RISK PREDICTION MODELS FOR BINARY RESPONSE VARIABLES FOR THE CORONARY BYPASS OPERATION

by

HONGBIN ZHANG

M.Eng., Institute of Software, Academia Sinica, 1987

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

Department of Statistics

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

June 1993

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

(Signature)

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date June 28 1993

DE-6 (2/88)

Abstract

The ability to predict 30 day operative mortality and complications following coronary artery bypass surgery in the individual patient has important implications clinically and for the design of clinical trials. This thesis focuses on setting up risk stratification algorithms.

Utilizing the binary feature of the response variables, logistic regression analyses and classification trees (recursive partitioning) were used with the variables identified by the Health Data Research Institute in Portland, Oregon. The data set contains records for 18171 patients who had coronary artery bypass surgery in one of several hospitals between 1968 to 1991. Statistical models are set up, one from each method, for six outcome variables of the surgery: 30 day operative mortality, renal shutdown complication, central nervous system complication, pneumothorax complication, myocardial infarction complication and low output syndrome.

The risk groups vary across different outcomes. The history of cardiac surgery has strong association with operative mortality and patients who suffer from a central nervous system disease tend to have higher risks for all the outcomes. Further study is necessary to consider the differences among hospitals and to divide the population according to the type of previous cardiac surgery.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my appreciation to my supervisor, Dr. Harry Joe, for his guidance, suggestions and assistance in producing this thesis. I also would like to give my thanks to Dr. Stan Page for his critical reading and helpful comments. I am indebted to the Health Data Research Institute for some financial support and for providing the data, to Scott Page, M.D., for advice on medical aspects of this study. The support of UBC Department of Statistics is also gratefully acknowledged. Finally, I thank my wife, Hua, for her continuous encouragement and support. Without her belief in me, it might have taken me longer to get here.

Chapter 1

Introduction

Since the beginning of coronary bypass surgery at the Good Samaritan Hospital and Medical Center, Portland, Oregon, in 1968, patient data have been recorded in order to manage the care of the patients. This management will be measured by outcome analysis of the 30 day operative mortality and complications arising from the surgery. Although of secondary concern, some complications directly affect the quality of life of patients, for example, renal shutdown may require dialysis treatment. From a clinical point of view, a patient with low risk of mortality or complications could be spared the discomfort and expense of unnecessary treatment in a coronary care unit; based on a prognostic assessment such patients could be placed into an intermediate care unit or a general ward, or be discharged early from the hospital.

However, for these concerns to be realized, it is necessary to establish some risk stratification algorithms. Two approaches have been devised for surgery patients depending on the strategy used for derivation. In one strategy, a panel of experts identifies and assigns weights to clinical variables believed to be associated with outcomes of interest. The second strategy uses statistical modeling to relate empirical data for many patient variables to outcomes of interest.

In this study, the odds ratio gives a simple summary for the binary risk factors; the logistic regression analysis and the tree-based methods are used to set up the stratification algorithms. While doing the analysis, answers to the following questions are sought:

1). What risk groups are most important in predicting the outcomes?

2). How does each stratification algorithm perform in predicting?

The clinical background of the coronary bypass procedure and the data description are presented in Chapter 2 while the initial data analysis is performed in Chapter 3. In Chapter 4 and Chapter 5 the statistical methodologies and results of analyses are described. Conclusions and suggestions are given in Chapter 6.

Chapter 2

Coronary Bypass Surgery and Data Source

2.1 Coronary Bypass Surgery

Coronary heart disease is the comprehensive term which includes all of the clinical manifestations that result from atherosclerotic narrowing or occlusion of the arteries which supply the heart muscle. In several countries of the industrialized world, it is the major cause of death in both men and women. Despite accumulating knowledge about the epidemiology and pathology of disease of the heart and coronary arteries, there is not, as yet, a way of intervening to definitely arrest the natural progression of atherosclerosis in order to prevent or cure this condition. Compared with medical treatment, the surgical technique is seemingly more radical for coronary artery disease. In 1963, the first coronary artery bypass surgery was performed in the United Kingdom. Nowadays, it is the most common form of elective surgery.

The surgical technique of coronary grafting involves opening the chest wall and temporarily stopping the heart while circulation is maintained with a heart-lung machine. A vein is removed to be used as the graft material. Each obstructed section of artery is then bypassed by attaching one end of vein to the aorta carrying blood for the heart, and the other end to the artery beyond the stenosis or occlusion. The heart is restarted, the chest wall closed, and the operation completed.

Coronary artery surgery has been described as relieving or very much reducing angina in over 90% of patients. Bad results are operative death and complications arising from the surgery although the mortality rate is believed to be decreasing with increasing surgical experience and skill.

2.2 Source of Data

MCR (Merged, Multi-Center, Multi-Specialty Clinical Registries) is an international database system developed by Health Data Research Institute (formerly Dendrite Systems, Inc.) in which information of patients who had heart related surgery were recorded. This database system is used by several hospitals and the contributors (patients, sometime the doctors) are encouraged to enter information. In doing that, MCR uses a long systematic set of questions to elicit information both prior to operation and after operation (see Appendix). The data set analyzed here consists of 18171 patients from the MCR who had coronary bypass surgery between 1968 to 1991.

The pre-operation information include date of operation, patient's age, gender, prior myocardial infarction, existence or non-existence of other diseases, body surface area, etc. The post-operation information include patient's status during or after the bypass surgery; for example, complications, such as renal or neurological problems, and survival status to 30 days following surgery.

The variables of primary interest in our analysis are those outcome variables indicating the patient's complications and survival status after the surgery. All variables studied and their abbreviations are listed in Table 1 to Table 3.

Table 1: Code Sheet for the MCR Data

Variable	Name	Codes/Values	Abbreviation
1	Age	Years	AGE
2	Sex	0=Male 1=Female	SEX
3	Prior Myocardial Infarction	0=No 1=Yes	PMI
	(Variables 4-23 pertain to other diseases)		
4	Obesity	0=No 1=Yes	OBE
5	Chronic Obstructive Pulmonary Disease	0=No 1=Yes	COP
6	Diabetes	0=No 1=Yes	DIA
7	Cholesterol Level$\geq$200	0=No 1=Yes	CH2
8	Cholesterol Level$\geq$300	0=No 1=Yes	CH3
9	Renal Disease	0=No 1=Yes	REN
10	Hypertension	0=No 1=Yes	HTN
11	Alcohol Abuse	0=No 1=Yes	ETO
12	Drug Abuse	0=No 1=Yes	DRU
13	Marfan's Syndrome (a skeletal abnormality)	0=No 1=Yes	MAR
14	HIV+	0=No 1=Yes	HIV
15	AIDS	0=No 1=Yes	AID
16	Cancer	0=No 1=Yes	CA
17	Anemia	0=No 1=Yes	ANE
18	Liver Disease	0=No 1=Yes	LIV
19	Central Nervous System Disease	0=No 1=Yes	CNS
20	Prior Cerebrovascular Accident	0=No 1=Yes	PCA

Table 2: Code Sheet for the MCR Data (continued)

Variable	Name	Codes/Values	Abbreviation
21	Rheumatic Heart Disease	0=No 1=Yes	RHE
22	Pulmonary Hypertension	0=No 1=Yes	PUL
23	Chronic Dialysis	0=No 1=Yes	CHR
	(Variables 24-29 pertain to type of prior cardiac surgery)		
24	Other Surgery	0=No 1=Yes	OTH
25	No Surgery	0=No 1=Yes	NON
26	Coronary Bypass Graft	0=No 1=Yes	CAB
27	Valve Replacement	0=No 1=Yes	VAL
28	Congenital	0=No 1=Yes	CON
29	Pacemaker	0=No 1=Yes	PAC
30	Left Ventricular Dysfunction	0=Normal 1= 40-49% 2= 30-39% 3= 20-29% 4= $\leq$ 20%	LVD
31	Prior Operation Status	1=Elective 2= Urgent 3= Emergency 4= Desperate	POS
32	Body Surface Area	Square meters	BSA

Table 3: Code Sheet for the MCR Data (continued)

Variable	Name	Codes/Values	Abbreviation
	(Variables 33-47 pertain to the complications after surgery)		
33	Reoperation for Bleeding	0=No 1=Yes	REP
34	Renal Shutdown (Mild)	0=No 1=Yes	REM
35	Renal Shutdown (Severe)	0=No 1=Yes	RES
36	Wound (Severe)	0=No 1=Yes	WOU
37	Neurological (Mild)	0=No 1=Yes	NEM
38	Neurological (Severe)	0=No 1=Yes	NES
39	Pulmonary (Mild)	0=No 1=Yes	PUM
40	Pulmonary (Severe)	0=No 1=Yes	PUS
41	Myocardial Infarction	0=No 1=Yes	MI
42	Low Output (Mild)	0=No 1=Yes	LOM
43	Low Output (Severe)	0=No 1=Yes	LOS
44	Clotting	0=No 1=Yes	CLO
45	Sepsis	0=No 1=Yes	SEP
46	Gastrointestinal Bleeding	0=No 1=Yes	GIB
47	Diffuse Intravascular Coagulation	0=No 1=Yes	DIC
48	Discharge/30 Day Status	1=Live 2= Died in OR 3= Died in Hosp/30D 4= Reop 5= Died Late Cardiac 6= Unrelated Death 9= Lost to Follow-up	STA

Chapter 3

Initial Data Analysis

3.1 Summary Statistics for the Data

Since the data are collected from several populations, missing values on several variables are inevitable. Of 18,171 MCR patients, 12,000 had complete data for the 32 variables selected. The missing data mostly occur in two variables *prior myocardial infarction*, 30% and *left ventricular dysfunction*, 29.3%. These are the only two which measure the previous damage of heart, so they should not be excluded.

20 kind of diseases/conditions are suspected to be related with the success of the surgery. But seven of them can be removed because of their lower incidences: *drug abuse, 0.3% (70 patients)*; *Marfan's syndrome, 0.1% (20 patients)*; *positive test for AIDS and AIDS, 0.0% (0 patients)*; *anemia, 0.0% (18 patients)*; *pulmonary hypertension, 0.3% (59 patients)*; *chronic dialysis, 0.0% (11 patients)*. The remaining 13 diseases are: OBE (obesity: 1.5x expected body weight), COP (patient with distinct limitations revealed at time of study or on treatment–bronchodilators, etc), DIA (diabetes: patient on oral medicine or insulin), CH2 (patient with cholesterol blood levels between 200-299), CH3 (patient with cholesterol blood level above 300), REN (renal failure: patients not on dialysis

with creatinines above 2.5), HTN (hypertension), ETO (patients who have undergone treatment for alcohol abuse or come in intoxicated), CA (history of malignant disease - cured or not), LIV (history of hepatitis, cholangitis, but not gall bladder disease), CNS (history of brain abscess, encephalitis, or clinical dementia), PCA (history of stroke with or without residual), RHE (history of rheumatic heart disease).

Prior cardiac operation makes the surgery more difficult technically. Six categories are distinguished: OTH (other cardiac surgery), NON (none surgery), CAB (coronary bypass surgery), VAL (valve operation), CON (congenital surgery) and PAC (pacemaker operation). The incidences of last two kinds of operation are lower, 0.2% and 0.6% respectively, so that these are removed as risk factors.

AGE and SEX are two important variables. BSA (body surface area) is calculated from height and weight using a NOMOGRAM formula and can be an important factor.

Although the *prior operation status* is important, from a decision point of view, we do not include it this time.

Post operation variables (outcomes) include 11 complications and the 30 day status. Among these 11 complications, we only consider those which are clinically related with the 30 day status. Hence we remove the following complications: REP (reoperation for bleeding, suspected tamponade), WOU (wound: dehiscence or infection), CLO (clotting: prolonged bleeding problems, low platelets), SEP (septicemia, pneumonia, wound infection, etc), GIB (gastrointestinal bleeding, perforated ulcer, cholecystitis) and DIC (diffuse intravascular coagulation). We study the remaining five complications: REMS (renal shutdown), NEMS (peripheral nerve, central nervous system defect), PUMS (pneumothorax, prolonged respiratory support), MI (intra- or post-operation myocardial infarction by EKG or enzymes) and LOMS (low output syndrome). These response variables were obtained by combining their two levels (mild and severe) into one so that the resulting variables are binary.

The response variable for the 30 day operative mortality was obtained by combining the case of original statuses 2, 3 and 5 into "1".

In Table 4 to Table 6, summary statistics are given as well as some special features such as missing value, etc.

3.2 Odds Ratio Analysis

A natural way to represent the association of a binary risk factor and a binary outcome is the 2x2 contingency table, as follows:

2x2 Contingency Table

	outcome1	outcome2	Sample Size
with risk factor	a	b	n_1
without risk factor	c	d	n_2

We suppose that such a table has been generated by drawing two independent binomial samples of sizes n_1 and n_2, with probabilities for outcome1 being p_1 and p_2 respectively. For example, in our study, an outcome variable is the status after operation, the samples correspond to the patients who have the presence or absence of a risk factor.

The odds ratio in such a table is defined as

$$\Psi = \frac{p_1(1-p_2)}{p_2(1-p_1)}$$

The odds ratio (as well as its logarithm) is widely used as a measure of association in 2x2 contingency tables due to its simple interpretability. For example, if outcome is the presence or absence of lung cancer and the populations are smokers and non-smokers, then $\hat{\Psi} = 2$ indicates that the odds of lung cancer among smokers is twice that among non-smokers in the study population. It has also been pointed out that the odds ratio forms a useful approximation to the relative risk in retrospective studies [Rothman, 1986]. The coefficients estimated in a logistic regression can also be interpreted as log-odds ratios (logarithm of odds ratio).

Table 4: Tabular summary for categorical variables

Variable	Heading	Code	Count	Frequency	Remark
V2	SEX	0	13084	72.0%	1 patient with no sex recorded
		1	5086	28.0%	
V3	PMI	0	7451	41.0%	5436 missing data (30.0%)
		1	5284	29.0%	coded -1
V4	OBE	0	16648	91.7%	
		1	1523	8.3%	
V5	COP	0	15977	88.0%	
		1	2194	12.0%	
V6	DIA	0	15276	84.1%	
		1	2895	15.9%	
V7	CH2	0	16008	88.1%	
		1	2163	11.9%	
V8	CH3	0	17373	95.7%	
		1	798	4.3%	
V9	REN	0	170561	93.9%	
		1	1115	6.1%	
V10	HTN	0	11175	61.5%	
		1	6996	38.5%	
V11	ETO	0	17454	96.1%	
		1	717	3.9%	
V12	DRU	0	18101	99.7%	ignored in future analysis
		1	70	0.3%	
V13	MAR	0	18151	99.9%	ignored in future analysis
		1	20	0.1%	
V14	HIV	0	18171	100.0%	ignored in future analysis
		1	0	0.0%	
V15	AID	0	18171	100.0%	ignored in future analysis
		1	0	0.0%	
V16	CA	0	17860	98.3%	
		1	311	1.7%	
V17	ANE	0	18153	100.0%	ignored in future analysis
		1	18	0.0%	

Table 5: Tabular summary for categorical variables (continued)

Variable	Heading	Code	Count	Frequency	Remark
V18	LIV	0	17946	98.8%	
		1	225	1.2%	
V19	CNS	0	17427	96.0%	
		1	744	4.0%	
V20	PCA	0	17700	97.5%	
		1	471	2.5%	
V21	RHE	0	17545	96.6%	
		1	626	3.4%	
V22	PUL	0	18112	99.7%	ignored in future analysis
		1	59	0.3%	
V23	CHR	0	18160	100.0%	ignored in future analysis
		1	11	0.0%	
V24	OTH	0	17649	97.2%	
		1	522	2.8%	
V25	NON	0	15418	84.8%	ignored in future analysis
		1	2753	15.2%	
V26	CAB	0	16525	91.0%	
		1	1646	9.0%	
V27	VAL	0	17733	97.6%	
		1	438	2.4%	
V28	CON	0	18140	99.8%	ignored in future analysis
		1	31	0.2%	
V29	PAC	0	18055	99.4%	ignored in future analysis
		1	116	0.6%	
V30	LVD	0	8972	49.3%	5320 (29.3%) missing data
		1	1854	10.2%	coded -1
		2	1524	8.3%	
		3	306	1.7%	
		4	195	1.0%	
V31	POS	1	14137	77.8%	433 (2.4%) missing data
		2	1935	10.6%	ignored in future analysis
		3	1501	8.3%	
		4	165	0.9%	
V33	REP	0	17471	96.2%	
		1	700	3.8%	
V34	REM	0	17623	97.0%	combined with RES in future analysis
		1	548	3.0%	to form a new variable REMS

Table 6: Tabular summary for categorical variables(continued)

Variable	Heading	Code	Count	Frequency	Remark
V35	RES	0	17881	98.4%	
		1	290	1.6%	
V36	WOU	0	17964	98.9%	
		1	207	1.1%	
V37	NEM	0	16803	92.5%	combined with NES in future analysis
		1	1368	7.5%	to form a new variable NEMS
V38	NES	0	17768	97.8%	
		1	403	2.2%	
V39	PUM	0	10173	56.0%	combined with PUS in future analysis
		1	7998	44.0%	to form a new variable PUMS
V40	PUS	0	15643	86.1%	
		1	2528	13.9%	
V41	MI	0	17194	94.7%	
		1	977	5.3%	
V42	LOM	0	17389	95.7%	combined with LOS in future analysis
		1	782	4.3%	to form a new variable LOMS
V43	LOS	0	17321	95.4%	
		1	850	4.6%	
V44	CLO	0	17803	98.0%	ignored in future analysis
		1	368	2.0%	
V45	SEP	0	17904	98.6%	ignored in future analysis
		1	267	1.4%	
V46	GIB	0	17415	95.8%	ignored in future analysis
		1	756	4.2%	
V47	DIC	0	18117	99.7%	ignored in future analysis
		1	54	0.3%	
V48	STA	0	10	0.0%	612 (3.4%) missing data
		1	15513	85.4%	only the cases 1, 2, 3 and 5
		2	519	2.9%	are considered. That is:
		3	234	1.3%	15513 (95.1%) alive, coded 0
		4	465	2.6%	794 (4.9%) died, coded 1
		5	41	0.2%	
		6	316	1.7%	
		9	10	0.0%	

There are several estimates of the odds ratio [Walter, 1987], but the most common one is the maximum likelihood estimate (MLE)

$$\hat{\Psi}_{MLE} = \frac{ad}{bc}$$

The derivation of this is simple: for a binomial distributed random variable with parameter p and n, the MLE of p is a/n where n is the sample size and a is the number of "successes". By the invariance principle, the MLE of the odds $p_1/(1\text{-}p_1)$ is a/b. Similarly the MLE of $p_2/(1\text{-}p_2)$ is c/d. Hence the above estimate obtains.

The estimate of odds ratio is more useful as an interval estimate or confidence interval (CI). A brief review of various methods for CI construction is given by Fleiss (1979). We use the result derived by Bishop et al. (1975) in which it is proved $\log\hat{\Psi}$ is asymptotically normal with mean $\log\Psi$ and variance $(n_1p_1(1-p_1))^{-1} + (n_2p_2(1-p_2))^{-1}$. This result follows from an application of the delta method. An estimates variance of $\log\hat{\Psi}$ is $1/a + 1/b + 1/c + 1/d = \widehat{SE}(\log\hat{\Psi})^2$.

So, a 100(1-α)% CI of $\hat{\Psi}$ is

$$\exp\{\log\hat{\Psi} \pm Z_{1-\alpha/2}\widehat{SE}(\log\hat{\Psi})\}$$

where Z_β is the upper β quantile of the standard normal distribution.

In Table 7 to Table 12, the odds ratios for each outcome variable are given.

Statistically, only those 95% CI not containing 1 are more strongly related with the outcome. In Table 13, the risk factors identified by odds ratio are listed.

If the estimated odds ratio is larger than 1, we say the variable is positively related with the outcome; otherwise, we say it negatively related. Nearly all binary explanatory variables (including the groups of other diseases, prior cardiac surgeries) are positively related with the 30 day mortality and complications except CH2 and CH3. Variable CH2 and CH3 measure high cholesterol blood levels. This may lead to increased risk of getting vascular disease in many organs. MI and LOMS are vascular related complications. Unfortunately, CH2 and CH3 have negative association with them so we have some doubt whether the measurement of cholesterol blood level is correct and

Table 7: Odds ratio for STA

Variable	Heading	$\hat{\Psi}$	$\widehat{SE}(\log \hat{\Psi})$	95% CI of Ψ	Remark
V2	SEX	1.17	0.24	(0.73, 1.90)	
V4	OBE	0.38	0.72	(0.09, 1.58)	
V5	COP	1.00	0.38	(0.47, 2.12)	
V6	DIA	0.31	0.72	(0.63, 2.16)	
V7	CH2	0.40	0.42	(0.17, 0.94)	negatively related
V8	CH3	1.25	0.41	(0.56, 2.80)	
V9	REN	1.95	0.38	(0.91, 4.19)	
V10	HTN	1.32	0.23	(0.84, 2.08)	
V11	ETO	1.18	0.60	(0.35, 3.86)	
V16	CA	1.61	0.74	(0.37, 6.95)	
V19	CNS	3.05	0.45	(1.24, 7.44)	positively related
V20	PCA	1.98	0.53	(0.68, 5.69)	
V21	RHE	1.26	0.61	(0.38, 4.14)	
V24	OTH	2.35	0.39	(1.09, 5.10)	positively related
V26	CAB	3.53	0.27	(2.07, 6.05)	positively related
V27	VAL	4.23	0.56	(1.40, 12.8)	positively related

Table 8: Odds ratio for REMS

Variable	Heading	$\hat{\Psi}$	$\widehat{SE}(\log \hat{\Psi})$	95% CI of Ψ	Remark
V2	SEX	1.48	0.21	(0.97, 2.26)	
V4	OBE	1.13	0.37	(0.54, 2.37)	
V5	COP	1.86	0.26	(1.11, 3.11)	positively related
V6	DIA	1.80	0.24	(1.11, 2.91)	positively related
V7	CH2	0.67	0.37	(0.32, 1.40)	
V8	CH3	0.88	0.46	(0.35, 2.20)	
V9	REN	9.78	0.22	(6.30, 15.2)	positively related
V10	HTN	1.75	0.20	(1.18, 2.62)	positively related
V11	ETO	1.34	0.47	(0.53, 3.95)	
V18	LIV	3.58	0.62	(1.04, 12.3)	positively related
V19	CNS	3.40	0.33	(1.75, 6.61)	positively related
V20	PCA	1.33	0.60	(0.40, 4.34)	
V21	RHE	0.88	0.59	(0.27, 2.85)	
V24	OTH	0.90	0.52	(0.32, 2.50)	
V26	CAB	0.98	0.35	(0.49, 1.98)	
V27	VAL	0.64	1.01	(0.08, 4.76)	

Table 9: Odds ratio for NEMS

Variable	Heading	$\hat{\Psi}$	$\widehat{SE}(\log\hat{\Psi})$	95% CI of Ψ	Remark
V2	SEX	1.34	0.15	(0.99, 1.81)	
V4	OBE	2.11	0.21	(1.37, 3.25)	positively related
V5	COP	1.22	0.21	(0.80, 1.84)	
V6	DIA	1.38	0.18	(0.96, 1.98)	
V7	CH2	0.69	0.25	(0.42, 1.14)	
V8	CH3	0.54	0.39	(0.25, 1.17)	
V9	REN	1.62	0.23	(1.02, 2.57)	positively related
V10	HTN	1.49	0.14	(1.12, 1.97)	positively related
V11	ETO	1.40	0.32	(0.73, 2.68)	
V16	CA	2.96	0.43	(1.26, 6.93)	positively related
V18	LIV	2.20	0.55	(0.74, 6.53)	
V19	CNS	3.92	0.24	(2.41, 6.39)	positively related
V20	PCA	2.02	0.37	(0.97, 4.19)	
V21	RHE	0.81	0.43	(0.34, 1.88)	
V24	OTH	0.94	0.35	(0.47, 1.90)	
V26	CAB	1.09	0.24	(0.67, 1.75)	
V27	VAL	0.58	0.73	(0.14, 2.46)	

Table 10: Odds ratio for PUMS

Variable	Heading	$\hat{\Psi}$	$\widehat{SE}(\log\hat{\Psi})$	95% CI of Ψ	Remark
V2	SEX	1.08	0.09	(0.90, 1.30)	
V4	OBE	3.38	0.16	(2.45, 4.66)	positively related
V5	COP	3.45	0.13	(2.66, 4.48)	positively related
V6	DIA	1.91	0.11	(1.52, 2.40)	positively related
V7	CH2	0.36	0.15	(0.26, 0.49)	negatively related
V8	CH3	0.49	0.20	(0.33, 0.74)	negatively related
V9	REN	3.32	0.16	(2.41, 4.58)	positively related
V10	HTN	1.86	0.08	(1.57, 2.20)	positively related
V11	ETO	2.21	0.21	(1.46, 3.36)	positively related
V16	CA	0.56	0.41	(0.25, 1.25)	
V19	CNS	2.12	0.21	(1.39, 3.21)	positively related
V20	PCA	0.67	0.29	(0.37, 0.91)	
V21	RHE	0.97	0.23	(0.61, 1.54)	
V24	OTH	0.58	0.22	(0.37, 0.91)	
V26	CAB	1.22	0.14	(0.92, 1.61)	
V27	VAL	1.03	0.34	(0.52, 2.01)	

Table 11: Odds ratio for MI

Variable	Heading	$\hat{\Psi}$	$\widehat{SE}(\log\hat{\Psi})$	95% CI of Ψ	Remark
V2	SEX	0.76	0.19	(0.52, 1.11)	
V4	OBE	1.70	0.26	(1.01, 2.85)	positively related
V5	COP	2.11	0.20	(1.40, 3.16)	positively related
V6	DIA	1.13	0.22	(0.73, 1.75)	
V7	CH2	0.39	0.36	(0.19, 0.80)	negatively related
V8	CH3	0.75	0.39	(0.34, 1.63)	
V9	REN	1.80	0.25	(1.09, 2.99)	positively related
V10	HTN	1.58	0.16	(1.14, 2.18)	positively related
V11	ETO	1.75	0.34	(0.89, 3.44)	
V18	LIV	6.43	0.46	(2.60, 15.86)	positively related
V19	CNS	1.35	0.37	(0.64, 2.83)	
V20	PCA	2.06	0.41	(0.92, 4.64)	
V21	RHE	1.12	0.43	(0.48, 2.61)	
V24	OTH	0.99	0.39	(0.45, 2.16)	
V26	CAB	1.28	0.26	(0.76, 2.13)	
V27	VAL	1.73	0.53	(0.60, 4.95)	

Table 12: Odds ratio for LOMS

Variable	Heading	$\hat{\Psi}$	$\widehat{SE}(\log\hat{\Psi})$	95% CI of Ψ	Remark
V2	SEX	1.69	0.15	(1.24, 2.32)	positively related
V4	OBE	1.08	0.28	(0.61, 1.91)	
V5	COP	1.06	0.23	(0.66, 1.68)	
V6	DIA	1.27	0.20	(0.85, 1.89)	
V7	CH2	1.20	0.22	(0.77, 1.89)	
V8	CH3	0.86	1.00	(0.01, 0.62)	negatively related
V9	REN	1.99	0.23	(1.24, 3.17)	positively related
V10	HTN	1.31	0.15	(0.96, 1.78)	
V11	ETO	1.01	0.40	(0.46, 2.21)	
V16	CA	2.33	0.49	(0.88, 6.13)	
V18	LIV	1.18	0.74	(0.27, 5.09)	
V19	CNS	1.17	0.37	(0.56, 2.46)	
V20	PCA	1.22	0.47	(0.48, 3.10)	
V21	RHE	2.20	0.32	(1.17, 4.15)	positively related
V24	OTH	1.60	0.31	(0.86, 2.99)	
V26	CAB	1.84	0.22	(1.19, 2.84)	positively related
V27	VAL	1.09	0.60	(0.33, 3.60)	

Table 13: Binary risk factors identified from odds ratio

Outcome	Binary Risk Factors					
STA	VAL (4.23)	CAB (3.53)	CNS (3.05)	OTH (2.35)		
REMS	REN (9.78)	LIV (3.58)	CNS (3.40)	COP (1.86)	DIA (1.80)	HTN (1.75)
NEMS	CNS (3.93)	CA (2.96)	OBE (2.11)	REN (1.62)	HTN (1.49)	
PUMS	COP (3.45)	OBE (3.38)	ETO (2.23)	CNS (2.12)	DIA (1.91)	HTN (1.36)
MI	LIV (6.43)	COP (2.11)	REN (1.80)	OBE (1.70)	HTN (1.58)	
LOMS	RHE (2.20)	REN (1.99)	CAB (1.84)	SEX (1.69)		

Table 14: Two-way table for PMI and LVD

	prior MI				
LVD	missing	no	yes	total	percent
missing	239	70	65	374	-
normal	59	413	183	655	30%
40-49%	3	61	78	142	56%
30-39%	40	24	40	104	63%
20-29%	0	5	14	19	74%
$\leq$ 20%	7	1	5	13	83%
total	384	574	385	1307	40%

consequently, we did not include CH2 and CH3 in our model building procedure. From a medical point of view, cholesterol blood level maybe a surrogate for nutritional level.

PMI indicates the health of the heart muscle while LVD measure the left ventricular function. So they essentially measure the same phenomenon although PMI is much less specific. In Table 14, their associations can be seen by a two-way table. The last column is the percent of YES among non-missing cases.

As we see, compared with the percentage in the normal category, as the left ventricular function is getting worse, the percentage of patients who had prior myocardial infarction is increasing. This confirms our knowledge about these two variables.

AGE and BSA are the only two continuous variables in the data set. Biologically, they maybe related with many other variables, but the most interesting ones (based on previous studies) are the following: the association of diabetes with AGE and BSA; the relationship between BSA and

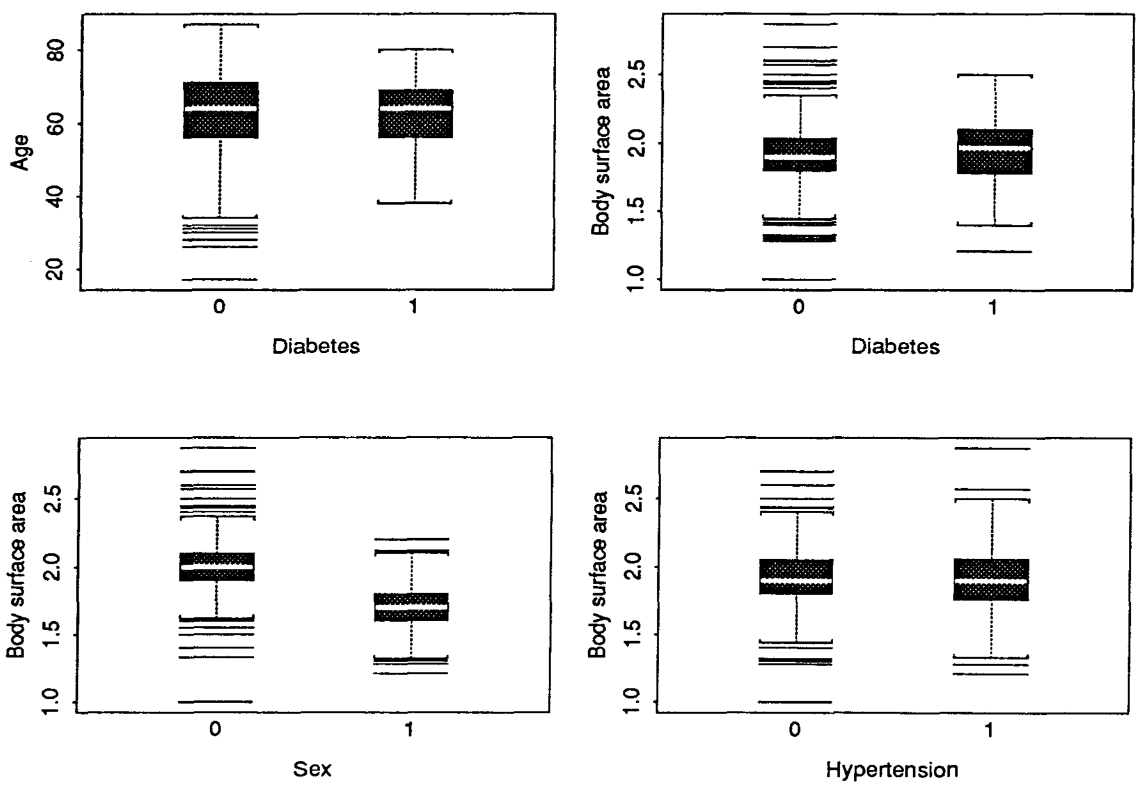


Figure 1: Boxplots for continuous variables AGE and BSA

gender as well as hypertension. In Figure 1, these relations are displayed by boxplots.

Boxplots have proven to be quite a good exploratory tool, especially when several boxplots are placed side by side for comparison as in the current cases. The most striking visual feature is the box which shows the limits of the middle half of the data (the white line inside the box represents the median and the ends of the box represent the lower and upper quartiles). The first horizontal lines beyond the box (which are called the whiskers) are drawn to the nearest value not beyond a standard span from the quartiles. Points beyond, which may be outliers, are drawn individually. The standard span is 1.5 times the difference of the upper and lower quartiles. [Hoaglin et al., 1983]

There is little difference in the distribution of age between the populations of patients with or without diabetes. Similarly, this holds for the distribution of body surface area with respect to diabetes and hypertension. Only the relation between gender and body surface area appears to be significant. Female patients tend to have small body surface area. Although this is a general truth, notice that by odds ratio analysis that female patients have higher risk of operative mortality. We need to investigate this relation further as some cardiologists believe that gender is a poor surrogate for body surface area.

Chapter 4

Logistic Regression Analysis

The methodology of logistic regression analysis has become extremely popular among biostatisticians in recent years, see for example Lemeshow et al. (1988).

Let y_i, $i = 1, \ldots, n$, be independent binary random variables. The logistic regression is a method for assessing the dependence of $\mu_i = \Pr(y_i = 1)$ on explanatory variables x_i. The dependence is postulated as

$$\mu_i = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}},$$

$$1 - \mu_i = \frac{1}{1 + e^{x_i^T \beta}},$$

for $i = 1, \ldots, n$, where $x_i{}^T = (x_{i_1}, \ldots, x_{i_p})$ is a row of known constants and $\beta = (\beta_1, \ldots, \beta_p)^T$ is a column of unknown parameters. The equations above are equivalent to

$$g(\mu_i) = x_i^T \beta$$

Then, $g(\mu) = \log(\mu/(1-\mu))$ is called the logistic transformation of the probability $\mu = (\mu_1, \ldots, \mu_n)$ and above equation is called a linear logistic model.

There are several ways to estimate the logistic parameters β [Hosmer and Lemeshow, 1989]. The maximum likelihood procedure is based on the conditional likelihood

$$L(\mu; \mathbf{y}) = \prod_{i=1}^{n} f(y_i; \mu_i | x_i)$$

where $f(y; \mu|x) = \mu^y(1-\mu)^{1-y}$. and $\mu = (\mu_1, \ldots, \mu_n)$, $\mathbf{y} = (y_1, \ldots, y_n)$ It is convenient to deal with the log-likelihood function. In our case, it is

$$l(\beta; \mathbf{y}) = \sum_{i=1}^{n} \{y_i x_i^T \beta - \log[1 + \exp(x_i^T \beta)]\}$$

To compute the maximum likelihood estimates, it is necessary to solve the score equations $\partial l(\beta; \mathbf{y})/\partial\beta{=}0$. Commonly, the Newton-Raphson method or the iteratively reweighted least squares method is used to calculate $\hat{\beta}$, the estimate of β [Rao, 1973].

Our goal is to use logistic regression to develop an objective model for prediction of 30 day operative mortality and complications among patients. Typically, the first step in this modeling process is data reduction; from all available predictor variables, only those most associated with outcome are selected for inclusion in the final model. If, after this step, there are still a large set of characteristics, a stepwise logistic regression analysis can be applied to reduce the number of predictor variables. An alternative method is the best subsets selection procedure which provides several candidate models.

4.1 Univariate Analysis and Comparison of Models

For continuous variables, the test of association of the outcome and the independent variable can be carried out using Student-t test analogous in linear regression [Weisberg, 1980]. For categorical variables, we use the likelihood ratio test which is defined as follows. The deviance function is defined as

$$D(\mu; \mathbf{y}) = 2\log L(\mathbf{y}; \mathbf{y}) - 2\log L(\mu; \mathbf{y})$$

The difference in deviance between two models measures the contribution of the parameters by which they differ. The distribution theory is asymptotic [McCullagh and Nelder, 1989]; for comparing 2 nested models with estimated mean $\hat{\mu}_1$ and $\hat{\mu}_2$, the difference in deviance

$$D(\hat{\mu}_1, \hat{\mu}_2) = D(\hat{\mu}_1; y) - D(\hat{\mu}_2; y)$$

has an asymptotic χ^2 distribution (under the null hyperthesis that the smaller model is correct) with degrees of freedom $\nu = \nu_1 - \nu_2$ equal to the difference in the dimensions of the parameter spaces implicit in the models with mean μ_1 and μ_2. Therefore, to test the association of a single variable x to the outcome, we only need to compare the model

$$g(\mu_{1i}) = \beta_0$$

with the model

$$g(\mu_{2i}) = \beta_0 + \beta_1 x_i$$

and find out how much the variable x improves the predictive value of the model.

4.2 Stepwise Logistic Regression

In stepwise logistic regression, models are built by adding in new variables and seeing how much they improve the fit, and by dropping variables that do not improve the fit by a "significant" amount. Usually the procedure starts with an arbitrary model and stops when no step will decrease the value of a selection criterion. The selection criterion used here is AIC (Akaike's Information Criterion) [Akaike, 1973]

$$AIC = D + 2p$$

where D is the deviance of the current model, p the dimension (number of variables) in the model. The changes in AIC due to augmenting or reducing a model by a given variable reflect both the change in deviance caused by the step, as well as the change in the dimension of the model. The rationale of AIC is that the more parameters a model contains, the less accurately they can be estimated and the predictive value of the model may get worse. AIC adjusts the deviance for the

Table 15: Stepwise regression procedure: a demonstration

Variables involved in the current model	operation	AIC
AGE, SEX, PMI, OBE, COP, ETO, CA, CNS, PCA, OTH, CAB, LVD, BSA		479.8
AGE, PMI, OBE, COP, ETO, CA, CNS, PCA, OTH, CAB, LVD, BSA	-SEX	477.9
AGE, PMI, COP, ETO, CA, CNS, PCA, OTH, CAB, LVD, BSA	-OBE	475.9
AGE, PMI, COP, CA, CNS, PCA, OTH, CAB, LVD, BSA	-ETO	473.5
AGE, PMI, CA, CNS, PCA, OTH, CAB, LVD, BSA	-COP	472.2
AGE, PMI, CNS, PCA, OTH, CAB, LVD, BSA	- CA	471.5

Table 16: Results of stepwise regression procedure for each outcome

Outcome	variable involved	AIC
STA	AGE, PMI, CNS, PCA, OTH, CAB, LVD, BSA	471.56
REMS	AGE, SEX, PMI, COP, DIA, REN, CNS, LVD	437.92
NEMS	AGE, PMI, OBE, ETO, CA, CNS	744.55
PUMS	PMI, DIA, HTN, CNS, BSA	1105.37
MI	PMI, LIV, PCA	684.66
LOMS	AGE, PMI, DIA, REN, RHE, CAB, LVD, BSA	784.16

number of parameters estimated. Thus, the model with the minimum AIC gives the best fit to the data according to the AIC criterion. Therefore, we think of AIC as a useful tool for the quick comparison of parametric models although it does not indicate that the better of two models is "significantly better".

Take STA as example. The initial model contains the 14 variables obtained from univariate screening. The first variable deleted was SEX leading to AIC=477.90; the second one deleted was OBE leading to AIC=475.97; . . .; the last one deleted was CA leading to AIC=471.56. The procedure is summarized in Table 15.

In Table 16, the results of stepwise logistic regression for various outcome variables are given with the corresponding best AIC values.

4.3 Best Subsets Selection

The best subsets selection is an alternative to the stepwise procedure for model building. This approach has been available for linear regression for years and makes use of the branch and bound algorithm of Furnival and Wilson (1974). Typical software implementing this method will identify a specified number of "best" models containing one, two, three variables, and so on, up to the single model containing all p variables. For the case of logistic regression, Hosmer and Lemeshow (1989) proposed a method which can be performed in a striaightforward manner using any program for the best subsets linear regression.

The best subsets selection procedure is regarded as a more reliable and informative method. This is because the the stepwise procedure lead to a single subset of variables and does not suggest alternative good subsets. In this procedure, C_p statistics are used for selecting the best subsets [Draper and Smith, 1981]; a model with a C_p value close to the number of predictors is better.

In the logistic model, let $\hat{\beta}$ be the maximum likelihood estimate and $\hat{\pi}_i$ be the estimated logistic probability computed using $\hat{\beta}$ and the data for the ith case, x_i. We define two matrix $\mathbf{X}$ and $\mathbf{V}$

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}$$

and

$$V = \begin{pmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \dots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{\pi}_n(1-\hat{\pi}_n) \end{pmatrix}$$

It may be shown [Pregibon, 1981] that $\hat{\beta} = (X'VX)^{-1}X'V\mathbf{z}$, where the vector $\mathbf{z}$ contains pseudovalues, $\mathbf{z} = X\hat{\beta} + V^{-1}\mathbf{r}$, and $\mathbf{r}$ is the vector of residuals, $\mathbf{r} = (\mathbf{y} - \hat{\pi})$. A computation for the best subsets logistic regression model can be performed using a best subsets linear regression program the

Table 17: Models obtained from best subsets procedure for each outcome

Outcome	Model Code	Variable included	C_p
STA	S_1	AGE, PMI, CNS, PCA, OTH, CAB, LVD, BSA	7.02
	S_2	AGE, CA, CNS, PCA, OTH, CAB, LVD, BSA	6.4
	S_3	AGE, COP, CNS, PCA, OTH, CAB, LVD, BSA	8.0
REMS	R_1	AGE, SEX, PMI,COP, DIA, REN, LIV, CNS	7.1
	R_2	AGE, SEX, PMI, COP, DIA, REN, CNS, LVD	8.4
	R_3	AGE, SEX, PMI, COP, DIA, REN, HTN, CNS	8.6
NEMS	N_1	AGE, PMI, OBE, ETO, CA, CNS	4.4
	N_2	AGE, PMI, ETO, CA, CNS, CAB	4.6
	N_3	AGE, PMI, ETO, CA, CNS, PCA	5.8
PUMS	P_1	AGE, PMI, REN, HTN, CNS, BSA	6.7
	P_2	PMI, OBE, REN, HTN, CNS, BSA	7.5
	P_3	AGE, PMI,OBE, HTN, CNS, BSA	7.9
MI	M_1	PMI ,COP, LIV, PCA	3.9
	M_2	PMI, COP, CNS, PCA	5.1
	M_3	PMI, LIV, CNS, PCA	5.3
LOMS	L_1	AGE, PMI, DIA, REN, RHE, OTH, CAB, LVD, BSA	11.0
	L_2	AGE, PMI, DIA, HTN, RHE, OTH, CAB, LVD, BSA	11.5
	L_3	AGE, SEX, PMI, DIA, RHE, OTH, CAB, LVD, BSA	12.2

dependent variable $\mathbf{z}$, case weights v_i, equal to the diagonal elements of $\mathbf{V}$, and original covariates $\mathbf{x}$.

In this study, for each outcome, we provide three candidate models produced by the best subsets selection procedure. One interesting finding is that the model obtained by stepwise procedure is among the three models.

4.4 Goodness-of-fit: Hosmer-Lemeshow Grouping Test

After the above procedures, we would like to know how effective the models we have are in describing the outcome variables. This is referred to as its *goodness-of-fit.*

One test was proposed by Hosmer and Lemeshow (1980). The Hosmer-Lemeshow grouping test creates groups based on the values of the estimated probabilities. Suppose we have n observations. With this method, use of $g = 10$ groups results in the first group containing the $n_1 = n/10$ subjects

Table 18: Hosmer-Lemeshow grouping test for selected models

		Decile of Risk												
Model		g1	g2	g3	g4	g5	g6	g7	g8	g9	g10	Total	$\hat{C}$	Prob
S_2	Obs	0	0	2	3	4	4	5	12	15	20	65	7.25	0.51
	Exp	0.7	1.4	2.0	2.6	3.3	4.3	5.7	7.7	11.4	25.8	65		
R_1	Obs	1	2	1	1	3	3	7	6	8	29	61	3.75	0.88
	Exp	0.5	1.1	1.6	2.2	2.8	3.6	4.9	6.9	9.8	27.3	61		
N_3	Obs	1	3	3	6	10	9	10	15	22	42	121	1.82	0.98
	Exp	1.2	2.6	4.1	5.8	7.7	9.8	12.2	15.9	21.1	40.7	121		
P_2	Obs	9	10	11	12	18	19	21	18	34	44	196	3.76	0.87
	Exp	9.9	11.2	12.1	13.2	14.5	16.8	19.4	23.4	31.0	45.2	196		
M_2	Obs	1	4	2	3	13	9	6	18	22	24	102	8.29	0.41
	Exp	3.1	3.1	3.1	3.2	7.5	9.1	9.0	16.3	20.4	27.2	102		
L_2	Obs	8	4	4	4	11	7	9	14	21	35	117	13.9	0.08
	Exp	2.9	4.6	5.8	7.0	8.4	10.1	11.9	14.4	18.9	33.0	117		

having the smallest estimated probabilities, and the last group containing the $n_{10} = n/10$ subjects having the largest estimated probabilities. The Hosmer-Lemeshow goodness-of-fit statistics $\hat{C}$ is obtained by calculating

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n_k \bar{\pi}_k)^2}{\bar{\pi}_k(1 - \bar{\pi}_k)},$$

where

$$o_k = \sum_{j \in A_k} y_j$$

and

$$\bar{\pi}_k = \sum_{j \in A_k} \hat{\pi}_j / n_k.$$

With A_k consisting of subjects in the k^{th} group. It can be shown that $\hat{C}$ is asymptotically well approximated by the chi-square distribution with $g - 2$ degrees of freedom, $\chi^2(g-2)$, if the model is correct.

A small value of $\hat{C}$ indicates a good fit. From the prediction point of view, we used this statistic as the final criterion for model selection. That is, among the three candidate models obtained from stepwise logistic regression and the best subsets selection procedure, we chose the one with smallest value of $\hat{C}$. In Table 18, the grouping tests for each selected model are shown.

Judging from the p-value, all the selected models fit quite well except possibly for the one with LOMS as outcome.

The final logistic regression models were given in Tables 19 to 24 together with the maximum estimated probability which was calculated by putting the higher value for all the risk factors in the model (for continuous variable, we use their mean values). This number indicates the range of probability that a model can predict. Since OTH and CAB cannot be 1 at same time, when these two appear together in the model, we use the one with larger coefficient. All the results are obtained using version 3.1 of the statistical software S/Splus. When coding dummy variables, treatment contrast was used [Chamber and Hastie, 1990].

As mentioned in section 3.2 the estimated coefficients here can be interpreted as log-odds ratio. We simply calculate $\exp(\hat{\beta})$ to give an odds ratio of each predictor with other factors held fixed. For example, in the STA model, variable OTH has a coefficient 1.77 which gives exp(1.77)=5.87. This means that the patient who had an other cardiac operation are 5.87 times more likely to have a mortality than those who had not. Another example is that AGE leads to an odds ratio of exp(0.048)=1.049. This means an additional multiplicative risk of 1.049 for each increase in age of one year, all other variables held fixed. Since this number larger than 1, we consider age is a contributor to operative mortality; older patients tend to have higher risk. For the categorical variable, the odds ratios should be interpreted as a comparsion to the first category. In both of the categorical variables PMI and LVD, the first category happens to represent the missing value and such a comparison can provide some insights to the missing value category.

Table 19: Final logistic regression model for STA

Variable	Heading	$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\beta}/SE(\hat{\beta})$	$\exp(\hat{\beta})$
Constant		-2.811	1.753	-1.603	
V24	OTH	1.770	0.444	3.981	5.871
V26	CAB	1.327	0.344	3.856	3.770
V32	BSA	-2.061	0.654	-3.148	0.127
V1	AGE	0.048	0.015	3.108	1.049
V30	LVD0	-0.092	0.333	-0.277	0.911
	LVD1	-0.430	0.541	-0.794	0.650
	LVD2	0.973	0.420	2.314	2.648
	LVD3	1.995	0.665	2.997	7.357
	LVD4	-3.914	5.808	-0.673	0.019
V19	CNS	1.160	0.496	2.337	3.190
V20	PCA	1.227	0.608	2.017	3.411
V16	CA	0.868	0.745	1.164	2.382
Maximum prediction probability: 0.97					

Table 20: Final logistic regression model for REMS

Variable	Heading	$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\beta}/SE(\hat{\beta})$	$\exp(\hat{\beta})$
Constant		-9.004	1.186	-7.586	
V9	REN	1.855	0.335	5.523	6.391
V1	AGE	0.065	0.016	4.020	1.067
V5	COP	1.032	0.345	2.988	2.806
V3	PMI1	0.597	0.422	1.415	1.817
	PMI2	1.146	0.407	2.813	3.146
V19	CNS	1.162	0.494	2.350	3.196
V2	SEX	0.577	0.288	2.000	1.780
V6	DIA	0.668	0.342	1.951	1.950
V18	LIV	0.836	0.935	0.893	2.307
Maximum prediction probability: 0.92					

Table 21: Final logistic regression model for NEMS

Variable	Heading	$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\beta}/SE(\hat{\beta})$	$\exp(\hat{\beta})$
Constant		-7.934	0.846	-9.377	
V1	AGE	0.089	0.012	7.362	1.093
V19	CNS	1.292	0.356	3.625	3.640
V3	PMI1	-0.811	0.241	-3.360	0.444
	PMI2	-0.861	0.261	-3.301	0.422
V16	CA	1.662	0.553	3.002	5.270
V11	ETO	0.591	0.415	1.425	1.807
V20	PCA	0.606	0.522	1.161	1.833
Maximum prediction probability: 0.76					

Table 22: Final logistic regression model for PUMS

Variable	Heading	$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\beta}/SE(\hat{\beta})$	$\exp(\hat{\beta})$
Constant		-0.720	0.690	-1.043	
V3	PMI1	-0.442	0.097	-4.530	0.642
	PMI2	-0.097	0.060	-1.620	0.907
V10	HTN	0.451	0.162	2.773	1.569
V32	BSA	-0.737	0.362	-2.037	0.478
V19	CNS	0.554	0.338	1.635	1.740
V9	REN	0.329	0.248	1.326	1.389
V4	OBE	0.286	0.269	1.063	1.331
Maximum prediction probability: 0.40					

Table 23: Final logistic regression model for MI

Variable	Heading	$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\beta}/SE(\hat{\beta})$	$\exp(\hat{\beta})$
Constant		-1.820	0.172	-10.53	
V3	PMI1	-1.991	0.303	-6.567	0.136
	PMI2	-0.899	0.248	-3.619	0.406
V20	PCA	1.469	0.524	2.801	4.348
V19	CNS	0.428	0.412	1.039	1.534
V5	COP	0.285	0.276	1.029	1.329
Maximum prediction probability: 0.59					

Table 24: Final logistic regression model for LOMS

Variable	Heading	$\hat{\beta}$	$SE(\hat{\beta})$	$\hat{\beta}/SE(\hat{\beta})$	$\exp(\hat{\beta})$
Constant		-1.916	1.265	-1.514	
V32	BSA	-1.603	0.485	-3.300	0.201
V3	PMI1	0.723	0.341	2.121	2.062
	PMI2	1.015	0.332	3.049	2.760
V26	CAB	0.838	0.0.293	2.855	2.312
V30	LVD0	-0.244	0.285	-0.854	0.783
	LVD1	-0.204	0.373	-0.548	0.814
	LVD2	0.608	0.355	1.711	1.838
	LVD3	1.214	0.553	2.193	3.367
	LVD4	1.338	0.817	1.638	3.813
V24	OTH	0.879	0.392	2.243	2.410
V21	RHE	1.000	0.454	2.200	2.720
V1	AGE	0.022	0.010	2.129	1.022
V10	HTN	0.315	0.209	1.503	1.370
V6	DIA	0.467	0.265	1.759	1.596
Maximum prediction probability: 0.93					

Chapter 5

The Tree-based Model

The tree-based model has gradually become a popular tool in clinical and epidemiological studies because of its clinical interpretability. The technique was introduced by Morgan et al. (1964), however, more ground-breaking ideas were introduced by Breiman et al. (1984) and the resulting computer program is named CART (Classification And Regression Tree).

The tree-based model procedure used in version 3.1 of S/Splus departs slightly from CART in the recursive partitioning (RP) method proposed by Ciampi et al. (1987). Also, compared with CART, the procedure is far less automatic in tree building, as the unbounding of procedures for growing, displaying and challenging trees requires user initiation in all phases.

5.1 Recursive Partitioning: Growing a Classification Tree

In general, the tree-based model is fitted by creating binary tree using a RP algorithm. The data have the form $(y^{(i)}, x^{(i)})$, $i = 1, \ldots, N$, where y is a multinomial distributed variable with s categories and $\mathbf{x}$ is assumed to be vector of categorical variables $\mathbf{x}=(x_1, \ldots, x_k)$ and for each j, x_j has a finite number of categories $l_1, \ldots, l_{m_j}$. The categories of x_j can be either ordered or unordered.

In what follows, we refer to y as criterion variable and to the components of $\mathbf{x}$ as predictor

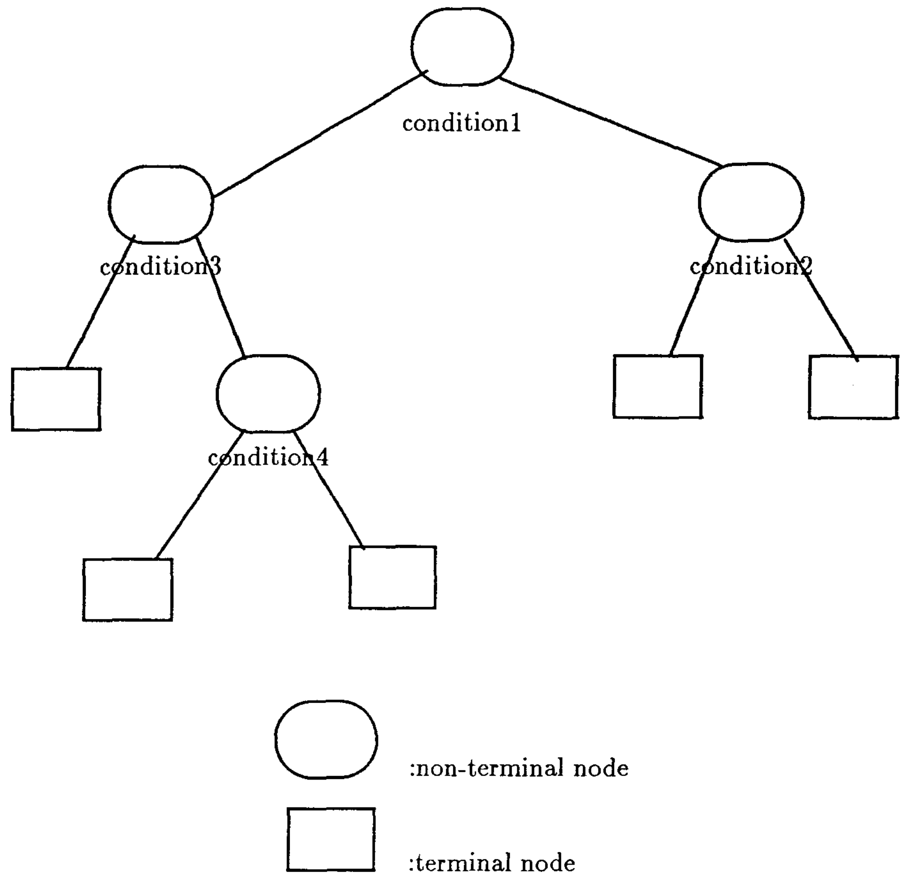


Figure 2: Binary tree: an example

variables. Predictors contain background information used to define strata which are homogeneous according to a criterion variable; for each homogeneous stratum one can define a unique criterion quantity independent of the $\mathbf{x}$ variables given the stratum.

In our study, the criterion quantity is the vector of probabilities of being assigned to each outcomes, i.e., $\mu = (p_1, \ldots, p_s)$ such that $\sum p_s = 1$.

Our aim is to grow a binary tree with nodes representing subsets of observations. In particular the root of the tree represents the entire set of observations and the terminal nodes represent strata that are more homogeneous (see Figure 2).

The tree is constructed based on a set of Split Defining Statements (SDS) such as $x_j \epsilon A_j$, where

A_j is a subset of the m_j categories of x_j. For x_j unordered, A_j can be any of the 2^{m_j-1} nontrivial subsets of $l_1, \ldots, l_{m_j}$; for x_j ordered, A_j can be any of the m_j-1 subsets of the form $A_j=[l_1,l]$, $l = l_2, \ldots, l_{m_j}$.

In an RP tree, each nonterminal node is split by a SDS into two nodes which represent subsets as dissimilar as possible from the point of view of the criterion quantity.

Ciampi et al. (1987) applied the likelihood ratio statistic (LRS) as a natural measure of dissimilarity as follows. Let P_1, P_2 be disjoint sets and let P denote their union. We shall assume that the criterion quantity is represented by a parameter θ which may take different values θ_1, θ_2 for P_1 and P_2 and that likelihood functions $L_1(\theta_1), L_2(\theta_2)$ can be defined for P_1, P_2. We shall also assume that the likelihood function for P is of the form:

$$L(\theta_1, \theta_2) = L_1(\theta_1)L_2(\theta_2)$$

Consider now the hypothesis

$$H_0 : \theta_1 = \theta_2 = \theta$$

and the alternative

$$H : \theta_1 \neq \theta_2$$

Then the log LRS of H versus H_0 is defined as:

$$P(H|H_0) = 2\log[L_1(\hat{\theta_1})L_2(\hat{\theta_1})]/[L_1(\hat{\theta})L_2(\hat{\theta})]$$

where $\hat{\theta_1}$, $\hat{\theta_2}$ are the MLEs of θ_1 and θ_2 under H and $\hat{\theta}$ is the MLE of θ underH_0. Clearly, the larger $p(H|H_0)$ is, the greater is the evidence in the data that P_1 and P_2 are heterogeneous with respect to the criterion quantity. It therefore provides a reasonable and general measure of dissimilarity:

$$d(P_1. P_2) = P(H|H_0)$$

In our case, the criterion quantity is $\mu = (p_1, \ldots, p_s)$ denotes the probability that y falls into each of the possible categories.

We have for a given subset or node:

$$L(\mu;\mathbf{y}) = \prod_{j=1}^{s} p_j^{n_j}$$

where n_j is the number of individuals at the node fall into jth category and $\mathbf{y} = (y_1, \ldots, y_n)$.

At a given node, let M be the observations in the node and n_j be the number of $y_i's$ falling into jth category, then the maximum likelihood estimate of μ is

$$\hat{\mu} = (\frac{n_1}{M}, \ldots, \frac{n_s}{M}).$$

A description of the RP algorithm is:

Denote by $N = \{N_1, \ldots, N_r\}$ the current collection of nodes.

(1) To initialize, set r=1 and let N_1 represent the observations

(2) For every $N_j \in N$ and every split P_{1_j},P_{2_j} defined by an element of SDS, compute $d(P_{1_j},P_{2_j})$.

(3) Among all nodes chose the node N_i^* corresponding to the split P_1^*,P_2^* with largest dissimilarity and replace N_i^* by two nodes representing $P_{1_j}^*$ and $P_{2_j}^*$. Use the resulting collection of nodes as current and go to (2) where r has increased by 1.

In the tree-based model in S/Splus, an intuitive way is used to implement above algorithm. A deviance of a node is defined

$$D(\mu;\mathbf{y}) = -2l(\mu;\mathbf{y})$$

where $l(\mu;\mathbf{y}) = \log L(\mu;\mathbf{y})$. It can be shown that the deviance is identically zero if all the y's are the same, and increases as the y's deviate from this case. The deviance $D_T(\mathbf{y})$ of a tree T is defined as the sum of deviance of all its terminal nodes, $\sum_{t\in T} D(\hat{\mu}_t;\mathbf{y})$, where $\hat{\mu}_t$ is the vector of the observed proportions of the s categories for node t. Splitting proceeds by comparing the deviance of the tree T, with that of larger trees T' in which a terminal node of T has been split into two. The split that maximizes the change in deviance

$$\Delta D = D_T(\mathbf{y}) - D_{T'}(\mathbf{y})$$

is the next split that is chosen.

5.2 Getting the Right Size Tree: Pruning the Classification Tree

The above discussion implies that nodes become more and more pure (homogeneous) as splitting progresses. In the limit, a tree can have as many terminal nodes as there are observations. In S/Splus, two thresholds are introduced to stop the splitting process;

(a). the node deviance is less than some small fraction of the root node deviance (say 1%); and

(b). the node is smaller than some absolute minimum size (say 10).

This also introduces another problem: if the threshold is set too high, good splits may be lost. There are two ways out of this dilemma: one is to use new (independent) data to guide the selection of the right size tree, and the other is to reuse the existing data by the method of cross-validation. In this case, S/Splus provides a function called "prune".

The idea of pruning is more easily described by tree terminology:

Notation 1. A binary tree is denoted by $\tilde{T}$. A node t on the tree $\tilde{T}$ is denoted by $t \in \tilde{T}$.

Definition 1: A branch $\tilde{T}_t$ of $\tilde{T}$ with node $t \in \tilde{T}$ consists of the node t and all descendents of t in $\tilde{T}$.

Definition 2: Pruning a branch $\tilde{T}_t$ from a tree $\tilde{T}$ involves cutting off $\tilde{T}_t$ just below the node t. The resulting tree is a subtree of $\tilde{T}$ denoted by $\tilde{T}$-$\tilde{T}_t$.

Definition 3: $\tilde{T}'$ is a pruned subtree of $\tilde{T}$ if $\tilde{T}'$ is obtained by successively pruning off the branches of $\tilde{T}$.

In S/Splus, the importance of a pruned subtree $\tilde{T}'$ is captured by the cost-complexity measure

$$D_\alpha(\tilde{T}') = D(\tilde{T}') + \alpha * \text{size}(\tilde{T}')$$

where $D(\tilde{T}')$ is the deviance of the subtree, size($\tilde{T}'$) is the number of terminal nodes of $\tilde{T}'$ and α is the cost-complexity parameter. For any specified α, cost-complexity pruning determines the subtree $\tilde{T}'$ that minimizes $D_\alpha(\tilde{T}')$ over all subtrees of $\tilde{T}'$.

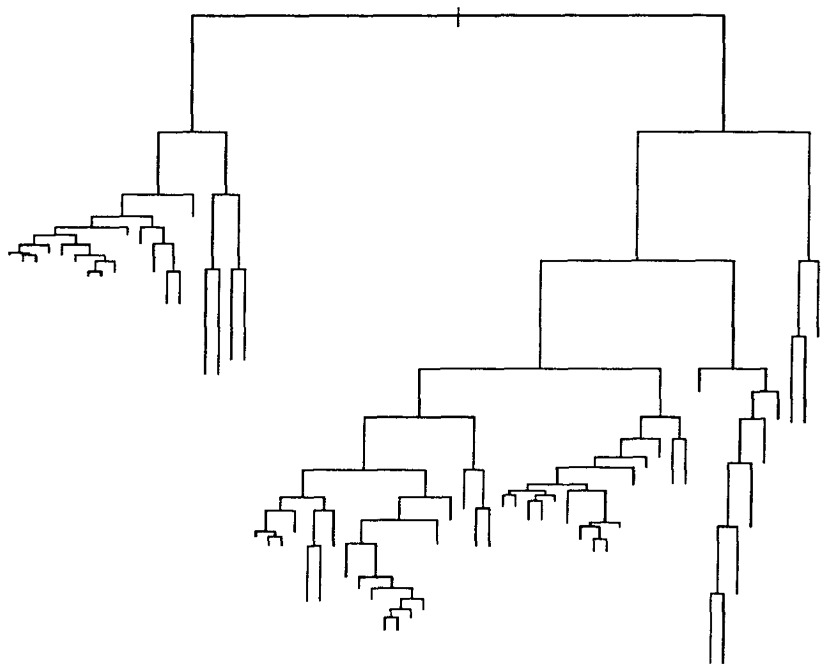


Figure 3: Original tree for STA.

As is known from the RP algorithm, the deviance of a tree $\tilde{T}$ is smaller than that of subtrees when α is set to zero. But when taking the size of tree into consideration, that is, α >0, pruning provides us an upward way to snip off the least important branches. In the extreme case, only the root node is left if α is set sufficiently large. A sequence of subtrees $\tilde{T} = \tilde{T}_0 \succ \tilde{T}_1 \succ \ldots \succ \tilde{T}_k$=root with decreasing size can be obtained while setting an increasing number of values of $\alpha : \alpha_0 = 0 \prec \alpha_1 \prec \ldots \prec \alpha_k$.

5.3 Applications and Results

For each outcome, the described recursive partitioning procedure was performed on a sample data set. The original tree underwent a cross-validation testing on a new data set by the pruning algorithm and the right size of the trees was decided.

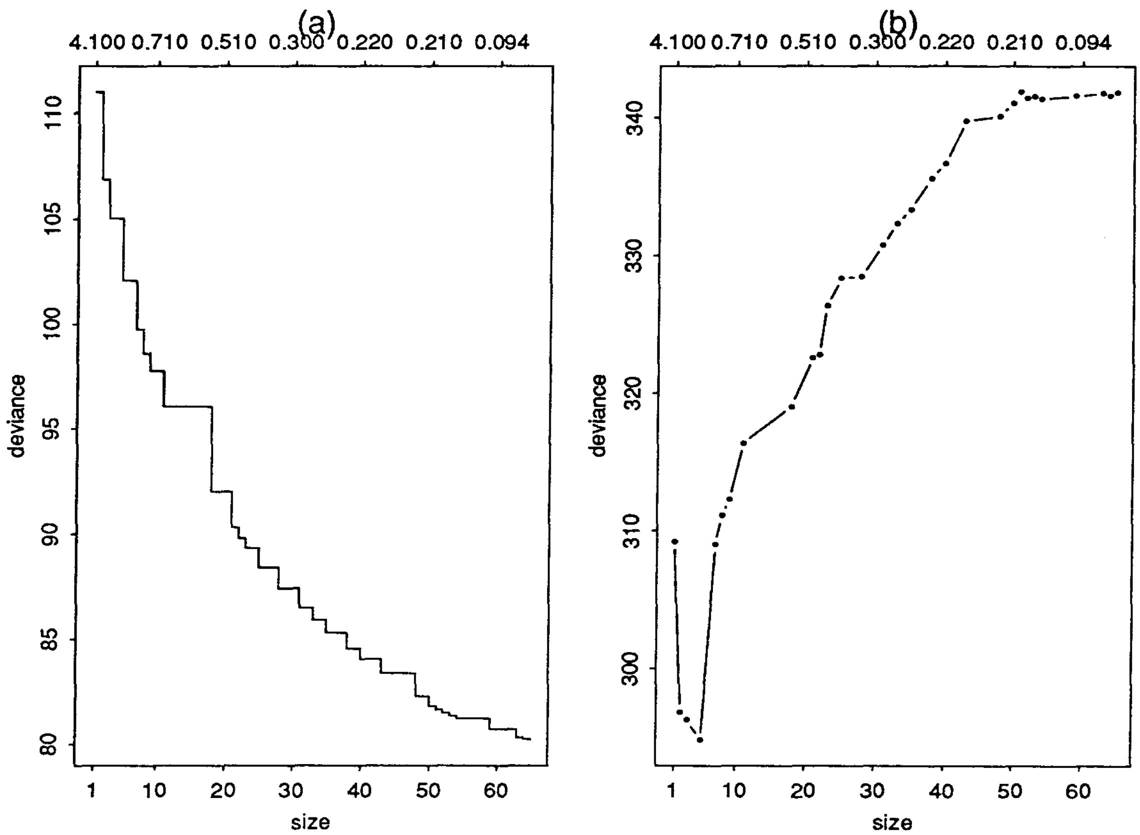
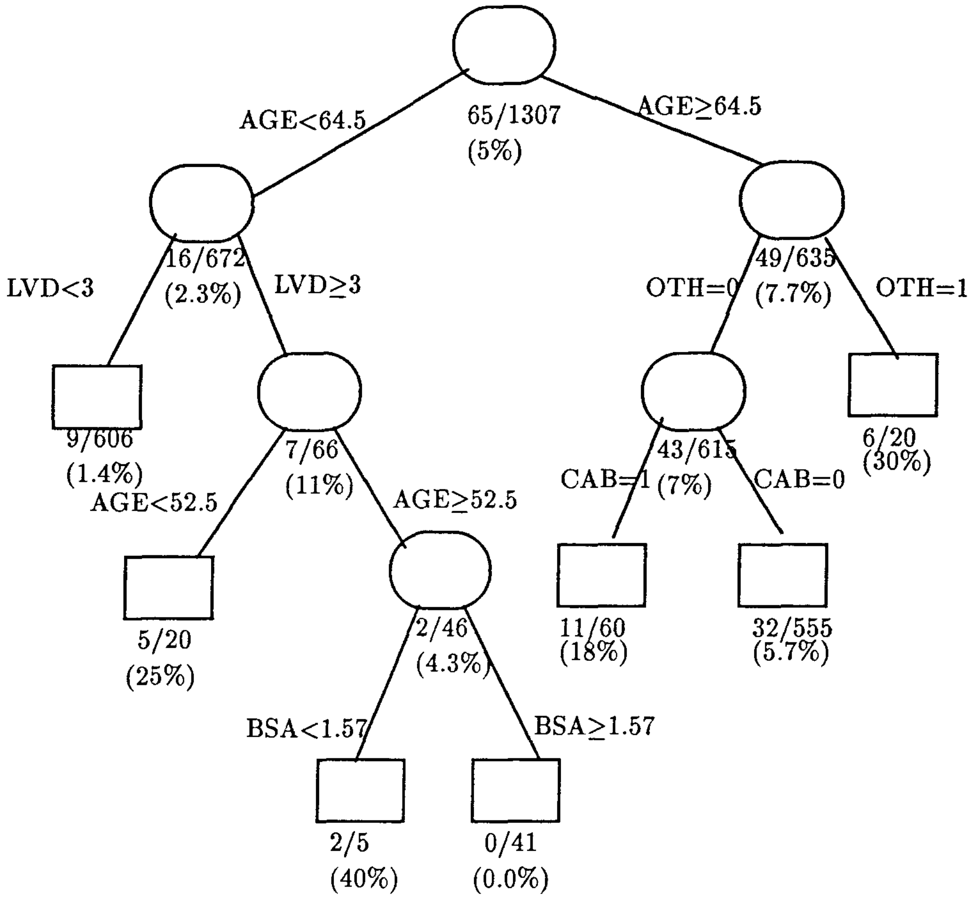


Figure 4: Plots of deviance versus size for sequences of subtrees. (a): sequence obtained from sample data; (b): sequence evaluated on test data

Also take STA for example. We use the variables obtained from the initial data analysis as the predictor variables. Each of these variables is considered to split the sample data set root node (with 1307 patients of which 65 died). In the first round, AGE is the variable leading to two nodes that are the most different (with mortality rates 16/672 and 49/635 respectively-refer Figure 5). We continue this procedure and use the same group of predictor variables to split each of the two nodes. For example, the winner for the left node is LVD while OTH is the best one for the right node. This process is continued until in each node there are less than 10 patients. See Figure 3 for the resulting tree. This tree has 63 terminal nodes and is obviously too large to use so that pruning is necessary.

Figure 4(a) displays the plot of deviance versus size (number of nodes) for the sequence of subtree of above tree. It should not be surprising that the sequence produced provide little guidance on what size tree is adequate. But we can use new data to guide the selection of the right size tree by using the pruning algorithm described in section 5.2. In S/Splus, this function provide a sequence of subtree and the deviance evaluated on the test data. Figure 4(b) illustrated this functionality for the STA data. Usually this sequence will not be monotone and the turning point will suggest the right size; for example, for the STA data, a seven-node tree is suggested and $\alpha = 1.125$.

The binary tree (see Figure 5) has three terminal nodes corresponding to *low* risk, and four terminal nodes corresponding to *high* risk. The size of the risk of a node is defined relatively to the sample population risk. Patients whose ages are over 64.5 years and have some other previous cardiac operation appear to have a relatively high risk of mortality. Those patients who are less than 64.5 years old and have normal condition of left ventricular function seem to be at much lower risk than those in the same age group but with a worse condition on LVD. Body surface area also plays an important role here. As we can see, with the same condition on age and LVD referred to above, patients with a smaller body surface area tend to face a higher risk of mortality. Similar interpretations can be made for the tree models with the response variables being one of the complication variables; see Figures 5 to 10.



The number under each terminal node is the observed proportion of the 30 day operative mortality; for example, in the leftmost node, which has low risk, 9 out of 606 patients in the node died after the operation.

Figure 5: Tree model for STA

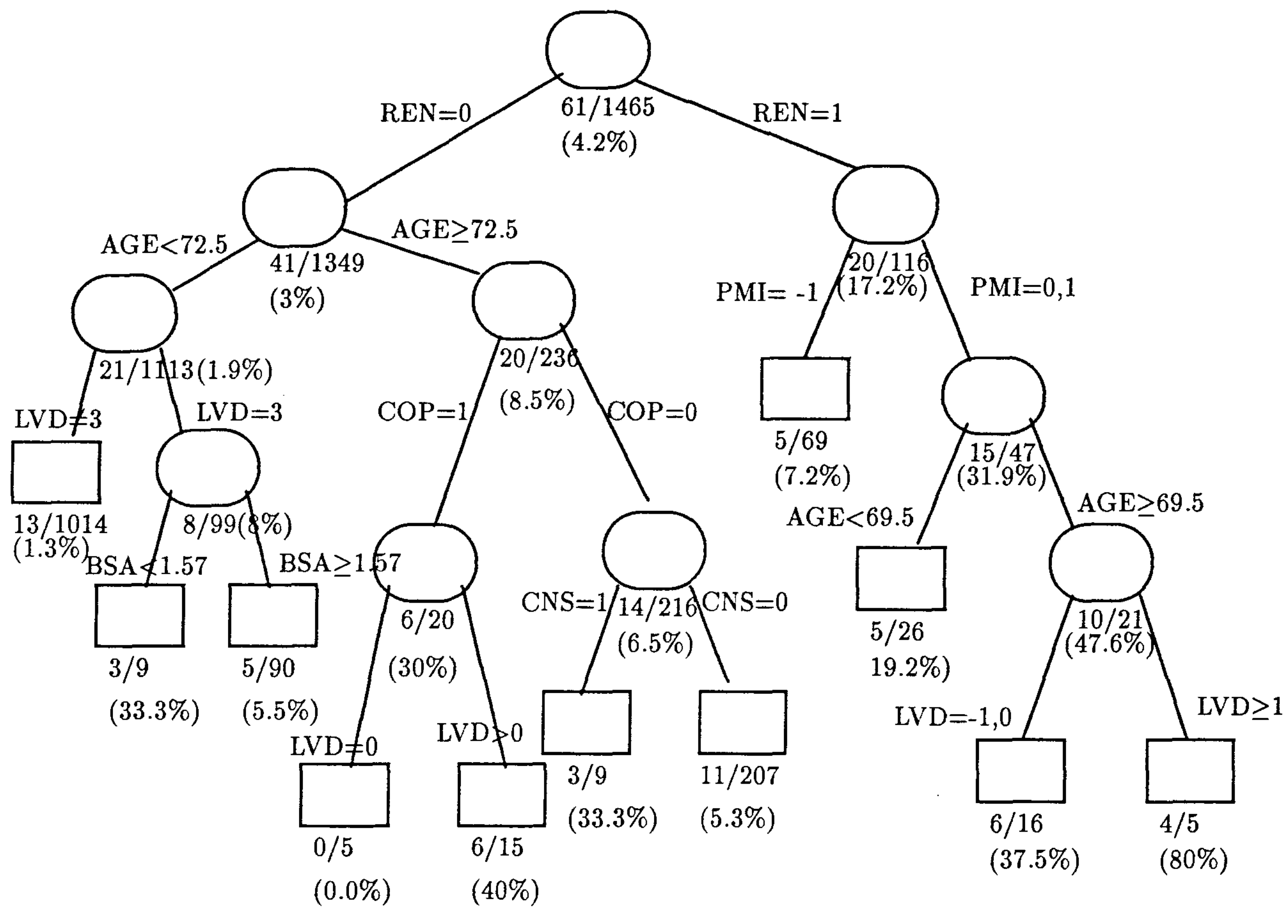


Figure 6: Tree model for REMS

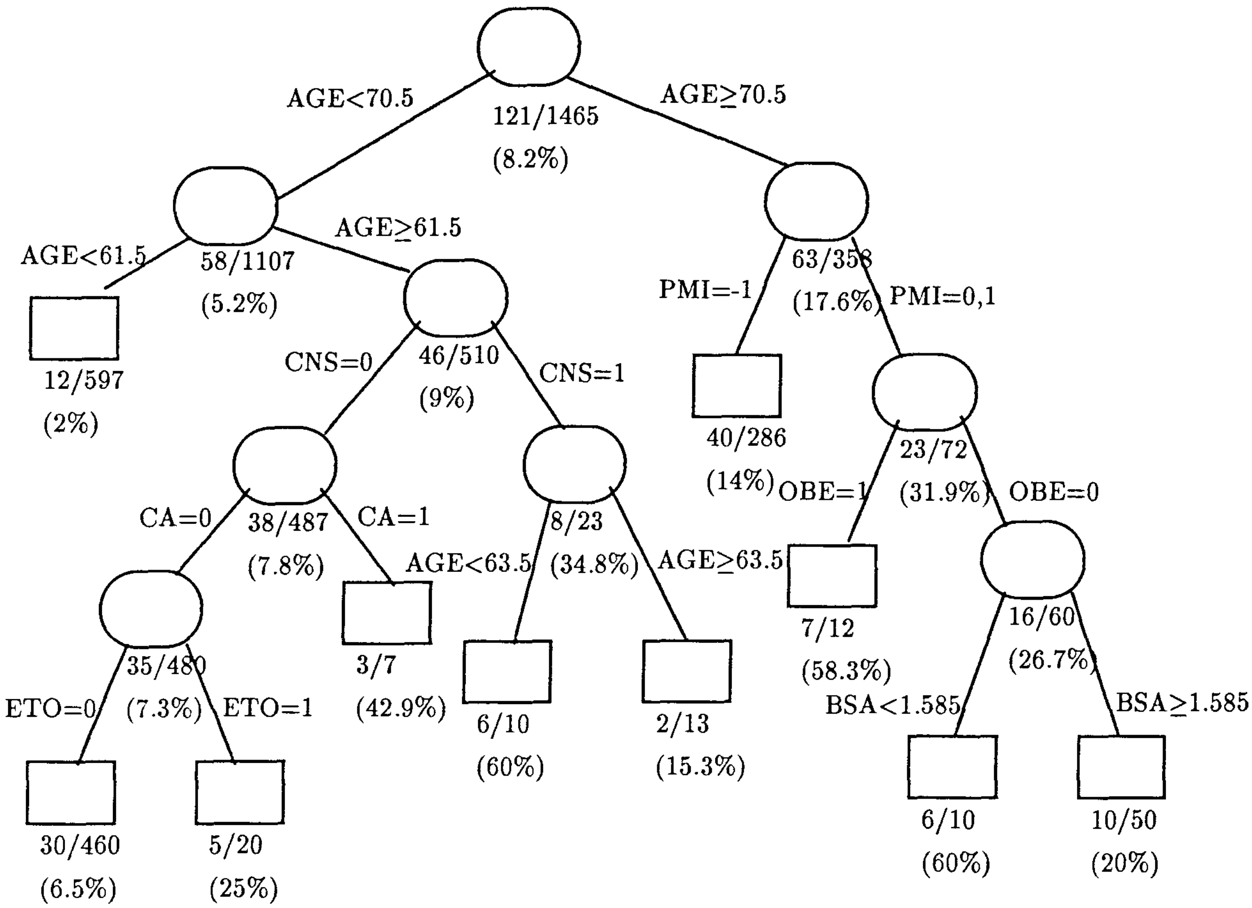


Figure 7: Tree model for NEMS

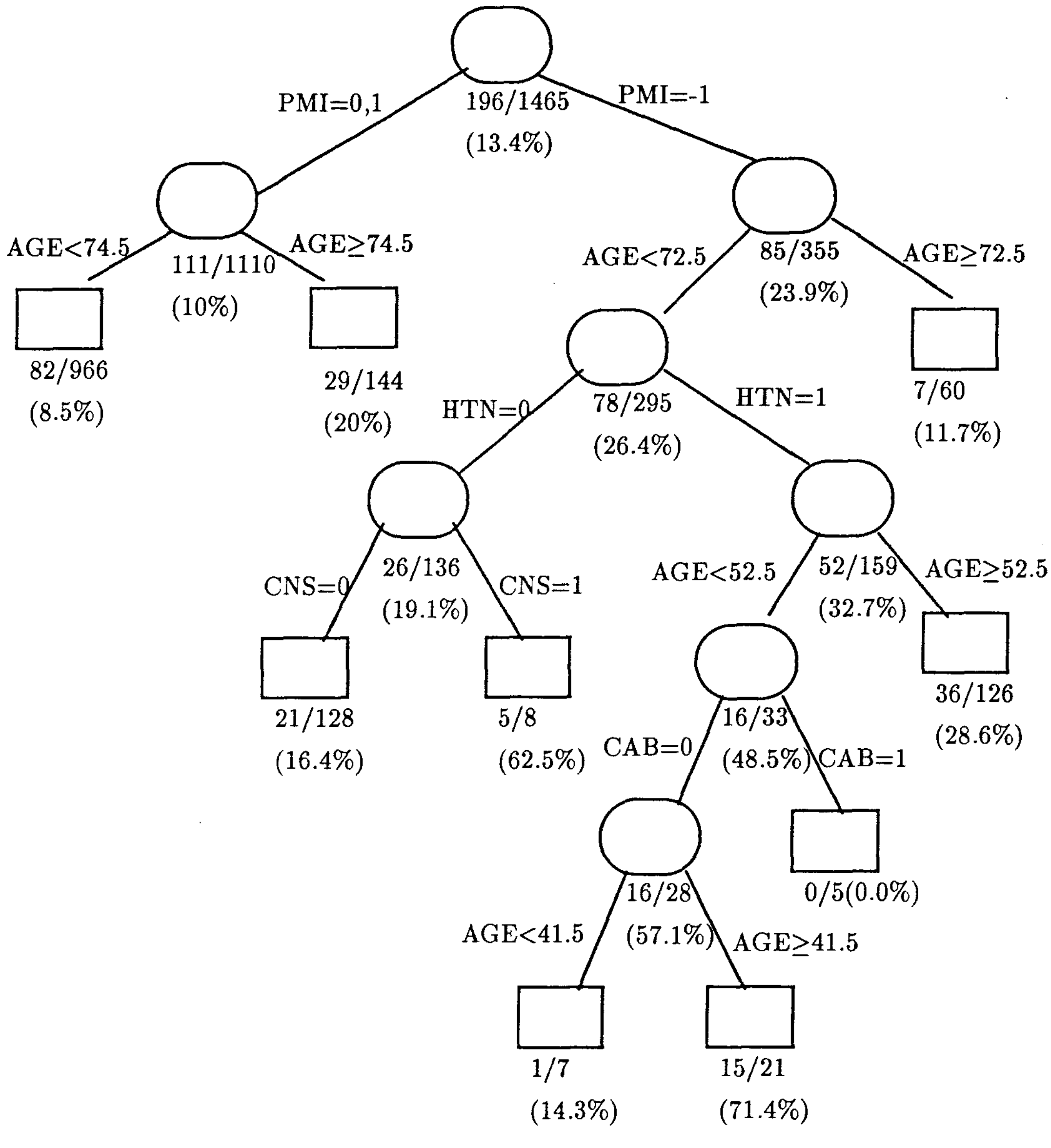


Figure 8: Tree model for PUMS

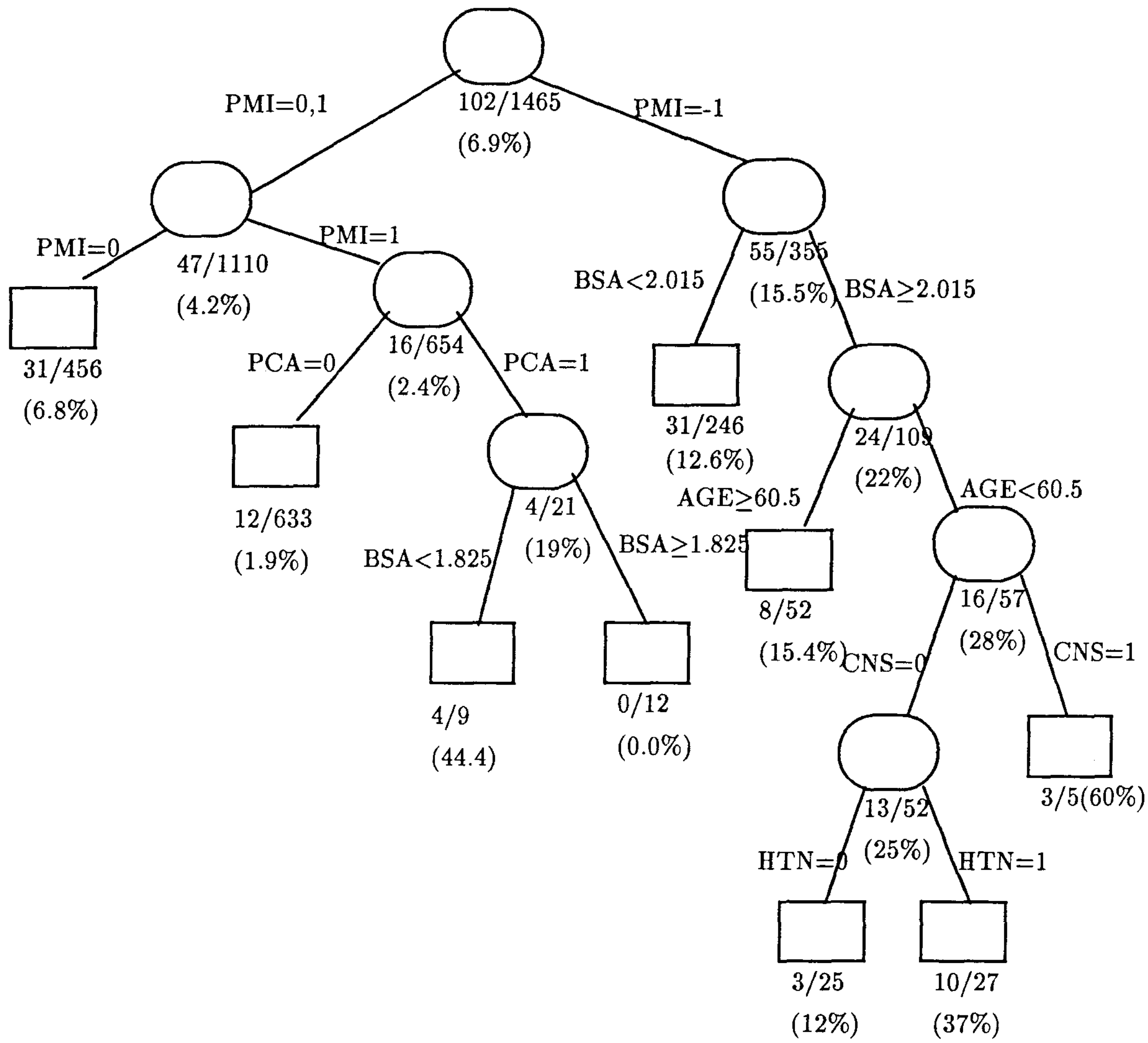


Figure 9: Tree model for MI

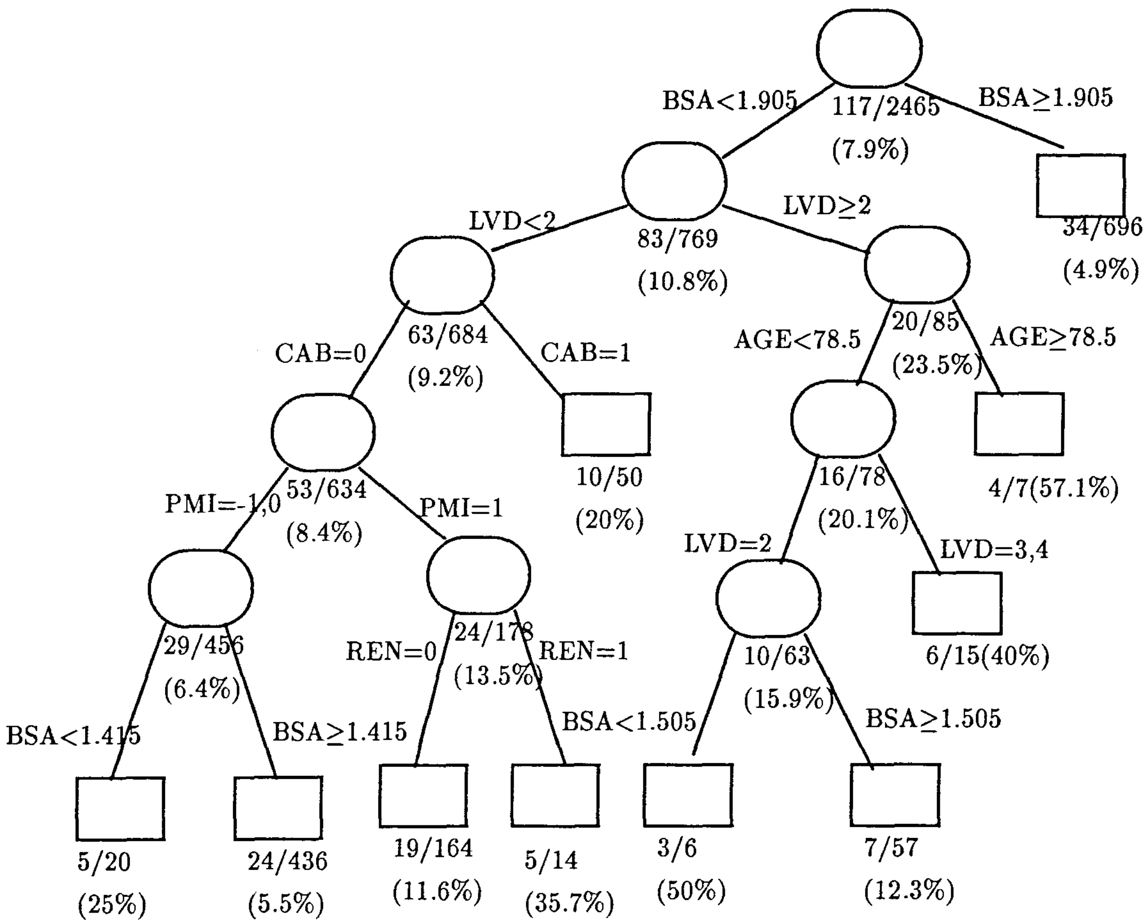


Figure 10: Tree model for LOMS

Chapter 6

Discussion and Conclusion

Our aim is to set up risk stratification models for some response variables. Beside STA, we are also interested in other variables, i.e., the complications. At first, using odds ratio analysis, we obtained some idea of the association between binary risk factors and the response variable. After that, two methods were applied using S/Splus. They are the logistic regression model and the tree-based model.

In the logistic regression procedure, there are several steps of work before achieving the final model. First, each predictor variable's association with the response variable was tested with the likelihood test. Only those which shows potential relation with the response variable (p-value $\leq$ 0.25) were selected. Linearity of the continuous variables to the response variable is checked and be confirmed. Secondly, if the dimension of variables space is still large ($\geq$ than 10), a stepwise procedure was used. This would usually reduce the dimension less than 10. Thirdly, a best subset program was run on the sample, this is to give some alternative candidate models. After that, the Hosmer-Lemeshow grouping test was applied to check the goodness-of-fit and finally the best models were selected.

In the tree-based model, a classification tree was grown based on a recursive partitioning method in S/Splus. After that, a pruning algorithm was applied to get a tree with 4 to 6 levels.

When doing statistical analysis, instead of using the entire data set, a sample subset was used. (This was done partly in order to reduce computational time.) While sampling, the 3000 data entries at the end of the data file were untouched. This latter part was kept for validation testing.

The missing values in PMI and LVD bring some troubles to the analysis. Fortunately, these are categorical variables and we can code an extra category labeled as "no information". Consequently, PMI and LVD become categorical variables with 3 and 6 categories respectively, and dummy variable are created to replace them in the logistic regression models, but not in the tree models.

A summary of important risk factors for the dependent status variable STA and the complication variables REMS, NEMS, PUMS, MI and LOMS is given in Table 25. The binary risk factors which have significant odds ratios, and the risk factors which are included in the final logistic model and the final pruned tree model are listed (in order of importance). There is substantial overlap in the important risk factors from the 3 methods. The range of predicted risks are summarized in Table 25 for each of the dependent variables. The range for the logistic regression model is wider than that of the tree model because the logistic regression separates out the cases more than the tree.

AGE, as we participated, is associated with all outcomes.

It seems that SEX is not strongly associated with operative mortality and complications since it does not appear as a predictor variable in any of the final logistic regression or tree models. This could be because the variable BSA accounts somewhat for the gender variable (that is BSA is a partial surrogate). Male and female are facing the same level of risk for the same value of BSA and other variables.

What about body surface area? BSA has strong association with the operative mortality. But with the complications, it is not always as important. It is an important risk factor to LOMS and takes a middle position of importance in predicting PUMS and MI. Its importance in the models for REMS and NEMS is much less.

Prior cardiac operation plays a very important role in predicting the 30 day operative mortality.

Table 25: List of risk factors and prediction range for the different methods and different outcomes

	Outcome					
Method	STA	REMS	NEMS	PUMS	MI	LOMS
Odds Ratio	VAL CAB CNS OTH	REN LIV CNS COP DIA HTN	CNS CA OBE REN HTN	COP OBE ETO CNS DIA HTN	LIV COP REN OBE HTN	RHE REN CAB SEX
Logistic Regression	OTH CAB BSA AGE LVD CNS PCA CA	REN AGE COP PMI CNS SEX DIA LIV	AGE CNS PMI CA ETO PCA	PMI PCA CNS REN OBE	PMI PCA CNS COP	BSA PMI CAB LVD OTH RHE AGE HTN DIA
Prediction Range	0 – 0.97	0 – 0.92	0 – 0.76	0 – 0.40	0 – 0.59	0 – 0.93
Tree-based Model	AGE OTH LVD CAB BSA	REN AGE PMI COP LVD CNS BSA	AGE PMI CNS OBE BSA CA ETO	PMI AGE HTN CNS CAB	PMI PCA BSA AGE CNS HTN	BSA LVD AGE CAB PMI REN
Prediction Range	0.0 – 0.4	0.01 – 0.80	0.02 – 0.6	0.0 – 0.625	0 – 0.60	0.04 – 0.57

Two out of three variables, CAB, OTH, VAL appear in logistic model and tree model. It is not surprising that these variables are mostly weighted by cardiologists. But they seem have weak association with some of the complications.

PMI and LVD are also important predictor variables. As we know, they are measuring roughly the same thing – damage of heart muscle; they seldom appear together in the same model and play the role alternately.

As to the diseases, CNS and HTN should be paid much attention to. CNS appears in all the logistic models except the one with LOMS as outcome, although in every appearance, its position in importance is around the middle. HTN's function is revealed when analyzing its relation with complications. For all complications, patients who possess hypertension will surely have higher risk. Actually, all the diseases studied appear as important predictors in different models for predicting the various complications. But they tend to be associated with particular outcomes, for example, REN (renal failure) is the most important risk factor in the REMS model and PCA seem closely related with the MI complication.

One of the difficulties in this study was that the patient data were from several populations. The technique and experience may vary across different hospitals. One possibility is to separate the patients and develop the models within one population and then seek generalization. Unfortunately, the MCR database did not provide such information. Another approach which may be more feasible is to include the variables which describe the operation such as X-clamp time, type of oxygenator use, etc., to capture the difference between populations since the database did record these information.

As we know, the logistic regression is more powerful to get prediction probabilities in the range of 0.1 to 0.9. So, although the logistic regression and tree models nearly identify the same group of risk factors, when predicting we suggest the latter be used since its prediction range is narrower. Another suggestion, also proposed by cardiologists, is to separate the population according to the prior cardiac operation done. Some such subpopulations are:

a). patient who had a coronary bypass operation,

b). patient who had a valve operation,

c). patient who had both operations.

Hopefully, the analyses based on these subpopulations will lead much more interesting and important findings.

Bibliography

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle, in *Second International Symposium on Information Theory.* Akademia Kiado, Budapest, 267-281.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press, Boston.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone C. J. (1984). *Classification And Regression Trees.* Wadsworth, Belmont, CA.

Chambers, John M. and Hastie, Trevor J. (1990). *Statistical Models in S.* Wadsworth, Belmont, CA.

Ciampi, A., Chang, C-H., Hogg, S. and Mckinney S. (1987) Recursive partitioning: a versatile method for exploratory data analysis in biostatistics, in *Biostatistics* (eds. I. B. MacNeill and G. J. Umphrey). D. Reidel Publishing, New York.

Draper, N.D. and Smith, H. (1981). *Applied Regression Analysis,* Second Edition. Wiley, New Yrok.

Fleiss, J. (1979). Confidence intervals for the odds ratio in case-control studies: state of art, *Journal of Chronic Diseases,* **32**, 69-77.

Furnival, G. M., and Wilson, R. W. (1974). Regression by leaps and bounds, *Technometrics,* **16**, 499-511.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors (1983). *Understanding Robust and*

Exploratory Data Analysis. Wiley, New York.

Hosmer, D. W. and Lemeshow, S. (1980). A goodness-of-fit testing for the multiple logistic regression model. *Communications in Statistics*, **A10**, 1043-1069.

Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression.* Wiley, New York.

Lemeshow, S., Teres, D., Avrunin, J. S., Pastides, H. (1988). Predicting the outcome of intensive care unit patients. *Journal of the American Statistical Association*, **83**, 348-356.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models.* Second Edition. Chapman Hall, London.

Morgan, J. N., and Sconquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of American Statistical Association*, **58**, 415-434.

Pregibon, D. (1981). Logistic refression diagnostics. *Annals of Statistics.* **9**, 705-724.

Rao, C.R. (1973). *Linear Statistical Inference and Its Application.* Second Edition. Wiley, New York.

Rothman, K. J. (1986). *Modern Epidemiology.* Little Brown, Boston.

Walter, S. D. (1987). Point estimation of the odds ratio in sparse 2x2 contingency tables, in *Biostatistics*(eds. I. B. MacNeill and G. J. Umphrey). D. Reidel Publishing, New York.

Weisberg, S. (1980). *Applied Linear Regression.* Wiley, New York.

Appendix A

Merged Cardiac Registry

MERGED CARDIAC REGISTRY
AN INTERNATIONAL DATABASE

DENDRITE SYSTEMS, INC.

___1: **DEMOG # :** *Entered by the system*

___2: **PRIOR ENTRY REGISTRY MCR :** *Entered by the system*

___3: **DATE OF SURGERY :** *Enter date MM/DD/YY*

___4: AGE : *Enter age in years*

___5: SEX : *Entered by the system*

___6: *(Reserved for future use)*

___7: **PRIOR MI :** 1=No, 2=Yes

___8: **MOST RECENT MI :** 1=0-6h, 2=6h-24h, 3=1d-7d, 4=1w-6w, 5=>6w

___9: **OTHER DISEASES :** 0[ ]=Other, 1[ ]=Obesity, 2[ ]=COPD, 3[ ]=Diab, 4[ ]=Chol>200
5[ ]=Chol>300, 6[ ]=Renal, 7[ ]=Htn, 8[ ]=ETOH, 9[ ]=Drug Abuse
10[ ]=Marfans, 11[ ]=HIV+, 12[ ]=AIDS, 13[ ]=CA, 14[ ]=Blood
15[ ]=Liver, 16[ ]=CNS, 17[ ]=Prior CVA, 18[ ]=RheumHD,
19[ ]=Pulm Htn, 20[ ]=Chronic Dialysis

___10: **SMOKING NOW :** 0=No, 1=Q>2y, 2=Y<1pk/d, 3=Y>1pk/d

___11: **PRIOR CARD SURG :** 0[ ]=Other, 1[ ]=None, 2[ ]=CABG, 3[ ]=Valve, 4[ ]=Cong
5[ ]=Pacemaker

___12: **LV DYSFUNCTION :** 1=Nor, 2=40-49%, 3=30-39%, 4=20-29%, 5=<20%

___13: **LVEF :** Ejection Fraction, enter %

___14: **CAD > 70% :** 1[ ]=No, 2[ ]=AD, 3[ ]=CX, 4[ ]=RC, 5[ ]=Branch, 6[ ]=L Main,
7[ ]=1 Vessel, 8[ ]=2 Vessel, 9[ ]=3 Vessel

___15: **OTHER CARD PATH :** 0[ ]=Other, 1[ ]=Ao St, 2[ ]=Ao Insf, 3[ ]=Mitr St,
4[ ]=Mitr Insf, 5[ ]=Tricusp, 6[ ]=Pulm, 7[ ]=Cong
8[ ]=Acq VSD, 9[ ]=LV Aneur, 10[ ]=Ao Aneur
11[ ]=Ascending diss, 12[ ]=Decending diss

___16: **DATE MOST RECENT PTCA :** Enter date, leave blank if none

___17: **PTCA RESULT :** 1[ ]=N/A, 2[ ]=Success, 3[ ]=Failed, 4[ ]=Had complication

___18: **NO. PTCA VESSELS :** Enter number of vessels dilated

___19: **REASON FOR OP :** 0[ ]=Other, 1[ ]=Ang, 2[ ]=Urgent, 3[ ]=Arrh, 4[ ]=Anat, 5[ ]=Fail PTCA, 6[ ]=Tumor, 7[ ]=Endocarditis, 8[ ]=Trauma, 9[ ]=Ao diss 10[ ]=Ao Aneur

___20: **PRE-OP STATUS :** 1=Elect, 2=Urgent, 3=Emerg, 4=Desperate

___21: **HEMODYNAMIC STAT:** 1=Stbl, 2=Stbl on meds, 3=Unstbl on meds 4=Cardiogenic shock on meds/IABP

___22: **OXYGENATOR :** 0=Other, 1=H1500, 2=Shiley, 3=TMO, 4=CMII, 5=Sci25 6=Sci35, 7=Maxima, 8=BCM7, 9=Sarns, 10=Terumo 11=SciUltra

___23: **OTHER OP DEVICES :** 0[ ]=Other, 2[ ]=Cel Sav, 3[ ]=HemoConcen 4[ ]=Dial Filter, 5[ ]=IABP, 6[ ]=BioMed Pump, 7[ ]=LHAD 8[ ]=RHAD, 9[ ]=Art Filter, 10[ ] Plasma Phor 11[ ]=MyoTempProb, 12[ ]=Cooling Pad, 13[ ]=Delphin Pump

___24: **THROMBOLYTIC Rx :** 1[ ]=tPA, 2[ ]=Strepto, 3[ ]=Urokin, 4[ ]=IntrCor, 5[ ]=IntrVein

___25: **CARDIOPLEGIA :** 1[ ]=None, 2[ ]=Cryst, 3[ ]=Blood, 4[ ]=Retro (Cor Sinus), 5[ ]= Intermit Clamp

___26: **X-CLAMP TIME :** Enter time in minutes

___27: **BODY SURFACE AREA :** Enter in square meters (e.g., 2.3)

___28: **NO. CABGs :** Enter number of distal anastomoses

___29: **VALVES REPLACED :** 1[ ]=Ao, 2[ ]=Mitr, 3[ ]=Tri, 4[ ]=Aocombined c Ao graft

___30: **REPAIRS :** 1[ ]=Ao, 2[ ]=Mitr, 3[ ]=Tri, 4[ ]=Cong, 5[ ]=Acq VSD, 6[ ]=LV Aneur

___31: **AORTIC PROSTHESIS :** 0=Other, 1=SE, 2=BS, 3=St. J, 4=Ed Porc, 5=Hancock 6=Froz Homo, 7=Medtronic

___32: **MITRAL PROSTHESIS :** 0=Other, 1=SE, 2=BS, 3=St. J, 4=Ed Porc, 5=Hancock 6=Froz Homo, 7=Ring, 8=Medtronic, 9=Omnisci

___33: (Reserved for future use)

___34: (Reserved for future use)

___35: **BLOOD PRODUCTS :** 1[ ]=Fresh Froz Plasma, 2[ ]=Platelets, 3[ ]=Cryo

___36: **DONOR TRANSFUSIONS :** Enter number of units

___37: **AUTOLOGOUS TRANSFUSIONS :** Enter number of units

___38: *(Reserved for future use)*

___39: **COMPLICATIONS :** 0[ ]=Other, 1[ ]=Reop/Bleed, 2[ ]=Renal/Mild, 3[ ]= Renal/Sev
4[ ]=Wound/Sev, 5[ ]=Neuro/Mild, 6[ ]=Neuro/Sev, 7[ ]=Pulm/Mild
8[ ]=Pulm/Sev, 9[ ]=MI, 10[ ]=Low Out/Mild, 11[ ]=Low Out/Sev
12[ ]=Clotting, 13[ ]=Sepsis, 14[ ]=GI/GB, 15[ ]=DIC

___*40: (Reserved for future use)*

___*41: (Reserved for future use)*

___42: **DAYS IN ICU :** Enter number of days

___43: **DAYS SURG/DISCH :** Enter number of days

___44: **PARSONNET RISK :** Calculated & entered by the system

___*45: (Reserved for future use)*

___46: **TRANSFER TO NEW ENTRY :** Entered by the system

___47: **DISCHG/30 DAY STATUS :** 0=UNK, 1=ALIVE, 2=DIED IN OR, 3=DIED IN HOSP/30D
4=REOP, 5=DIED LATE CARD, 6=UNREL DEATH
9=LOST TO FU

Appendix B

Expanded Definitions for Merged Cardiac Registry

MERGED CARDIAC REGISTRY

Expanded Definitions for Version 2

1. **DEMOG #:**
   (There is no user correspondence file set up for this question.) The program will use your Demographic number. It is the only patient identification that is sent to Dendrite. At Dendrite an offset will be added which is group specific. The offsets will not be published.

2. **PRIOR ENTRY REGISTRY MCR:**
   (There is no user correspondence file set up for this question.) This information is entered at the time of transfer. It follows reoperations for both valves and bypasses.

3. **DATE OF SURGERY:**
   (There is no user correspondence file set up for this question.) This is the date for this procedure. If a patient is in more than one Source Registry on the same date, the entries will be merged. If the patient has two entries in the same Source Registry on the same date, they will also be merged.

4. **AGE:**
   This is the patient's age in years at the time of operation. If you don't have this question in your Source Registry(ies), you should consider adding it.

   For a minimal fee, we can give you the ability to calculate a default answer for age if registry question *"DATE OF SURGERY"* and demographic question *"DATE OF BIRTH"* have been entered. Please call Dendrite for information on this feature.

5. **SEX:**
   (There is no user correspondence file set up for this question.) The program gets this information from your demographic file automatically.

6. *Reserved for future use.*

7. **PRIOR MI:**
   1=No (If no clinical MI.)
   2=Yes (One or more clinical MIs.)
   Do not include silent MIs diagnosed only on angiography.

8. **MOST RECENT MI:**
Select the insure that reflects the interval from the most recent MI to this operation. This could be important as a risk factor. If you don't have this question in your Source Registry(ies), you should consider adding it.

9. **OTHER DISEASES:**
0[ ]=Other (Use "other" to record a disease not listed in the answers but that you feel is significant.)
1[ ]=Obesity (1.5x expected body weight.)
2[ ]=COPD (Patient with distinct limitations revealed at time of study or on treatment - bronchodilators, etc.)
3[ ]=Diab (Patient on oral meds or insulin.)
4[ ]=Chol >200 (Patients from 200 - 299.)
5[ ]=Chol >300 (Patients above 300>)
6[ ]=Renal (Patients with creatinines above 2.5 not on dialysis.)
7[ ]=Hypertension (History of treatment.)
8[ ]=ETOH (Patients who have undergone treatment or come in intoxicated.)
9[ ]=Drug Abuse (History or current use of cocaine, heroin, etc.)
10[ ]=Marfans (Patient with diagnosis or you diagnose.)
11[ ]=HIV+ (Positive test for AIDS. Not clinical disease.)
12[ ]=AIDS (Clinical disease.)
13[ ]=CA (History of malignant disease - cured or not.)
14[ ]=Blood (History of anemia not related to blood loss; e.g., sickle cell. Also, leukemia or lymphoma even if in remission.)
15[ ]=Liver (History of hepatitis, cholangitis, but not gall bladder disease.)
16[ ]=CNS (History of brain abscess, encephalitis, or clinical dementia.)
17[ ]=Prior CVA (History of stroke with or without residual.)
18[ ]=RheumHD (History of Rheumatic Heart Disease.)
19[ ]=Pulm Htn (PA pressures >60mmHG systolic.)
20[ ]=Chronic Dialysis (Not successful transplants.)

10. **SMOKING NOW:**
Smoking now is within ten (10) days or at the time of catheterization. Consider answer 2 to mean mild and answer 3 to mean heavy. Do not count pipe smoking or chewing tobacco.

11. **PRIOR CARDIAC SURG:**
Use "Other" for tumors, stab wounds, vinebergs, etc. We have added answer 5[ ]=Pacemaker.

12. **LV DYSFUNCTION:**
Select the answer that reflects the estimate from non-planimetry or echo, gated, etc.

13. **LVEF:**
This is the actual left ventricular ejection fraction. We only consider planimetry by angiography a valid means to answer this question. For other means (gated blood pool, echo, use question #12.)

14. **CAD >70%:**
Since it is possible that the answer for this question could come from multiple Source Questions, a no answer will be considered the same as none.
1[ ]=No (None/no coronary disease)
2[ ]=LAD
3[ ]=Cx (Includes the large OM as well if >70%.)
4[ ]=RCA (Includes PDA.)
5[ ]=Branch (Includes intermediate, large diagonal but does not define which system.)
6[ ]=L. Main
7[ ]=1 Vessel Disease
8[ ]=2 Vessel Disease
9[ ]=3 Vessel Disease

15. **OTHER CARDIAC PATHOLOGY**
0[ ]=Other (For dissections of the aorta, tumors of the heart.)
1[ ]=Ao St (Aortic stenosis with a gradient >60mmHG or valve area <.8CM.)
2[ ]=Ao Insuf (Aortic insufficiency moderate or great.)
3[ ]=Mitr St. (Mitral stenosis with a gradient >60mmHG.)
4[ ]=Mitr Insuf (Significant mitral leak with V-waves.)
5[ ]=Tricusp (Either stenosis, leak, or both.)
6[ ]=Pulm (Valve stenosis.)
7[ ]=Cong (Any diagnosis of congenital heart disease.)
8[ ]=Acq VSD (VSD post MI or surgery.)
9[ ]=LV Aneur (Localized paradoxical segment.)
10[ ]=Ao Aneur (Ascending, arch, or descending aneurysm.)
11[ ]=Asc Diss (Ascending dissection of the aorta.)
12[ ]=Dsc Diss (Descending thoracic aortic dissection.)

16. **DATE MOST RECENT PTCA:**
Enter date. This new format will allow date arithmetic later. If you have data in the old format, we can help you transform it.

17. **PTCA RESULT:**
Enter the initial 5-day result judged by the surgeon. A complication would include MI, MI in progress, perforation, etc., within this 5-day period.

18. **NUMBER OF VESSELS PTCA'd:**
Enter the number of vessels dilated prior to this operation. (A triple PTCA would count as 3.)

19. REASON FOR OP:

0[ ]=Other (Use "Other" to record an answer not listed, but that you feel is significant.)
1[ ]=Ang (Angina uncontrollable with meds.)
2[ ]=CHF (Congestive heart failure - low output state.)
3[ ]=Arrh (Arrhythmia.)
4[ ]=Anat (Anatomy; left main, etc. in otherwise stable patient.)
5[ ]=Failed PTCA (PTCA that was performed within five (5) days if you are treating the same vessel.)
6[ ]=Tumor
7[ ]=Endocarditis (Patient has had positive cultures.)
8[ ]=Trauma
9[ ]=Ao Dissection
10[ ]=Ao Aneurysm

20. PREOP STATUS:

1=Elect (Elective scheduled case.)
2=Urgent (Case moved up on schedule.)
3=Emerg (Emergency case-- do ASAP.)
4=Desperate (Case that has arrested, is very near death, or in severe low output.)

21. HEMODYNAMIC STAT:

1=Stbl (Stable patient.)
2=Stbl on meds (CI >2 on meds or IABP.)
3=Unstbl on meds (CI <2 on meds or IABP.)
4=Cardiogenic shock on meds/IABP (CI <2 and falling.)

22. OXYGENATOR:

0=Other (Use "Other" to record a answer that is not listed, but that you feel is significant.)
1=H1500 (Harvey bubbler includes H1300.)
2=Shiley (Shiley bubbler.)
3=TMO (Travenol membrane.)
4=CMII (Cobe membrane.)
5=Sci25 (SciMed SM25.)
6=Sci35 (SciMed SM35.)
7=Maxima (J&J (Medtronic) membrane.)
8=BCM7 (Bentley membrane.)
9=Sarns (Membrane.)
10=Terumo (Membrane.)
11=SciUltra (SciMed Ultrox I.)

23. OTHER OPERATIVE DEVICES:

0[ ]=Other (Use "Other" to record an answer that is not listed, but that you feel is significant.)
Answer 1 has been deleted.
2[ ]=Cell Saver (Any brand.)
3[ ]=HemoConcen (Ultra filtration device to remove H2O.)
4[ ]=Dial Filter (Renal dialysis filter in circuit.)
5[ ]=IABP (Intra or post op.)
6[ ]=BioMed Pump (Biomedicus pump rather than roller pump.)
7[ ]=LHAD (Any long term use of left heart assist device post bypass.)
8[ ]=RHAD (Any long term use of right heart assist device post bypass.)
9[ ]=Art Filter (Any filter in the arterial line.)
10[ ]=Plasma Phor (For the use of plasma phoresis for platelet rich plasma.)
11[ ]=MyoTmpProb (For the use of myocardial temps where monitored.)
12[ ]=Cooling Pad (If a cooling pad is placed under or on the heart during crossclamp.)
13[ ]=Delphin Pump (Sarns centrifical pump rather than roller pump.)

24. THROMBOLYTIC Rx:

1[ ]=tPA (tPA used within 24 hours.)
2[ ]=Strepto (Streptokinase used within 36 hours.)
3[ ]=Urokin (Urokinase infused.)
4[ ]=IntraCor (Intracoronary infusion used.)
5[ ]=IntraVein (Intravenous.)

26. CARDIOPLEGIA:

This question has been changed to a type 7.
1[ ]=None (None or just slush.)
2[ ]=Cryst (For cold +/- high k+.)
3[ ]=Blood (For cardioplegia solutions containing blood.)
4[ ]=Retro-cor sinus (Any use of retrograde perfusion.)
5[ ]=Intermit Clamp (Can be combined with any of the above answers.)

26. X-CLAMP TIME:

Enter your answer in minutes.

27. BODY SURFACE AREA:

This is a new question. Enter your answer in square meters (e.g., 2.3)

28. NO CABGs:

Enter the total number of distal coronary anastomoses for this procedure.

29. **VALVES REPLACED:**
Enter those valves replaced with a prosthesis. If this information is in your subprocedure section, you may be required to set up one or more secondary questions to create the criteria for transfer.

30. **REPAIRS:**
This includes debridement, commissurotomy, partial resection. May be combined with questions 31-32 if the repair fails or needs supplement.

31. **AORTIC PROSTHESIS:**
Enter the type of prosthesis used.

32. **MITRAL PROSTHESIS:**
Enter the type of prosthesis used.

33. Reserved for future use.

34. Reserved for future use.

35. **BLOOD PRODUCTS:**
Enter any use of the blood products listed on the MCR.

36. **DONOR TRANSFUSIONS:**
Enter the number of units of bank blood/packed cells on this admission.

37. **AUTOLOGOUS TRANSFUSIONS:**
Enter the number of units of blood drawn 5 to 30 days preop for elective use at surgery. Do not enter blood withdrawn at the time of surgery or plasma phoresis.

38. Reserved for future use.

39. **COMPLICATIONS:**
0[ ]=Other (Use "Other" to record an answer not listed, but that you feel is significant.)
1[ ]=Reop/Bleed (Reoperation for bleeding, suspected tamponade.)
2[ ]=Renal/Mild (Mild renal shutdown not requiring dialysis.)
3[ ]=Renal/Sev (Severe renal shutdown requiring dialysis.)
4[ ]=Wound/Sev (Dehiscence or infection - for sternal wounds only.)
5[ ]=Neuro/Mild (Peripheral nerve, brachial plexus, confusion or CNS defect that clears before discharge.)
6[ ]=Neuro/Sev (CNS defect that does not clear in 7 days.)
7[ ]=Pulm/Mild (Pneumothorax, hemothorax, atelectasis, air leak.)
8[ ]=Pulm/Sev (Prolonged respiratory support, ARDS or pneumonia requiring antibiotics.)

COMPLICATIONS continued

9[ ]=MI (Intra- or post-op MI by EKG or enzymes.)
10[ ]=Low Output/Mild (Low output syndrome postop requiring drugs for a short time.)
11[ ]=Low Output/Sev (Severe low output syndrome postop with prolonged use of drugs or IABP.)
12[ ]=Clotting (Prolonged bleeding problems, low platelets, etc.)
13[ ]=Sepsis (Septicemia, pneumonia, wound infection, etc.)
14[ ]=GI/GB (GI bleed, perforated ulcer, cholecystitis, hepatitis, etc.)
15[ ]=DIC (Diffuse intravascular coagulation.)

40. Reserved for future use.

41. Reserved for future use.

42. **DAYS IN ICU:**
Enter the number of days - round off (e.g., 27 hrs. = 1 day, 30 hrs. = 2 days).

43. **DAYS SURG/DISCH:**
Enter the number of days from surgery to discharge.

44. **PARSONNET RISK:**
(There is no user correspondence file set up for this question.) Use the "R" option to get risk calculations once you have transferred your data to the MCR.

45. Reserved for future use.

46. **TRANSFER TO NEW ENTRY:**
(There is no user correspondence file set up for this question.) This is a system-generated question to follow re-entries in your registry(ies).

47. **DISCH/30D STATUS:**
0=UNK (This is a historical answer in use before the addition of follow-up.) **DO NOT USE THIS ANSWER WHEN CREATING YOUR CORRESPONDENCE FILE.**
1=ALIVE
2=DIED IN OR (Died in the operating room.)
3=DIED IN HOSP/30D (Died in or out of hospital within 30 days of surgery.)
4=REOP (Your reoperations only.)
5=DIED LATE CARDIAC (Died after 30-day interval, cardiac-related.)
6=UNRELATED DEATH (Died after 30-day interval, non-cardiac-related.)
9=LOST TO FU (Patient who can no longer be followed because cannot be located.)

The only follow-up transferred at this time is survival status. This is moved automatically if your Source Registry(ies) have follow-up.