL | 1.000 | 0 | 26 | | 136 | 16 | |

Splitting Rule: Twoing
Priors: Equal
Costs: User Specified

| | | CROSS VALIDATION | | | LEARNING SAMPLE | | |
|---|---|---|---|---|---|---|---|
| CLASS | PRIOR PROB. | NO. OF CASES | NO. MIS-CLASSIFIED | COST | NO. OF CASES | NO. MIS-CLASSIFIED | COST |
| 1 | 0.250 | 67 | 13 | 0.61E+00 | 67 | 7 | 0.34E+00 |
| 2 | 0.250 | 7 | 2 | 0.14E+01 | 7 | 1 | 0.71E+00 |
| 3 | 0.250 | 4 | 1 | 0.25E+00 | 4 | 0 | 0.00E+00 |
| 4 | 0.250 | 58 | 53 | 0.12E+01 | 58 | 58 | 0.12E+01 |
| TOTAL | 1.000 | 0 | 69 | | 136 | 66 | |

Figure 5.1

Classification Tree for LRC CPPT Fatal End Points

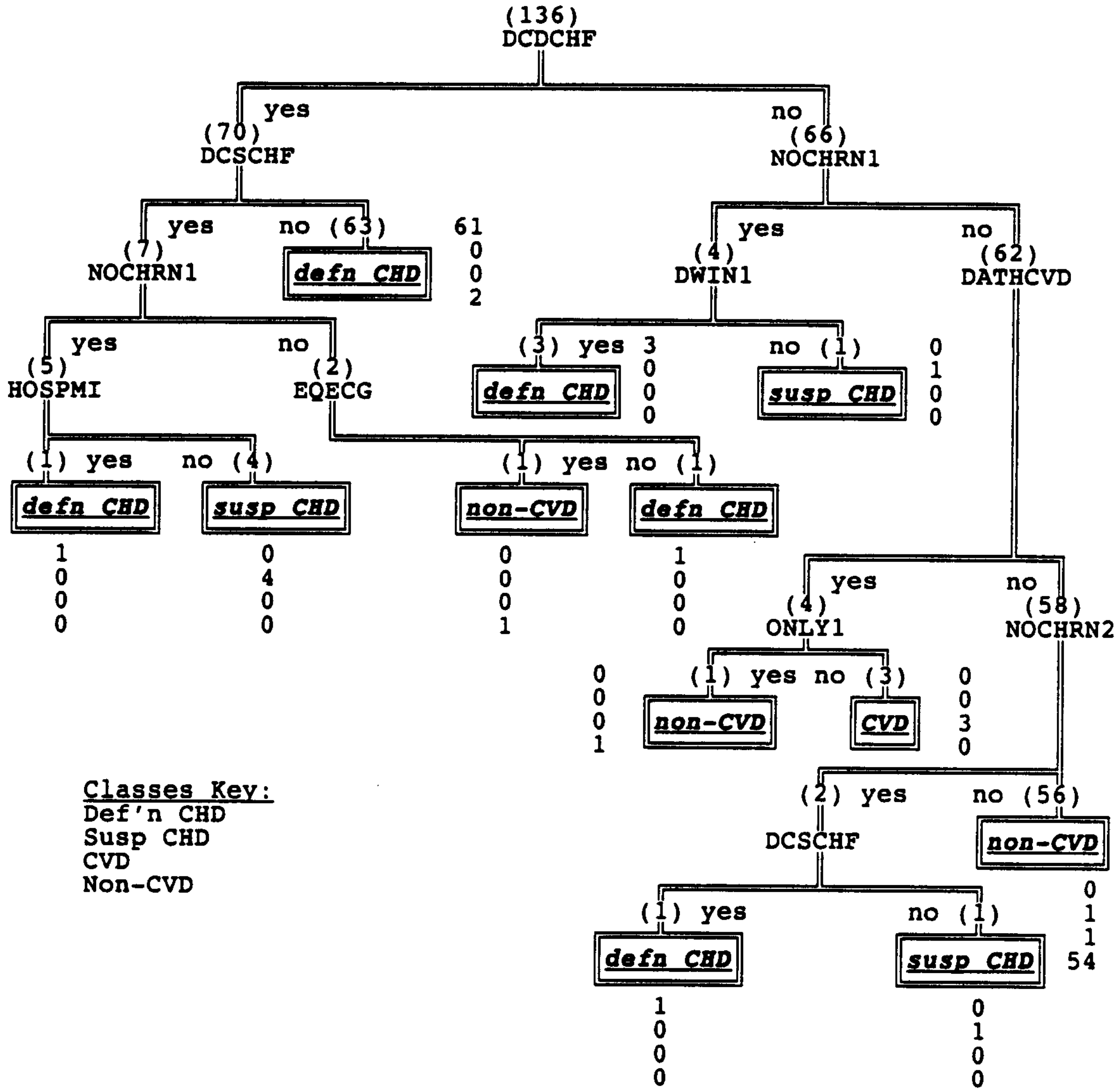

Figure 5.1 Key

DCDCHF: has death certificate consistent with underlying immediate cause (atherosclerotic coronary heart disease)

DCSCHF: has death certificate with consistent underlying or immediate cause (atherosclerotic coronary heart disease) but neither adequate preterminal documentation of the event nor previous atherosclerotic coronary heart disease

NOCHRN1: No known non-atherosclerotic acute or chronic process or event that could have been potentially lethal

HOSPMI: pre-terminal hospitalization with definite or suspect myocardial infarction

EQECG: equivocal enzymes present

DATHCVD: death certificate with consistent underlying or immediate cause (atherosclerotic cerebrovascular disease)

DWIN1: death occurring within one hour after the onset of severe symptoms or having last been seen without symptoms

ONLY1: suspect transient cerebral ischemic attack has occurred only one time (non-fatal end point item)

NOCHRN2: no known non-atherosclerotic acute or chronic process or event that could have been potentially lethal

The results of the most accurate CART tree compared to the LRC classification are displayed in Table 5.4. The tree algorithm agreed with 132 of the 136 LRC fatal end points. The four end points which were misclassified all required the full Cardiovascular Endpoints Committee adjudication, due to disagreement with the two primary reviewers. It appears there was some disagreement regarding the clinical interpretation of certain events used in the classification, or that the documentation of the presence or absence of these events was missing.

Table 5.4
LRC Coronary Primary Prevention Trial
Classification of Fatal End Points
Comparison of CART with LRC Classification

| CART Classification | LRC Classification: Definite CHD | Suspect CHD | CVD | Non-CVD | Total |
|---|---|---|---|---|---|
| Definite CHD | 67 | 0 | 0 | 2 | 69 |
| Suspect CHD | 0 | 6 | 0 | 0 | 6 |
| CVD | 0 | 0 | 3 | 0 | 3 |
| Non-CVD | 0 | 1 | 1 | 56 | 58 |
| Unknown | 0 | 0 | 0 | 0 | 0 |
| Total | 67 | 7 | 4 | 58 | 136 |

## D. Logistic Discriminant Analysis Results

The third method used in this study to classify fatal end points was logistic discrimination. A stepwise procedure was used to select the set of variables to use in the discrimination equation. A 10-fold cross-validation method was then used to estimate the misclassification based on this set of variables. The results of the maximum likelihood estimates for the final logistic model are presented in Table 5.5. The final classifications based on the logistic discriminant function compared to the LRC classification are presented in Table 5.6. The logistic discriminant function did poorly compared to previous methods. All suspect CHD and CVD end points were misclassified. Furthermore, there was an excess in the number of Non-CVD end points classified into the definite CHD group.

Table 5.5
Results of Logistic Discrimination Model
Analysis of Maximum Likelihood Estimates

| Variable | Parameter Estimate | Standard Error | Wald Statistic | Pr > Chi-Square |
|---|---|---|---|---|
| Intercept1 | -8.3674 | 2.0006 | 17.4924 | 0.0001 |
| Intercept2 | -7.3243 | 1.8995 | 14.8674 | 0.0001 |
| Intercept3 | -6.8197 | 1.8597 | 13.4479 | 0.0002 |
| DCDCHF | 4.0925 | 0.6645 | 37.9334 | 0.0001 |
| DWIN1 | 3.6510 | 1.1575 | 9.9500 | 0.0016 |
| DEATHC | -1.8357 | 0.6309 | 8.4651 | 0.0036 |

Table 5.6
LRC Coronary Primary Prevention Trial
Classification of Fatal End Points

Comparison of Logistic Discrimination with LRC Classification

| Logistic Classification | LRC Classification: Definite CHD | Suspect CHD | CVD | Non-CVD | Total |
|---|---|---|---|---|---|
| Definite CHD | 66 | 4 | 0 | 7 | 77 |
| Suspect CHD | 0 | 0 | 0 | 0 | 0 |
| CVD | 0 | 0 | 0 | 0 | 0 |
| Non-CVD | 1 | 3 | 4 | 51 | 59 |
| Unknown | 0 | 0 | 0 | 0 | 0 |
| Total | 67 | 7 | 4 | 58 | 136 |

E. Non-Automated LRC CPPT Trials Results

The results of other methods of end point classification were examined as a basis for comparison. Each of these methods compared to the LRC classifications are displayed in the next three tables. Table 5.7 compares the results of the clinic cardiologist's classification. Table 5.8 contrasts a randomly selected reviewer's (a cardiologist member of the Cardiovascular Endpoints Committee) blinded review. Table 5.9 presents the nosologist's classification of the fatal event based on the death certificate compared to the LRC classification. The clinic cardiologist's review was entirely consistent with the final LRC classification; in this dataset, there were no misclassifications of the fatal event by the clinic cardiologist. However, the blinded review by a member of the Cardiovascular Endpoints Committee, contained some misclassification. There was a tendency for these reviews to err on the side of not classifying definite CHD end points when they were classified as such according the LRC classification. Five out of 67 definite CHD end points and only 2 out of 58 Non-CVD end points were misclassified. Three end points were not given a classification, presumably due to incomplete or insufficient information. Using death certificates, there was not enough information to permit a distinction between definite CHD and suspect CHD. For purposes of comparison, all suspect CHD end points were considered misclassified as definite CHD events. This method had the largest number of Non-CVD end points misclassified with 8 out of 58 events. Four of the 67 definite CHD end points were misclassified; one of these four, and another event could not be classified due to a lack of a death certificate.

Table 5.7
LRC Coronary Primary Prevention Trial
Classification of Fatal End Points
Comparison of Clinic Cardiologist with LRC Classification

| Clinic's Classification | LRC Classification: Definite CHD | Suspect CHD | CVD | Non-CVD | Total |
|---|---|---|---|---|---|
| Definite CHD | 67 | 0 | 0 | 0 | 67 |
| Suspect CHD | 0 | 7 | 0 | 0 | 7 |
| CVD | 0 | 0 | 4 | 0 | 4 |
| Non-CVD | 0 | 0 | 0 | 58 | 58 |
| Unknown | 0 | 0 | 0 | 0 | 0 |
| Total | 67 | 7 | 4 | 58 | 136 |

Table 5.8
LRC Coronary Primary Prevention Trial
Classification of Fatal End Points
Comparison of Random Cardiologist with LRC Classification

| Random Cardiologist's Classification | LRC Classification: Definite CHD | Suspect CHD | CVD | Non-CVD | Total |
|---|---|---|---|---|---|
| Definite CHD | 62 | 0 | 0 | 0 | 62 |
| Suspect CHD | 2 | 6 | 0 | 1 | 9 |
| CVD | 0 | 0 | 4 | 1 | 5 |
| Non-CVD | 1 | 0 | 0 | 56 | 57 |
| Unknown | 2 | 1 | 0 | 0 | 3 |
| Total | 67 | 7 | 4 | 58 | 136 |

Table 5.9
LRC Coronary Primary Prevention Trial
Classification of Fatal End Points
Comparison of Nosologist with LRC Classification

| | LRC Classification | | | | |
|---|---|---|---|---|---|
| Nosologist's Classification | Definite CHD | Suspect CHD | CVD | Non-CVD | Total |
| Definite CHD | 63 | 6 | 1 | 3 | 73 |
| Suspect CHD | 0 | 0 | 0 | 0 | 0 |
| CVD | 2 | 0 | 3 | 4 | 9 |
| Non-CVD | 1 | 1 | 0 | 50 | 52 |
| Unknown | 1 | 0 | 0 | 1 | 2 |
| Total | 67 | 7 | 4 | 58 | 136 |

## F. Misclassification and Kappa Analyses

Misclassifications are compared in Table 5.10 for all methods considered in this study. Both the number of events and the percent of events correctly and incorrectly classified are displayed. The clinic cardiologist's classification agreed with the LRC classification 100 percent of the time. CART trees only misclassified three percent of the LRC fatal end points. Six percent of these fatal events were misclassified by a randomly selected Cardiovascular Endpoints Committee cardiologist's review. The expert system, logistic discrimination function, and the nosologist had approximately the same number and rate of misclassifications. Each of these methods had particular difficulties with the methodology or dataset. For the expert system, 15 of the 22 total misclassified deaths could not be classified into a category; all of these 15 events had missing data for all of the criteria used in the knowledge base used for classification.

Because of small sample sizes of some of the classification groups, logistic discrimination had little power to differentiate end points into these classes. The nosologist did not have a differentiation of two of the final fatal events classification. These were considered misclassifications.

Table 5.10
LRC Coronary Primary Prevention Trial
Misclassification of End Points with Various Methodologies

n [%]

Correctly Classified

| | | End Point | | | | |
|---|---|---|---|---|---|---|
| Method | Statistic | Defn CHD | Susp CHD | CVD | Non-CVD | Total |
| Clinic | n | 67 | 7 | 4 | 58 | 136 |
| | % | *100* | *100* | *100* | *100* | *100* |
| Random Cardio-logist | n | 62 | 6 | 4 | 56 | 128 |
| | % | *95* | *86* | *100* | *97* | *94* |
| Expert System | n | 65 | 6 | 3 | 40 | 114 |
| | % | *97* | *86* | *75* | *69* | *84* |
| CART | n | 67 | 6 | 3 | 56 | 132 |
| | % | *100* | *86* | *75* | *97* | *97* |
| Logistic | n | 66 | 0 | 0 | 51 | 117 |
| | % | *99* | *0* | *0* | *88* | *86* |
| Nosologist | n | 63 | 0 | 3 | 50 | 116 |
| | % | *94* | *0* | *75* | *86* | *85* |

Table 5.11 presents a summary of the weighted kappa statistics and their standard errors. This table compares each of the methods discussed in this chapter with the final LRC fatal classification. The fatal events in this table are the 136 deaths for which data were available for the classification by these methods, except in the

following noted cases. There were two events for which a nosologist and three events for which the Cardiovascular Endpoints Committee cardiologist's review classification were not available. These events were considered unknown. Four different weights for the kappas statistics in Table 5.11 were used: (1) all four classifications plus unknown were distinct (i.e., five categories); (2) the unknown and Non-CVD death categories were considered equivalent (i.e., four categories); (3) CVD, Non-CVD, and unknown death categories were considered equivalent (i.e., three categories); and (4) Definite CHD and Suspect CHD deaths were equivalent, and CVD, Non-CVD, and unknown death categories were considered equivalent (i.e., two categories).

The results of this table of weighted kappa statistics in Table 5.11 match the results of misclassifications in Table 5.10. In general, the clinic cardiologists agreed perfectly with the LRC classification. This was followed by a very close match with the CART tree methodology. The Cardiovascular Endpoints Committee individual clinician's review was less accurate then CART, but still had quite a close agreement. The other three methods, the expert system, logistic discrimination function, and nosologist were much less accurate, and approximately the same. The other result displayed in Table 5.11 is that the less distinction between the different classes of fatal end points that is required, the higher the agreement with the LRC classification. If only two classifications are sought, then almost any method produces good results, and the expert system and individual cardiologist's review almost matched the CART tree results.

Several hypotheses of interest were identified for each of these kappa statistics. They can be analyzed by using contrasts within a weighted least squares generalized linear model framework with one modification. The results of these calculations are presented in Table 5.12. Because all of the weighted kappa statistics of clinic cardiologist with the LRC final classification were identical and two of the measures with the logistic discrimination function were also identical, singularities existed. Furthermore, many of the kappa statistics within and across methodologies were very similar, and close to 1.0 . In this situation where the kappa statistics calculated are very near the extremes of it's range [0,1], very large sample sizes may be necessary for these statistics to have approximately normal distributions. The stability of these estimates become questionable.

These two difficulties were handled in the following manner. With singularities in a weighted-least-squares analysis on a contingency table, the log transformation cannot be done. A small number, 0.0625, was added to each cell. This had the effect of adding one observation uniformly distributed across the contingency table. This was implemented using SAS's PROC CATMOD, using an ADDCELL option to the MODEL statement. For the second problem of the kappa statistics near to the end of the range, a second method of addressing this problem was used. These analyses used Wald statistics using weighted least squares methods. Bloch and Kraemer (1989), Garner (1991), and Carr et al. (1989) all have suggested, in an analogous situation with rank correlation coefficients, to transform the parameters using a Fisher's z transformation. This method was used simultaneously use with the ADDCELL option and was implemented using a RESPONSE statement in PROC CATMOD. The contrasts of interest are: (1) all methods examined: clinic, reviewing cardiologists, expert system, CART, logistic discrimination, and nosologist are equivalent; (2) the clinic's classification is equivalent to the reviewing cardiologist; (3) the clinic's classification is equivalent to the nosologist; (4) the reviewing cardiologist classification is equivalent to the nosologist; and (5) all the automated methods: expert system, CART, and logistic discrimination are equivalent.

Table 5.11
LRC Coronary Primary Prevention Trial
Weighted Kappa Statistics* of Final Classification with Various Methodologies

Kappa
[standard error]

| Methodology | Measure One | Measure Two | Measure Three | Measure Four |
|---|---|---|---|---|
| Clinic | 1.000 [1.28E-9] | 1.000 [1.28E-9] | 1.000 [1.28E-9] | 1.000 [1.28E-9] |
| Random Cardiologist | 0.901 [0.033] | 0.899 [0.034] | 0.907 [0.034] | 0.926 [0.032] |
| Expert System | 0.742 [0.045] | 0.886 [0.036] | 0.907 [0.033] | 0.926 [0.033] |
| CART | 0.948 [0.025] | 0.948 [0.025] | 0.959 [0.023] | 0.956 [0.025] |
| Logistic Discrimination | 0.739 [0.051] | 0.739 [0.051] | 0.789 [0.049] | 0.836 [0.047] |
| Nosologist | 0.742 [0.049] | 0.752 [0.049] | 0.790 [0.049] | 0.867 [0.043] |

*Measure
- One: All classifications distinct (including unknown)
- Two: Unknown equivalent to Non-CVD
- Three: Two plus Non-CVD equivalent to CVD
- Four: Three plus Def'n CHD equivalent to Susp CHD

Table 5.12
LRC Coronary Primary Prevention Trial
Hypothesis Tests of z-transformed Weighted Kappa Statistics* of Final Classification with Various Methodologies

Contrasts of Interest

Wald Statistic
[P-Value]

| Contrast | df | Measure One | Measure Two | Measure Three | Measure Four |
|---|---|---|---|---|---|
| 1) All methods equal | 5 | 26.69 [.0001] | 25.14 [.0001] | 18.89 [.0020] | 10.13 [.0716] |
| 2) Clinic = Cardiologist | 1 | 4.52 [.0335] | 4.53 [0.0333] | 3.91 [.0479] | 2.82 [.0933] |
| 3) Clinic = Nosologist | 1 | 10.52 [.0012] | 9.80 [.0017] | 7.70 [.0055] | 4.77 [.0289] |
| 4) Cardiologist = Nosologist | 1 | 6.08 [.0136] | 5.13 [.0235] | 3.49 [.0618] | 1.08 [.2983] |
| 5) Expert System = CART = Logistic Dis | 2 | 13.43 [.0012] | 1.68 [.4321] | 1.57 [.4553] | 0.53 [.7656] |

*Measure
One: All classifications distinct (including unknown)
Two: Unknown equivalent to Non-CVD
Three: Two plus Non-CVD equivalent to CVD
Four: Three plus Def'n CHD equivalent to Susp CHD

Table 5.12 confirms much of the same results comparing misclassifications and kappa statistics summaries. The answer to the question: are all seven methods of classifying death equivalent in this dataset, is no. Across three of the four, there was substantial evidence that the results were different. There was also strong evidence there were differences among the three methods not developed in this study: clinic versus cardiologist, clinic versus nosologist, and cardiologist versus nosologist. Simultaneously comparing the three methods in this study, there was only evidence of differences among the three methods for the situation where it is important to distinguish unclassifiable deaths/unknowns from Non-CVD deaths. This may likely be due to many unclassifiable events in the expert system, and to the large number of events with missing data for all criteria used in the knowledge base algorithm.

G. Recalculation of LRC CPPT Primary End Point Results

Table 5.13 presents a reanalysis of the final LRC CPPT results when the classification used was the result of the classification of the fatal events of each of the methods examined in this chapter: LRC classification (i.e., manuscript), clinic's classification, and single reviewing cardiologist, expert system, CART, logistic discrimination, and the nosologists. Very small differences exist when repeating the analyses using reclassified fatal end points. This reflects the small number of changes to the classifications of end points used for the primary hypothesis of the LRC manuscript. The LRC CPPT hypothesis was a decrease in definite CHD deaths or definite myocardial infarctions. Observations would change classification only if a fatal event classified by the LRC as a definite CHD death was classified as something else or another fatal event was classified by the alternative methodology as a definite CHD death. This could also change the length of follow-up time. For the clinic, there was complete agreement so the repeat analysis was identical. For the expert system, four events changed classification: two events became positive end points while two positive end points became censored observations. For the CART tree, two additional deaths became classified as definite CHD deaths, thus becoming positive end points. For the logistic discriminant function, one positive end point, a definite CHD death became a censored observation, and eleven deaths were reclassified as definite CHD deaths. For the single cardiologist's review, five definite CHD deaths by LRC classification were reclassified as other than definite CHD deaths; however one of these patients had a definite MI prior to the definite CHD death, and thus only four events changed classification for the primary hypothesis. For nosologist reclassified fatal events, four definite CHD deaths became censored while ten deaths were reclassified as definite CHD deaths. The recalculated relative risks changed no greater than from 20.0 to 23.2 percent. All changes were well with the 95 percent confidence limits of the manuscript calculated relative risk.

Table 5.13
LRC Coronary Primary Prevention Trial
Simulated Results Using Fatal End points as Reclassified by Various Methodologies

| Methodology | Beta | S.E. Beta | RR (%) | Z Score | 95% C.I.(RR) |
|---|---|---|---|---|---|
| Manuscript | -.2228 | .1087 | 20.0 | 2.05 | (1.0, 35.3) |
| Clinic Cardiologist | -.2228 | .1087 | 20.0 | 2.05 | (1.0, 35.3) |
| Random Cardiologist | -.2251 | .1094 | 20.2 | 2.06 | (1.1, 35.6) |
| Expert System | -.2231 | .1087 | 20.0 | 2.05 | (1.0, 35.4) |
| CART | -.2332 | .1084 | 20.8 | 2.15 | (2.0, 36.0) |
| Logistic Discriminant | -.2639 | .1074 | 23.2 | 2.46 | (5.2, 37.8) |
| Nosologist | -.2551 | .1080 | 22.5 | 2.36 | (4.3, 37.3) |

## VI. Discussion and Implications

The goal of this research was to evaluate three methods for end point classification requiring no or minimal human review. No attempt was made to ascertain the correct classification, only to compare the results of each method with that of the LRC CPPT. The dataset used for this purpose must be one which contained several varied methods for end point classification, including one which was rigorous. The LRC CPPT was an excellent dataset for this purpose. One problem in using this dataset was the quality of the individual responses from the independent reviews. Although these data were collected on data collection forms, and were keypunched, they had not undergone any clean up, clarification, or review. Thus, many hours were required for this process. In addition, many questions which arose could not be answered, and thus remained missing or unknown responses. This contributed to the error rates for those methods which would be more sensitive to missing or unknown values for responses.

The process undertaken in this study was to develop, and perform the analyses first which may be influenced by knowing the results of other analyses. Therefore, the development of the expert system was undertaken first. This should have simulated this process if it had been done at the beginning of a trial without the benefit of prior data. The rules for the knowledge base of the expert system used was based entirely on the algorithm of the LRC CPPT documented in the Policy and Procedures Manual of the LRC. Second, the CART methodology was used as some qualitive decision required in the choice of various options in the growing of classification trees. Finally, logistic discrimination was used as little to no decision was required in the process.

The dataset used was identical for all three methods, thus missing values and unresolved responses were classified with values of no and not present. Some of the additional advantages of the CART procedures in handling missing data were not fully utilized. This was done so that identical data could be used for all classification methods. The LRC algorithm also based it's rules on the presence of various conditions. Only after these three automated end point classification procedures were completed was there any comparison with human methods (i.e., the LRC CPPT study method, the clinic review, a randomly selected independent review, and a nosological death certificate review).

As stated previously, expert systems would have done better if there had not been so much missing data. Expert systems usually are developed with the goal that the use will be interactive in nature. Thus any unknown, missing, or otherwise data which were available to the software can be obtained by asking the user. Results of expert systems documented in the literature usually give good results in terms of discriminative accuracy [Davis (1986), Friedman and Frank (1983), Hutchinson et al. (1991), Shortliffe (1984), Spiegelhalter (1983)]. This is not different from my results if the 15 deaths with missing data are eliminated from consideration. Of the remaining 125 deaths, only seven were misclassified. This would be 5.6 percent misclassified, a rate which is comparable with a randomly selected reviewing cardiologist and only slightly higher than CART.

There was a deliberate effort not to involve probability judgement into the development of this expert system. This topic has been one of the continuing controversies statisticians have with expert systems. How one develops this system, assigns probabilities (or confidence factors, or belief factors), particularly when these are based on a human expert (or group of experts) subjective impressions has been the primary statistical concern. Speigelhalter has done much work on the development of expert systems for the diagnosis of disease, diagnostic aids, and classification of adverse drug reactions based on a Bayesian approach [Hutchinson et al. (1991), Speigelhalter (1983), Speigelhalter and Knill-Jones (1984)]. When developing expert systems with confidence factors, he bases these probabilities on data, not the subjective judgement of experts.

Expert systems also have several advantages. They are rapid, completely coherent, and have perfect memory. They may produce results which are difficult to arrive at by intuition due to complexity of the reasoning involved. They work 24 hours a day, do not take vacations, and allow one to reclassify events rapidly and consistently with advances in scientific knowledge and understanding. They can be developed without pre-existing data, if probability judgements are not used in the knowledge base. They can also be used by individuals with relatively little knowledge in a specific domain. They can sometimes be difficult to develop, particularly when probability judgements are used, thus necessitating the need for pre-existing data.

CART or recursive partitioning, had extremely good results. In addition, the full power of CART was not used. As previously stated, missing data were assigned the value *absence/no/unknown* to use identical data for each method. Despite this, only 4 of 136 end points were misclassified in the learning sample. These four end points were particularly difficult to classify; all four of these required adjudication with the full Cardiovascular Endpoints Committee.

Choice of a splitting rule, priors, and costs of misclassification provided some flexibility in the development of these classification trees. The strategy used was to use all combinations of these three choices: Gini and twoing splitting rules, proportional to the data and equal priors, and equal and user specified costs of misclassification. Trees were grown for the eight combinations on a set of data with each end point having one randomly selected from the two (or more) independent cardiologist's review. These observations were the same set as was used for the randomly selected individual cardiologist review. Although not presented, the remaining observations from individuals' review (i.e., not from a full Cardiovascular Endpoints Committee review) were then reclassified with the trees developed as a test sample. With the best classification tree, there was 100 percent agreement with the LRC classification for all 135 end points classified.

I made several observations about the choices of the three parameters used in the growing of classification trees. The choice of a splitting rules makes very little difference to the outcome of the tree and the misclassification rates. Breiman et al. (1984), and Stewart and Stamm (1991) state similar results will occur with either choice. Choice of the priors, and of the cost of misclassification make a larger difference. Unit cost of misclassification provides a more accurate classification than a scheme of higher costs of misclassification for those end points which are more dissimilar (e.g., definite CHD death and Non-CVD death). More importantly, the priors should be proportional to those in the dataset. Two of the four end point classifications have every small numbers of members in the classes. Out of 136 end points, there are only seven suspect CHD deaths and four CVD deaths. Because of these small percentages, a more reliable result is obtained using priors reflecting this distribution. CART is known to give more accurate trees using priors based on the data if the distributions of classes are not equal. Also, if the distribution was more uniform, a priors proportional to the data would be more numerically similar to an equal priors. Using a priors proportional to the data would also be justified in the development of a tree for another study in a similar population, as a similar distribution of end points would be expected.

A few investigators have compared CART to humans in the classification of medical events or identified CART as a potentially useful method for screening observations. These results, when documented, are quite similar in accuracy to the results of this study. Wheeless et al. (1986) have documented using CART trees as a useful prescreen alarm for discriminating normal and abnormal gynecologic specimens. Goldman et al. (1982) compared CART to physicians diagnoses of acute myocardial infarctions in patients with acute chest pain. These investigators grew a classification tree with data from 482 patients from one hospital, and tested it with 468 patients from a second hospital. They found that CART had an equal sensitivity to physicians (91 percent) and a higher specificity (70 versus 67 percent). Goldman et al. (1988) expanded this study to grow a tree with 1370 patients at two hospitals. They prospectively tested this on 4770 patients at six hospitals. Similar results were obtained with a higher specificity with CART compared to physicians (74 versus 71 percent) with similar sensitivities (88.0 versus 87.8 percent). Giampaolo et al. (1991) used CART to identify chemical characteristics of compounds with induce duodenal ulcer and adrenal necrosis in pharmaceutical drug development.

The results of this study are slightly better than most comparisons of CART in the literature. However, many references compare CART to other classification methods in ways which make comparisons difficult. Many authors will eliminate observations for those methods which will not handle missing data, or may impute missing values. However, the CART method is usually at least comparable, if not superior, with other more standard methods. The most common methods used in these comparisons are logistic discriminant, and/or linear discriminant analyses.

Gammerman and Thatcher compare the application of CART diagnoses using Bayes' Theorem in two manners, assuming independence or *simple Bayes*, and a strict application not assuming independence or *proper Bayes*. In this study, CART and the proper Bayes were equivalent with 65 percent correct and the simple Bayes at 74 percent. Several problems exits with this study in that many of the diseases were relatively rare, and no documentation on the splitting rule or more importantly the priors or costs of misclassification were given. Streitberg et al. (1989) reported using CART for the selection of highly discriminating parameters for the analysis of circadian blood pressure using a spline model. They report essentially identical results with CART as with a *robustified* version of Fisher's LDF. Using a pruned tree, they report 15.7 percent false-positive specimens and 5.9 percent false-negative specimens. Stewart and Stamm (1991) used the UNC Caries Risk Assessment Study database to compare the application of CART to linear discriminant analysis and logistic discriminant analysis. They found CART to produce models containing fewer variables than either of the two more classical models with similar sensitivities and specificities. The sensitivities were in the range of 56 to 66 percent in one dataset and 64 to 72 percent in a second dataset with specificities in the range of 71 to 80 percent in one dataset and 75 to 86 percent in a second dataset. They chose to eliminate cases with missing values. MacEntee et al. (1991) compared these same three classification techniques with dental data in an elderly Canadian population. With three objectives, these investigators found a wide range of correct predictions. Using a survey of institutionalized elderly patients, they found the range of predictions from these models of 62 to 74 percent for dental complaints, of 62 to 91 percent for the use of desire for treatment, and 63 to 97 percent for the use of dental services. These results become difficult to interpret given different methods were used to handle missing data for each of these three methods. Harrell et al. (1985) compared four methods of classification for prediction in cancer patients: three methods using logistic regression (i.e., a stepwise variable selection, sickness score as a clinical index, and using incomplete principal components), and CART. They found CART to predict better than stepwise logistic discriminantion with a small training sample (n=110, 61 versus 58 percent). When a larger training sample was used, logistic discriminant predicted better then CART (n=224, 67 versus 56 percent). Altman et al. (1986) directly compared using stepwise logistic discriminantion versus CART for the prediction of idiopathic osteoarthritis of the knee. They used three criteria in their analyses: one based on clinical and laboratory data, one based on clinical and radiographic and one based on clinical data alone. They found logistic and CART methods to be equally sensitive, with sensitivity of 89 to 95 percent, but that CART was more specific. Specificities for logistic methods was 69 to 86 percent whereas for CART they were all 88 percent. They also validated these results with an additional dataset, and found similar results: sensitivities of over 90 percent for both methods, with specificities of 60 to 85 percent for logistic methods and 71 to 84 percent for CART.

Logistic discrimination was chosen as one of the methods of this study because it had some important advantages over other more classical discriminant methods. Many investigators using simulations and data have shown that logistic discrimination is preferred over Fisher's LDF when the data are not normal [Halperin et al. (1971), Press and Wilson (1987), Schmitz et al. (1983), Schmitz et al. (1983), Baron (1991), Campbell et al. (1991)]. Schmitz et al. (1983b) have suggested that there is little difference between LDF and logistic discriminantion in one set of data. They prefer LDF only because of software availability. This is less of a problem today. Campbell et al. (1991) in a simulation study comparing five ordinal models, multinomial logistic, and normal discriminant procedures, found logistic and normal discriminant procedures to be superior and roughly equivalent to a variety of datasets where the variables were independent and dichotomous. When the data were normal, normal discriminant procedure always performed well, ranking first or second. Whereas the multinomial logistic usually performed well with normal data, there were circumstances in which it did perform poorly. Baron (1991) found logistic discrimination preferable to other methods for multiple group classification with non-normal data, and comparable to classification by multinomial LDF for normal data.

In this research, logistic discrimination did not classify as well as the other automated classification methods. It's rates of misclassification and kappas were very comparable to nosologist's classification only using death certificates. Another result of this study, also seen by other investigators, is the number of variables which usually form the logistic discriminant function is small. In this study, only three variables entered the function. Other investigators have reported developing logistic models with the number of variables generally between two to five, with one investigator reported one model with eight variables [Chivot et al. (1990), Ettinger and Jakobsen (1990), Hung et al. (1985), Kansal et al. (1983), Lin and Reschke (1987), Rigby (1991), Roberts et al. (1991), Schmitz et al. (1983b)].

There are many reports in the literature of logistic discriminant analysis not useful as a classification method. Ettinger and Jakobsen (1990) reported very high specificity (98.5 percent) using a logistic discriminant function to report caries in adult patients, but very low sensitivity (7.9 percent). They conclude that this model was not useful as it could not predict persons who would be a high risk for dental caries. Lin and Reschke (1987) reported a comparison of logistic discriminant models with six variables compared to a Bayes linear discriminant model for the prediction of susceptibility and seriousness of space motion sickness. Using a cross-validation method, they found logistic discriminant predicted 53 to 83 percent correct while the Bayes linear discriminant procedure ranged from 48 to 66 percent correct. They conclude that the logistic model predicted correctly more often (an average of 13 percent in model building), and made better use of the information (correctly classifying four to five percent more cases in cross-validation).

Spiegelhalter has continued to do research with automated classification systems in two general areas: in the development of scoring systems based on Bayes' theorem, and on probability judgements in expert systems, also based on the application of Bayes' theorem. Spiegelhalter (1985) notes that these methods are preferable to logistic discrimination, because the number of variables using this method or Fisher's LDF is small and may not be enough for adequate discrimination. Rigby (1991) has used

Spiegelhalter's scoring system to develop a Bayes' Discriminant scoring system, but has not yet reported on a comparison of these methods to logistic discrimination.

It was not the intent of this study to compare the various classification scheme with human *experts* or to validate their classifications. The intent was to compare automated, statistical and other, computer analysis methods of classification. Nevertheless, I observed these three non-automated methods also differed. In general, there was about the same amount of variation with non-automated methods of classifications involving human experts as there was with the automated methods. It was impressive, that the clinic cardiologist, who saw and knew individual patients and their histories agreed with the LRC classification on all fatal end points. However, this observation is difficult to interpret because the clinic cardiologist's classification was a part the LRC classification scheme. Thus, this result is not independent. In fact, the clinic cardiologist could have influenced the final classification, by the selection of medical records send to the CPR. There are no comparable reports in the literature to substantiate this result, although the SOLVD trials used this method to document cause-specific mortality [SOLVD (1990, 1991)].

Comparing human and automated methods, CART did better than the single independent cardiologist's review, and comparable to the expert system if missing data are eliminated from consideration. Using an independent single cardiologist's review, twice as many misclassifications occurred as CART. This should be considered an acceptable rate with a 94 percent overall accuracy. It is also not surprising that the nosologist, relying only on death certificates did the poorest, and that the single cardiologist's review of clinical case histories did better than the nosologist's classification. This has been widely considered a less accurate method for classifying fatal events [Alderson and Meade (1967), Andersen et al. (1991a, 1991b), Beadenkopf et al. (1963), Curb et al. (1983), Florey et al. (1969), Gittlesohn and Senning (1979), Glasser (1981), Hook et al. (1977), James et al. (1955), Kohn (1982), Kuller et al. (1967), Paton (1957), Percy and Dolman (1978), Percy et al. (1981), Rigdon (1981), Schoenberg and Powell (1968), Wingrave et al. (1981)]. It is often used because the only information regarding the death is the death certificate and because of the expense of classifying fatal end points. For these reasons, this method was chosen for comparison. Andersen et al. (1991a, 1991b) report that with perinatal and neonatal deaths, the observed agreement was about 50 percent based on the death certificate and a more extensive cause of death. They also note that when the death certificate was correctly completed, there was considerable improvement in the accuracy of classification using only the death certificate. The nosologist's classification produced an overall accuracy and kappa statistics which were comparable to logistic discrimination, and the expert system when missing data are included.

Finally, only when we must differentiate between unknown and non-CVD fatal events was there any statistical difference among the three automated methods. Thus, a reanalysis of the trial manuscript data for the primary end point would have been the same, regardless of the method, human or automated chosen. All methods examined were comparable with regard to the misclassification of definite CHD deaths with misclassification rates no higher then six percent.

Based on these results, I would make several recommendations. The most important factor in selecting a method, human or otherwise, is to determine how specific and accurate the final classification needs to be. The less accurate the final end points, the more acceptable most methods become. However, if preexisting data are available, or can be reasonably simulated based on summary data, I would advise using CART. This method is fast and is easily implemented. If carefully developed and used, it is also highly accurate. It is easily understood by clinicians, and thus can be more acceptable to clinicians. It also can easily and accurately handle data with missing values.

If preexisting data do not exist or cannot be derived, and missing data will not be a problem (e.g., data are analyzed at the time interactively), then the development of an expert system with clinician input, and careful monitoring would be a reasonable approach. I would rarely recommend a death certificate-based classification by a nosologist be used as the primary end point classification. Only if costs are an important consideration and accuracy is not would a nosologist's classification alone suffice. Finally, it would be prudent if any human *expert* system is used to carefully monitor it and supplement it with an automated system.

## VII. Summary and Future Research

### A. Summary

A critical element of any prevalence study or clinical trial is the validity and reproducibility of the identification and classification of end points. This is often expensive and time-consuming activity. This study was undertaken to examine several automated methods of classifying end points in an accurate, reliable, and reproducible manner. No attempt was made to assess the correctness of the original classification, only to compare the automated methods with those made by human experts. Three methods were chosen: logistic discrimination; recursive partitioning or binary tree structure classifiers, using the software CART; and expert systems. Logistic discrimination was chosen as a standard, well accepted discriminant procedure. It was considered a more appropriate technique than Fisher's LDF given the non-normal dataset used. The CART methodology was chosen as a new and promising technique with some substantial advantages. CART makes no assumptions regarding the distributions of the data. It produces a discriminant function in the form of a decision algorithm or tree which is easily understood by non-statisticians, and easily handles missing data. Expert systems was chosen as a third automated, and non-statistical approach which has gained some popularity in the medical community.

The Lipid Research Clinics Coronary Primary Prevention Trial was a study which enrolled 3,806 men with Type II hypercholerolemia unresponsive to control with diet alone. The primary end points of this study were definite coronary heart disease death or definite myocardial infarction. End points of this trial were classified using a well defined and elaborate scheme. This procedure required the classification of an end point by the clinic cardiologist, and two independent cardiologists' reviews to agree, or go to a full Cardiovascular Endpoints Committee for adjudication. The fatal end points of this trial were used, as well as the clinic cardiologist's classification, and one randomly selected of the two independent cardiologist's review's classification. There was also a nosologist's classification based on the death certificates alone. These were compared to the three automated methods. Unweighted and weighted kappa statistics comparing final LRC classification with each of the study and automated methods were calculated and compared. A reanalysis of the primary end point from the manuscript was done using each of the methods examined in this investigation.

The LRC CPPT dataset was found to have a moderate number of missing values. Because the algorithm used in the expert system, and logistic discriminantion case-wise deletes missing data, for comparison, missing data was combined with response of *no/unknown/absences* responses. CART was the automated method with the high overall accuracy, with only three percent of the data misclassified. Only the clinic cardiologist's classification produced a more accurate classification; no fatal end points were misclassified by this method. The independent cardiologist's classification doubled the rate of misclassification with six percent of the cases misclassified. Other methods performed less well. The expert system had a sixteen percent misclassification of the cases, although if end points with all missing data are eliminated from consideration, the misclassification rate drops to 6 percent. Logistic discriminantion had a 14 percent overall misclassification rate, and the nosologist's classified 15 percent of the end points incorrectly. Unweighted kappa statistics reflect these same relationships; however, if weights are applied to give less importance to similar or unimportant differences, all methods give kappas better then 0.84, and a test for equality of the three automated methods cannot be rejected (with a p-value of 0.43 to 0.76, depending on the weighting).

Based on these results, I would make the following recommendations: use CART whenever possible. CART is fast and is easily implemented. If carefully developed and used, it is also highly accurate. It is easily understood by clinicians, and thus can be more acceptable. It also can easily and accurately handle data with missing values. It does require that previous data exist, if only to reconstruct or simulate observations based on summary data. If this is not possible, the careful development of an expert system is a second alternative. This would require the careful monitoring of both results and data to minimize missing values. I would recommend using logistic discrimination, only as a third line alternative, when precise accuracy is not necessary. This might be done in conjunction with some sort of human classification system or review. Certainly automated systems of classification could be used alone, particularly CART, or secondarily or as part of a scheme which could also include human classification systems.

## B. Future Research

There are two main directions available for future research. One direction is to apply other similar methods to the same dataset, the LRC CPPT. The same three methods could be used on the non-fatal end points of this dataset. A newer recursive partitioning method using the logistic discriminant method could be investigated. This method, FACT or Fast Algorithm for Classification Trees, is computationally faster, but handles missing data by imputation [Loh and Vanichsetakul (1988)]. Another recursive partitioning method has recently been developed for use with censored survival data and categorical response data. This method, RECPAM or recursive partition and amalgamation, has taken the CART methodology and generalized some of the ideas [Ciampi et al. (1989a), Ciampi et al. (1989b), Ciampi et al. (1989)]. It has recently been used on a number of survival trials reported in the medical literature [Ciampi et al. (1990), Marubini et al. (1983), Nadal et al. (1988), Sagman et al. (1991), Shuster et al. (1990), Weisdorf et al. (1991)]. This methodology, either using the survival data, or using fatal end points as a categorical response variable would be an interesting investigation. It could be used to recalculate the final LRC primary statistic and also cluster patients into similar groups based on survival times. A scoring system using Spiegelhalter methods could be developed, using the other cardiologist's review of the fatal end points (i.e., the reviews not selected in this study), although this number is small. This could also be used as probability judgements for an expert system [Speigelhalter and Knill-Jones (1984)]. Finally, these non-selected end points review could be used to develop probability judgements for incorporation into an expert system.

The second major direction for future research is to apply the methods used in this study to other datasets. The Atherosclerosis Risk In Communities (ARIC) is a dataset which might provide baseline data and classifications on fatal end points. The LRC Follow-Up Study (FUS) of the North American Prevalence Study is another LRC study. In this study, a very similar procedure was used to that of the LRC CPPT with regards to how end points were classified. It should also have similar proportions of deaths in each classification. The UNC Collaborative Studies Coordination Center was involved with a TIA/Stroke study at Bowman Gray School of Medicine in Winston-Salem. This study had an elaborate flow chart for classifying end points for which I have already developed an expert system. The Evans County dataset may have enough data, and may have end points classified in a rigorous manner, such that this dataset might be used. The Framingham data also might have fatal end points classified, and certainly would have more than ample data to be useful to investigate these classification methods. Although events leading up to the fatal end points of SOLVD may not be available, other data collected could be used to classify end points with CART, RECPAM, and logistic discriminant analysis. A dataset could be simulated, with approximate proportions seen in the LRC CPPT, with varying degrees of missing data. Finally, I am currently involved in a newly started survival trial in severe congestive heart failure using a continuously administered intravenous prostaglandin. Although the primary end point is all-cause mortality, cause-specific mortality is being collected, and is an ideal clinical trial to utilize some the methodologies investigated in this study.

# Appendix I
## Knowledge Base for Expert System
## Language: VP Expert

```
ACTIONS

          FOR RECNO= 1 TO 306
          GET RECNO=RECORD_NUM,B:EPVVPX,ALL
          FIND VPXDEATH
          PUT B:EPVVPX
          RESET VPXDEATH
          END ;

RULE 1
IF        DCDCHF = P AND
          HOSPMI = Y
THEN      VPXDEATH = DEF_CHF ;

RULE 1A
IF        DCDCHF = P AND
          CHOSPMI = Y
THEN      VPXDEATH = DEF_CHF ;

RULE 1B
IF        DIAGECG = Y
THEN      HOSPMI = Y;

RULE 1C
IF        CARDPN = Y AND
          ABENZ = Y
THEN      HOSPMI = Y ;

RULE 1D
IF        CARDPN = Y AND
          EQENZ = Y AND
          EQECG = Y
THEN      HOSPMI = Y ;

RULE 1E
IF        CARDPN = Y
THEN      HOSPMI = Y ;

RULE 1F
IF        ABENZ = Y
THEN      HOSPMI = Y ;

RULE 1G
IF        EQECG = Y
THEN      HOSPMI = Y ;

RULE 2
IF        DCDCHF = P AND
          PREVANG = Y
THEN      VPXDEATH = DEF_CHF ;
```

```
RULE 3
IF          D_WIN_1 = Y AND
            NO_CHRN1 = Y AND
            NO_HOSP1 = Y
THEN        VPXDEATH = DEF_CHF ;

RULE 4
IF          DCSCHF = Y
THEN        VPXDEATH = SUSP_CHF ;

RULE 5
IF          DCDCHF = P AND
            HOSPMI < > Y AND
            PREVANG < > Y
THEN        VPXDEATH = SUSP_CHF ;

RULE 6
IF          D_1_24 = Y AND
            NO_CHRN2 = Y AND
            NO_HOSP2 = Y
THEN        VPXDEATH = SUSP_CHF ;
RULE 7
IF          DPREFGAN = P OR
            DARTRUP = P OR
            DATHCVD = P
THEN        VPXDEATH = CVD ;

RULE 8
IF          DEATHC = P AND
            DPREFGAN < > P AND
            DARTRUP < > P AND
            DATHCVD < > P AND
            DCDCHF < > P AND
            DCSCHF < > Y
THEN        VPXDEATH = NON_CVD ;
```

**References**

Adams ID, Chan M, Clifford PC, Cooke WM, Dallow V, DeDombal FT, Edwards MH, Hancock DM, Hewett DJ, McIntyre N, Somerville PG, Spiegelhalter DJ, Wellwood J, and Wilson DH (1986). Computer aided diagnosis of acute abdominal pain: a multicentre study. British Medical Journal 293:800-804.

Aikins JS, Kunz JC, and Shortliffe EH (1983). PUFF: An Expert System for Interpretation of Pulmonary Function Data. Computers and Biomedical Research 16:199-208.

Alderson MR, and Meade TW (1967). Accuracy of Diagnosis on Death Certificates Compared with That in Hospital Records. British Journal of Prevention and Social Medicine 21:22-29.

Altman R, Asch E, Bloch D, Bole G, Borenstein D, Brandt K, Christy W, Cooke TD, Greenwald R, Hochberg M, Howell D, Kaplan D, Koopman W, Longley III S, Mankin H, McShane DJ, Medsger Jr T, Meenan R, Mikkelsen W, Moskowitz R, Murphy W, Rothschild B, Segal M, Sokoloff L, and Wolfe F (1986). Development of Criteria for the Classification and Reporting of Osteoarthritis, Classification of Osteoarthritis of the Knee. Arthritis and Rheumatism 29:1039-1049.

Andersen KV, Helweg-Larsen K, and Lange AP (1991). Validiteten af Sundhedsstyrelsens perinatalklassifikation [The validity of the perinatal classification by the Health Department] (in Danish). Ugeskr-Laeger 153:1575-1577.

Andersen KV, Helweg-Larsen K, and Lange AP (1991). Kvaliteten af dodsarsagsregistreringen ved peri- og neonatale dodsfald. [The quality of registration of causes of death in perinatal and neonatal deaths] (in Danish). Ugeskr-Laeger 153:1577-1581.

Anderson TW (1984). An Introduction to Multivariate Statistical Analysis. 2nd Edition. John Wiley and Sons, New York, New York.

Anderson JA (1972). Separate sample logistic discrimination. Biometrika 59:19-35.

Bangdiwala SI, Cohn R, Hazard C, Davis CE, and Prineas RJ (1989). Comparisons of Cause of Death Verification Methods and Costs in the Lipids Research Clinics Program Mortality Follow-up Study. Controlled Clinical Trials 10:167-187.

Baron AE (1991). Misclassificaion amoung Methods Used for Multiple Group Discrimination - The Effects of Distributional Properties. Statistics in Medicine 10:757-766.

Beadenkopf WG, Abrams M, Daoud A, and Marks RU (1963). An Assessment of Certain Medical Aspects of Death Certificate Data for Epidemiologic Study of Arteriosclerotic Heart Disease. Journal of Chronic Diseases 16:249-262.

Beta Blocker Heart Attack Trial Research Group (1982). A Randomized Trial of Propranolol in Patients with Acute Myocardial Infarction. I. Mortality Results. Journal of the American Medical Association 247:1707-1714.

Bloch DA, and Kraemer HC (1989). 2 X 2 Kappa Coefficients: Measures of Agreement of Association. Biometrics 45:269-287.

Bobrow DG, and Stefik MJ (1986). Perspective on Artificial Intelligence Programming. Science 231:957-963.

Brand DA, Frazier WH, Kohlhepp WC, Shea KM, Hoefer AM, Ecker MD, Kornguth PJ, Pais MJ, and Light TR (1982). A protocol for selecting patients with injured extremities who need x-rays. New England Journal of Medicine 306:333-339.

Breiman L, Friedman JH, Olshen RA, and Stone CJ (1984). Classification and Regression Trees. Wadsworth International Group, Belmont, California.

Buchanan BG (1986). Expert systems: working systems and the research literature. Expert Systems 3:32-51.

Campbell MK, Donner A, and Webster KM (1991). Are Ordinal Models Useful for Classification? Statistics in Medicine 10:383-394.

Carr GJ, Hafner KB, and Kock GG (1989). Analysis of Rank Measures of Association for Ordinal Data from Longitudinal Studies. Journal of the American Statistical Association 84:797-804.

Central Patient Registry and Coordinating Center for Lipid Research Clinics (1977). Reference Manual for Lipid Research Clinics Program Prevalence (Phase IV) Follow-up Study. Vol III. April 1977.

Central Patient Registry and Coordinating Center for Lipid Research Clinics (1980). Protocol for the Lipid Research Clinics Type II Coronary Primary Prevention Trial. August 1980.

Chivot L, Mainard F, Bigot E, Bard JM, Auget JL, Madec Y, and Fruchart JC (1990). Logistic discriminant analysis of lipids and apolipoproteins in a population of coronary bypass patients and the significance of apolipoprotiens C-III and E. Atherosclerosis 82:205-211.

Choi SC (1986). Discrimination and Classification:Overview. Computers and Mathematics with Applications 12A:173-177.

Ciampi A. Lawless JF, McKinney SM, and Singhal K (1988). Regression and Recursive Partition Strategies in the Anlysis of Medical Survival Data. Journal of Clinical Epidemiology 41:737-748.

Ciampi A, Hogg SA, McKinney S, and Thiffault J (1988). RECPAM: A Computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. I. Methods and program features. Computer Methods and Programs in Biomedicine 26:239-256.

Ciampi A, Thiffault J, and Sagman U (1989). RECPAM: A computer program for recursive partition amalgamation for censored survival data and other situations frequently occurring in biostatistics. II. Applications to data on small cell carcinoma of the lung (SCCL). Computer Methods and Programs in Biomedicine 30:283-296.

Ciampi A, Schiffrin A, Thiffault J, Quintal H, Weitzner G, Poussier P, and Lalla D (1990). Cluster Analysis of an Insulin-Dpeendent Diabetic Cohort towards the Definition of Clinical Subtypes. Journal of Clinical Epidemiology 43:701-715.

Cochran WG (1977). Sampling Techniques. 3rd Ed. John Wiley & Sons, New York, New York.

Collaborative Studies Coordinating Center (1986). UNC Department of Biostatistics. Protocol. Studies of Left Ventricular Dysfunction (SOLVD)- Coordinating Center.

Cook EF, and Goldman L (1984). Empiric Comparison of Multivariate Analytic Techniques: Advantages and Disadvantages of Recursive Partitioning Analysis. Journal of Chronic Diseases 37:721-731.

Curb JD, Babcock C, Pressel S, Tung B, Remington RD, and Hawkins CM (1983). Nosological Coding of Cause of Death. American Journal of Epidemiology 118:122-128.

d'Anincourt L (1986). OPAL Speeds the Development of Artificial Intelligence System. Research Resources Reporter X(9):1-4.

Davis R (1986). Knowledge-Based Systems. Science 231:951-957.

deDombal FT, Leaper DJ, Staniland JR, McCann AP, and Horrocks JC (1972). Computer-aided Diagnosis of Acute Abdominal Pain. British Medical Journal 2:9-13.

Denning PJ (1986). The Science of Computing Expert Systems. American Scientist 74:18-20.

Diehr P, Wood RW, Barr V, Wolcott B, Slay L, and Tompkins RK (1981). Acute headaches: presenting symptoms and diagnostic rules to identify patients with tension and migraine headache. Journal of Chronic Diseases 34:147-158.

Dillman RO, and Koziol JA (1983). Statistical approach to immunosuppression classification using lymphocyte surface markers and functional assays. Cancer Research 43:417-421.

Duda RO, and Shortliffe EH (1983). Expert Systems Research. Science 220:261-268.

Efron B (1975). The efficiency of logistic regression compared to normal discrimination analysis. Journal of the American Statistical Association 70:892-898.

Ettinger RL, and Jakobsen J (1990). Caries: a Problem in an overdenture population. Community Dentistry and Oral Epidemiology 18:42-45.

Evans RS, Gardner RM, Bush AR, Burke JP, Jacobson JA, Larsen RA, Meier FA, and Warner HR (1985). Development of a Computerized Infectious Disease Monitor (CIDM). Computers and Biomedical Research 18:103-113.

Expert Systems (1985). Directory of microcomputer-based software for expert systems work. Expert Systems 2:222-228.

Feinstein AR, Schimpff CR, and Hull EW (1975). A reappraisal of staging and therapy for patients with cancer of the rectum. I. Development of two new systems of staging. Archives of Internal Medicine 135:1441-1453.

Feinstein AR, Schimpff CR, Andrews JF Jr, and Wells CK (1977). Cancer of the larynx; a new staging system and re-appraisal of prognosis and treatment. Journal of Chronic Diseases 30:277-305.

First JB, Soffer LJ, and Miller RA (1985). QUICK(QUick Index to Caduceus Knowledge): Using the Internist-1/Caduceus Knowledge Base as an Electronic Textbook of Medicine. Computers and Biomedical Research 18:137-165.

Fleiss JL, Spitzer RL, Cohen J, and Endicott J (1977). Three computer diagnosis methods compared. Archives of General Psychiatry 27:643-649.

Florey CDV, Senter MG, and Acheson RM (1969). A Study of the Validity of the Diagnosis of Stroke in Mortality Data. II Comparison by Computer of Autopsy and Clinical Records with Death Certificates. American Journal of Epidemiology 89:15-24.

Friedman RH, and Frank AD (1983). Use of Conditional Rule Structure to Automate Clinical Decision Support: A Comparison of Artificial Intelligence and Deterministic Programming Techniques. Computers and Biomedical Research 16:378-394.

Gammerman A, and Thatcher AR (1991). Bayesian Diagnostic Probabilities without Assuming Independence of Symptoms. Methods of Information in Medicine 30:15-22.

Garner BJ (1991). The Standard Error of Cohen's Kappa. Statistics in Medicine 10:767-775.

Giampaolo C, Gray AT, Olshen RA, and Szabo S (1991). Predicting chemically induced duodenal ulcer and adrenal necrosis with classification trees. Proceedings of the National Academy of Sciences (USA) 88:6298-6302.

Gilpin E, Olshen R, Henning H, and Ross J Jr (1983). Risk prediction after myocardial infarction: comparison of three multivariate methodologies. Cardiology 70:73-84.

Gittlesohn A, and Senning J (1979). Studies on the Reliability of Vital and Health Records: I. Comparison of Cause of Death and Hospital Record Diagnoses. American Journal of Public Health 69:680-689.

Glasser JH (1981). The Quality and Utility of Death Certificate Data. American Journal of Public Health 71: 231-233.

Goldman L, Weinberg M, Weisberg M, Olshen R, Cook EF, Sargent RK, Lamas GA, Dennis C, Wilson C, Deckelbaum L, Fineberg H, Stiratelli R, and the Medical House Staffs at Yale-New Haven Hospital and Brigham and Women's Hospital (1982). A Computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. New England Journal of Medicine 307:588-596.

Goldman L, Cook EF, Brand DA, and Lee TH, Rouan GW, Weisberg MC, Acampora D, Stasiulewicz C, Walshon J, Terranova G, Gottlieb L, Kobernick M, Goldstein-Wayne B, Copen D, Daley K, Brandt AA, Jones D, Mellors J, and Jakubowski R (1988). A Computer Protocol to Predict Myocardial Infarction in Emergency Department Patients with Chest Pain. New England Journal of Medicine 318:797-803.

Goodall A (1985). The Guide to Expert Systems. Learned Information, Oxford England and Medford, New Jersey.

Gordon T (1974). Hazards in the Use of the Logistic Function with Special Reference to Data from Prospective Cardiovascular Studies. Journal of Chronic Diseases 27:97-102.

Halperin M, Blackwelder WC, and Verter JI (1971). Estimation of the Multivariate Logistic Risk Function: A Comparison of the Discriminant Function and Maximum Likelihood Approaches. Journal of Chronic Diseases 24:125-158.

Harmon P, and King D (1985). Expert Systems - Artificial Intelligence in Business. John Wiley and Sons, New York.

Harrell FE, Lee, KL, Matchar DB, and Reichert TA (1985). Regression Models for Prognostic Prediction: Advantages, Problems, and Suggested Solutions. Cancer Treatment Reports 69:1071-1077.

Hayes-Roth F, Waterman DA, and Lenat DB (eds) (1983). Building Expert Systems. Addison-Wesley Publishing Company, Reading Massachusetts.

Helliwell J (1986). Guru: Brave New Expert System? PC Magazine 5(10, May 27):151-163.

Henning H. Gilpin EA, Covell JW, Swan EA, O'Rourke RA, and Ross J Jr. (1976). Prognosis after acute myocardial infarction: multivariate analysis of mortality and survival. Circulation 59:1124-1136.

Hjermann I, Velve Byre K, Holme I, and Leren P (1981). Effect of Diet and Smoking Intervention on the Incidence of Coronary Heart Disease. Lancet II:1303-1310.

Honigfeld G, Klein DF, and Feldman S (1969).. Prediction of psychopharmacologic effect in man: Development and validation of a computerized diagnostic decision tree. Computers in Biomedical Research 2:350-361.

Hook EB, Farina MA, and Hoff MB (1977). Death Certificate Reports of Cardiovascular Disorders in Children: Comparison with Diagnoses in a Pediatric Cardiology Registry. Journal of Chronic Diseases 30:383-391.

Hung J, Chaitman BR, Lam J, Lesperance J, Dupras G, Fines P, Cherkaoui O, Robert P, and Bourassa MG (1985). A logistic regression analysis of multiple noninvasive tests for the prediction of the presence and extent of coronary artery disease in men. American Heart Journal 110:460-469.

Hutchinson TA, Dawid AP, Spiegelhalter DJ, Cowell RB, and Roden S (1991). Computerized Aids for Probabilistic Assessment of Drug Safety II: An Expert System. Drug Information Journal 25:41-48.

Imrey PB, Koch GG, Stokes ME, Darroch JN, Freeman DH, and Tolley HD (1981). Categorical Data Analysis:Some Reflections on the Log Linear Model and Logistic Regression. Part I: Historical and Methodological Overview. International Statistical Review 49:265-283.

Imrey PB, Koch GG, Stokes ME, Darroch JN, Freeman DH, and Tolley HD (1982). Categorical Data Analysis:Some Reflections on the Log Linear Model and Logistic Regression. Part I: Data Analysis. International Statistical Review 50:35-63.

Jacobs DR, Mebane IL, Bangdiwala SI, Criqui MH, and Tyroler HA (1990). High Density Lipoprotein Cholesterol as a Predictor of Cardiovascular Disease Mortality in Men and Women: The Follow-up Study of the Lipid Research Clinics North American Prevalence Study. American Journal of Epidemiology 131:32-47.

Jackson PC Jr (1985). Introduction to Artificial Intelligence. 2nd Ed. Dover Publication, New York.

Jackson P (1986). Introduction to Expert Systems. Addison-Wesley Publishing Company, Reading Massachusetts.

Jagannathan V, Bourne JR, Jansen BH, and Ward JW (1982). Artificial intelligence methods in quantitative electroencephalogram analysis. Computers and Biomedical Research 15:249-258.

James G, Patton RE, and Heslin AS (1955). Accuracy of Cause-of Death Statements on Death Certificates. Public Heath Reports 70:39-51.

Kansal S, Roitman D, Bradley EL, and Sheffield T (1983). Enhanced Evaluation of Treadmill Tests by Means of Scoring Based on Multivariate Analysis and Its Clinical Application: A Study of 608 Patients. American Journal of Cardiology 52:1155-1160.

Klein DF, Honigfeld G, and Feldman S (1972). Prediction of drug effect by diagnostic decision tree. Diseases of the Nervous System 29(Suppl):159-187.

Kleinbaum DG, Kupper LL, and Chambless LE (1982). Logistic Regression Analysis of Epidemiologic Data: Theory and Practice. Communications in Statistics-Theory and Methods 11:485-547.

Knill-Jones RP (1987). Diagnositc systems as an aid to clinical decision making. British Medical Journal 295:1392-1396.

Koch GG, Amara IA, Davis GW, and Gillings DB (1982). A Review of Some Statistical Methods for Covariance Analysis of Categorical Data. Biometrics 38:563-595.

Kohn, RR (1982). The Cause of Death in Very Old People. Journal of the American Medical Association 247:2793-2797.

Kuller L, Blanch T, and Havlik R (1967). Analysis of the Validity of Cerebrovascular Disease Mortality Statistics in Maryland. Journal of Chronic Diseases 20:841-851.

Kuller L, Neaton J, Caggiula A, and Falvo-Gerard L (1980). Primary Prevention of Heart Attacks: the Multiple Risk Factor Intervention Trial. American Journal of Epidemiology 112:185-199.

Lachenbruch PA (1975). Discriminant Analysis. Hafer, New York, New York.

Lehner PE, and Barth SW (1985). Expert systems on microcomputers. Expert Systems 2:198-208.

Lemmon H (1986). Comax: An Expert System for Cotton Crop Management. Science 233:29-33.

Liebowitz J (1986). Useful approach for evaluating expert systems. Expert Systems 3:86-96.

Lin KK, and Reschke MF (1987). The Use of the Logistic Model in Space Motion Sickness Prediction. Aviation, Space, and Environmental Medicine 58:A9-A15.

Lindley DV (1987). The Probability Approach to the Treatment of Uncertainty in Artificial Intelligence and Expert Systems. Statistical Science 2:17-24 (with Comments and Rejoinder pp 30-44).

Lipid Research Clinics Program (1979). The Coronary Primary Prevention Trial: Design and Implementation. Journal of Chronic Diseases 32:609-631.

Lipid Research Clinics Program (1984). The Lipid Research Clinics Coronary Primary Prevention Trial Results I. Reduction in Incidence of Coronary Heart Disease. Journal of the American Medical Association 251:351-364.

Loh WY, and Vanichsetakul N (1988). Tree-Structured Classification Via Generalized Discriminant Analysis. Journal of the American Statistical Association 83:715-725 (with Comment and Rejoinder 725-727)

MacEntee MI, Hill PM, Wong G, Mojon P, Berkowitz J, and Glick N (1991). Predicting Concerns for the Mouth among Institutionalized Elders. Journal of Public Health Dentistry 51:82-90.

Marubini E, Morabito A, and Valsecchi MG (1983). Prognostic Factors and Risk Groups: Some Results Given by Using an Algorithm Suitable for Censored Survival Data. Statistics in Medicine 2:295-303.

Masarie FE Jr, Miller RA, and Myers JD (1985). INTERNIST-I Properties:Representing Common Sense and Good Medical Practice in a Computerized Medical Knowledge Base. Computers and Biomedical Research 18:458-479.

Meyers A, Brand DA, Dove HG, and Dolan TF (1978). A Technique for analyzing clinical data to provide patient management guidelines: A study of meningitis in children. American Journal of Diseases in Children 132:25-29.

Michaelsen RH, Michie D, and Boulanger A (1985). The Technology of Expert Systems. Byte 10:303-312.

Miller PL (1984). Medical plan-analysis by computer. Computers and Biomedical Research 18:15-20.

Miller PL, Blumenfrucht SJ and Black HR (1984). An Expert System Which Critiques Patient Workup: Modeling Conflicting Expertise. Computers and Biomedical Research 17:554-569.

Miller PL (1985). Goal-Directed Critiquing by Computer:Ventilator Management. Computers and Biomedical Research 18:422-438.

Miller RA, Pople HE Jr, and Myers JD (1982). INTERNIST-I, An Experimental Computer-Based Diagnostic Consultant for General Internal Medicine. New England Journal of Medicine 307:468-476.

Nadal MJ, Jouve M, Mosseri V, Asselain B, and Pouillart P (1988). Cancer Metastase du sein Traite Par Polychimiotherapie: Une Nouvelle Approche du Pronostic [Polymetastatic breast carcinoma treated by polychemotherapy: a new prognostic approach] (in French). Bulletin of Cancer (Paris) 75:757-769.

O'Connell DL, Davis CE, Hazard C, Irving S, and Rubenstein C (1985). Classification of Endpoints in the LRC CPPT. Contributed Paper. Presented at the 6th Annual Society for Clinical Trials Meetings May 12-15, 1985, New Orleans, LA.

O'Hare GM, and Bell DA (1985). The coexistence approach to knowledge representation. Expert Systems 2:230-238.

Paton BC (1957). The Accuracy of Diagnosis of Myocardial Infarction. A Clinicopathologic Study. American Journal of Medicine 15:761-768.

Percy C, and Dolman A (1978). Comparison of the Coding of Death Certificates Related to Cancer in Seven Countries. Public Health Reports 93:335-350.

Percy C, Stanek E, and Gloeckler L (1981). Accuracy of Cancer Death Certificates and Its Effect on Cancer Mortality Statistics. American Journal of Public Health 71:242-250.

Press SJ, and Wilson S (1987). Choosing Between Logistic Regression and Discriminant Analysis. Journal of the American Statistical Association 73:699-705.

Remington RD (1984). Who Should Code Cause of Death in a Clinical Trial? Controlled Clinical Trials 5:241-244.

Rich E (1983). Artificial Intelligence. McGraw-Hill Book Company, New York, New York.

Rigby AS (1991). Development of a Scoring System to Assist in the Diagnosis of Rheumatoid Arthritis. Methods of Information in Medicine 30:23-29.

Rigdon RH (1981). Problems in a Statistical Study of Disease Based on Death Certificates. Southern Medical Journal 74:1104-1106.

Robert AR, Melin JA, and Detry JMR (1991). Logistic Discrimination Analysis Improves Diagnostic Accuracy of Exercise Testing for Coronary Artery Disease in Women. Circulation 83:1202-1209.

Sagman U, Maki E., Evans WK, Waar D, Shepherd FA, Sculier JP, Haddad R, Payne D, Pringle JF, Yeoh JL, Ciampi A, DeBoer G, McKinney S, Ginsberg R, and Feld R (1991). Small-Cell Carcinoma of the Lung: Derivation of a Prognositic Staging System. Journal of Clinical Oncology 9:1639-1649.

SOLVD Investigators, The (1990). Studies of Left Ventricular Dysfunction (SOLVD) - Rationale, Design and Methods: Two Trials That Evaluate the Effect of Enalapril in Patients with Reduced Ejection Fraction. American Journal of Cardiology 66:315-322.

SOLVD Investigators, The (1991). Effect of Enalapril on Survival in Patients with Reduced Left Ventricular Ejection Fractions and Congestive Heart Failure. New England Journal of Medicine 325:293-302.

Schafer DG (1985). Microcomputer-based expert systems;where we are, where we are headed. Expert Systems 2:188-195.

Schmitz PIM, Habbema JDF, and Hermans J (1983). The Performance of Logistic Discrimination on Myocardial Infarction Data, in Comparison with Some Other Discriminant Analysis Methods. Statistics in Medicine 2:199-205.

Schmitz PIM, Habbema JDF, Hermans J, and Raatgever JW (1983). Comparative Performance of Four Discrimination Analysis Methods for Mixtures of Continuous and Discrete Variables. Communications in Statistics-Simulation and Computations 12:727-731.

Schoenberg, BS and Powell JM (1968). Statistics on Stroke. A Pilot Study of the Clinical Evidence Justifying the Reporting of Stroke on Death Certificates in Alameda County, California. California Medicine 109:19-23.

Shafer G (1987). Probability Judgment in Artificial Intelligence and Expert Systems. Statistical Science 2:3-16 (with Comments and Rejoinder pp 30-44).

Shortliffe EH (1984). Reasoning methods in medical consultation systems: Artificial intelligence approaches. Computer Programs in Biomedicine 18:5-14.

Shuster JJ, Falletta JM, Pullen DJ, Crist WM, Humphrey GB, Dowell BL, Wharam MD, and Borowitz M (1990). Prognostic Factors in Childhood T-Cell Acute Lymphoblastic Leukemia: A Pediatric Oncology Group Study. Blood 75:166-173.

Spiegelhalter DJ (1983). Evaluation of Clinical Decision-Aids, with an Application to a System for Dyspepsia. Statistics in Medicine 2:207-216.

Speigelhalter DJ, and Knill-Jones, RP (1984). Statistical and Knowledge-based Approaches to Clinical Decision-support Systems, with an Application in Gastroenterology. Journal of the Royal Statistical Society Series A 147:35-77.

Spiegelhalter DJ (1985). Statistical Methodology for Evaluationg Gastrointestinal Symptoms. Clinics in Gastroenterology 14:489-515.

Spiegelhalter DJ (1987). Probabilistic Expert Systems in Medicine: Practical Issues in Handling Uncertainty. Statistical Science 2:25-30 (with Comments and Rejoinder pp 30-44).

Spitzer RL, and Endicott J (1968). DIAGNO: A computer program for psychiatric diagnosis utilizing the differential diagnostic procedure. Archives of General Psychiatry 18:746-756.

Spitzer RL, and Endicott J (1969). DIAGNO II: Further development in a computer program for psychiatric diagnosis. American Journal of Psychiatry 125(Suppl 7):12-21.

Stern RB, Knill-Jones RP, and Williams R (1974). Clinician Versus Computer in the Choice of 11 Differential Diagnoses of Jaundice Based on Formalised Data. Methods of Information in Medicine 13:79-82.

Stewart PW, and Stamm JW (1991). Classification Tree Prediction Models for Dental Caries from Clinical, Microbiological, and Interview Data. Journal of Dental Research 70:1239-1251.

Streitberg B, Meyer-Sabellek W, and Baumgart P (1989). Statistical analysis of circadian blood pressure recordings in controlled clinical trials. Journal of Hypertension 7:S11-S17.

Sutherland DH, Olshen R, Cooper L, and Woo SL-Y (1980). The development of mature gant. Journal of Bone and Joint Surgery (American) 62:336-353.

Theil H (1969). A multinomial extension of the linear logit model. International Econometric Review 10:251-259.

Thompson BA, and Thompson W (1985). Inside an Expert System. Byte 10:315-330.

Vanker AD, and Van Stoecker W (1984). An Expert Diagnostic Program for Dermatology. Computers and Biomedical Research 17:241-247.

Walker SH, and Duncan DB (1967). Estimation of the probability of an event as a function of several independent variables. Biometrika 54:167-179.

Weisdorf D, Hakke R, Blazar B, Miller W, McGlave P, Ramsay N, Kersey J, and Filipovich A (1991). Risk Factors for Acute Graft-Versus-Host Disease in Histocompatible Donor Bone Marrow Transplantation. Transplantation 51:1197-1203.

Wheeless LL, Robinson RD, Cox C, Berkan TK, and Reeder JE (1986). A Statistical Analysis of Prescreening Alarms in a Population of Normal and Abnormal Gynecologic Specimens. Cytometry 7:205-211.

Wingrave SJ, Beral V, Adelstein AM, and Kay CR (1981). Comparison of cause of death coding on death certificates with coding in the Royal College of General Practitioners Oral Contraception Study. Journal of Epidemiology and Community Health 35:51-58.

Winston PH (1984). Artificial Intelligence. 2nd Edition. Addison-Wesley Publishing Company, Reading Massachusetts.